UNIVERSIDAD NACIONAL
PEDRO RUIZ GALLO
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA PROFESIONAL DE MATEMÁTICA

TESIS

"Estimación de parámetros probabilísticos en modelos de mezclas gausianas para la segmentación en imágenes usando el algoritmo Expectation-Maximization"

Para optar el título profesional de
Licenciada en Matemáticas

INVESTIGADORA:

ATOCHE BRAVO MARIA JACQUELINE

ASESOR:

M.Sc. PERALTA LUI MARCO ANTONIO MARTIN

LAMBAYEQUE – PERÚ
2021

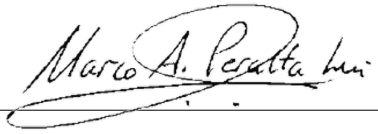
UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA PROFESIONAL DE MATEMÁTICA

Los firmantes, por la presente certifican que han leído y recomiendan a la Facultad de Ciencias Físicas y Matemáticas la aceptación de la tesis titulada "**Estimación de parámetros probabilísticos en modelos de mezclas Gausianas para la segmentación en imágenes usando el Algoritmo Expectation-Maximization**", presentada por la bachiller en matemáticas, Atoche Bravo María Jacqueline, en el cumplimiento parcial de los requisitos necesarios para la obtención del título profesional de licenciado en matemáticas.

Dr. Leandro Agapito Aznarán Castillo
Presidente Jurado de Tesis

M.Sc. Elmer Lluén Cumpa
Secretario Jurado de Tesis

Lic. Mat. Juan Antonio Cornetero Capitán
Vocal Jurado de Tesis

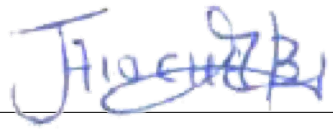


M.Sc. Marco Antonio Martín Peralta Lui
Asesor

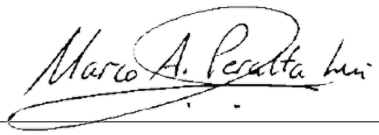
Fecha de Defensa: 19 de enero de 2021

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA PROFESIONAL DE MATEMÁTICA

"Estimación de parámetros probabilísticos en modelos de mezclas gausianas para la segmentación en imágenes usando el algoritmo Expectation-Maximization"



Bach. Mat. ATOCHE BRAVO MARIA JACQUELINE
Autor



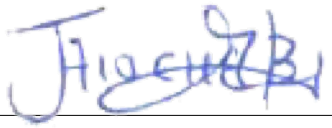
M.Sc. PERALTA LUI MARCO ANTONIO MARTIN
Asesor

LAMBAYEQUE – PERÚ
2021

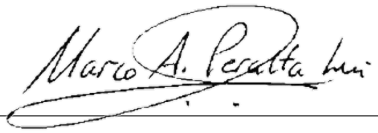
Declaración Jurada de Originalidad

Yo, Bach. María Jacqueline Atoche Bravo investigadora principal, y M. Sc. Marco Antonio Martín Peralta Lui asesor del trabajo de investigación "Estimación de parámetros probabilísticos en modelos de mezclas Gausianas para la segmentación en imágenes usando el Algoritmo Expectation-Maximization", declaramos bajo juramento que este trabajo no ha sido plagiado, ni contiene datos falsos. En caso se demostrara lo contrario, asumo responsablemente la anulación de este informe y por ende el proceso administrativo, a que hubiera lugar, que puede conducir a la anulación del título emitido como consecuencia de este informe.

Lambayeque, 19 de enero del 2021.



Bach. Mat. María Jacqueline Atoche Bravo
Autor



M.Sc. Marco Antonio Martín Peralta Lui
Asesor

Dedicatoria

A mis padres Gloria y Nelson, los grandes amores de mi vida.

A mi querida hermana Magdalena.

A mis abuelos Genoveva y Agustín.

AGRADECIMIENTOS

Me gustaria agradecer a todos que, de alguna manera, contribuyeron para que este trabajo pudiese ser realizado. En especial agradezco:

Primeramente a Dios, por haberme permitido llegar hasta este punto y haberme dado salud para lograr mis objetivos, además por su infinita bondad y amor.

A mis maestros, por su gran apoyo y motivación para la culminación de mis estudios profesionales. Gracias por su tiempo, por su apoyo así como por la sabiduría que me transmitieron en el desarrllo de mi formación profesional, en especial: al M. Sc. Marco Peralta Lui, por haberme guiado en el desarrollo de este trabajo y llegar a la culminación del mismo.

A toda mi família, principalmente a mis padres, Gloria y Nelson, por su inmenso amor que recibo todos los dias, por la comprensión y cariño en todos los momentos difíciles, por el apoyo en mis decisiones. A mi querida hermana Magdalena, por todos los consejos, apoyo, incentivo, que me animaron siempre a seguir adelante.

A todos mis amigos, que sempre creyeron en mí, que me incentivaron en el desarrollo de este trabajo.

RESUMEN

En el presente trabajo, se desarrolla una revisión sobre los principales modelos probabilísticos para el Reconocimiento de Patrones. Centrándonos en el estudio del Reconocimiento de Patrones aplicado a la segmentación de imágenes. La segmentación de imágenes digitales es el proceso de dividir o segmentar dicha imagen en varios grupos; es decir los píxeles de la imagen se dividen en distintios grupos o categorias. Este proceso, se realiza con el fin de simplificar o modificar la representación de una imagen en grupos más significativos o más fáciles de analizar. Nuestro trabajo propone estudiar el algoritmo *Expectation-Maximization* (EM) en particular en Modelos de Mezclas Gaussianas.

El algoritmo EM es un método iterativo, usado frecuentemente para estimar los valores de los parámetros en modelos probabilísticos en problemas de datos incompletos.

Para evaluar el desempeño del algoritmo, utilizamos imágenes de texturas, reales y de satélite. Después de procesar las imágenes por el algoritmo EM, las imágenes segmentadas pueden ser utilizadas en diferentes aplicaciones, como por ejemplo, reconstrucción de imágenes.

Palabras-claves: Modelos de mezclas Gaussianas, Algoritmo *Expectation-Maximization*, Máximima verosimilitud, Segmentación de imágenes.

ABSTRACT

In this work, a review is carried out about of the main probabilistic models for Pattern Recognition. We focus on the study of Pattern Recognition applied to image segmentation. Digital image segmentation is the process of to divide or to segment the digital image into several groups; this is, the pixels of the image are divided into different groups or categories. This process is carried out in order to simplify or modify the representation of an image in groups that are more significant or easier to analyze. Our work proposes to study the Expectation-Maximization (EM) algorithm in particular in Gaussian Mixture Models.

The EM algorithm is an iterative method, frequently used to estimate the values of the parameter in probabilistic models in incomplete data problems.

To evaluate the performance of the algorithm, we use images of textures, real and satellite images. After processing the images by the EM algorithm, segmented images can be used in different applications, such as image reconstruction.

Keywords: Gaussian mixture models, Expectation-Maximization Algorithm, Maximum likelihood, Image segmentation.

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DECANATO
Ciudad Universitaria - Lambayeque

ACTA DE SUSTENTACIÓN VIRTUAL N° 003-2021-D/FACFyM

Siendo las 9 10 am del dia martes 19 de Enero de 2021, se reunieron via plataforma virtual, https://meet google.com/ift-ucqr-sib, los miembros del jurado evaluador de la Tesis titulada:

ESTIMACIÓN DE PARÁMETROS PROBABILÍSTICOS EN MODELOS DE MEZCLAS GAUSIANAS PARA LA SEGMENTACIÓN EN IMÁGENES USANDO EL ALGORITMO ESPECTATION - MAXIMIZATION,

Designados por Resolución N° 874 – 2019-D/FACFyM de fecha 01 de Julio de 2019. Con la finalidad de evaluar y calificar la sustentación de la tesis antes mencionada, conformada por los siguientes docentes:

Dr. Leandro Agapito Aznarán Castillo	Presidente
M.Sc. Elmer Lluén Cumpa	Secretario
Lic. Mat. Juan Antonio Cornetero Capitán	Vocal

La tesis fue asesorada por el M. Sc Marco Antonio Peralta Lui, nombrado por Resolución N° 569 – 2019-D/FACFyM de fecha 09 de Mayo de 2019. El Acto de Sustentación fue autorizado por Resolución N° 015 – 2021 – Virtual - D/FACFyM de fecha 14 de Enero de 2021.

La Tesis fue presentada y sustentada por la Bachiller: Maria Jacqueline Atoche Bravo, y tuvo una duración de 40 minutos.

Después de la sustentación y absueltas las preguntas y observaciones de los miembros del jurado se procedió a la calificación respectiva, otorgándole el Calificativo de 18 (Dieciocho) en la escala vigesimal, con mención Muy Bueno.

Por lo que queda apta para obtener el Título Profesional de **Licenciada en Matemáticas,** de acuerdo con la Ley Universitaria 30220 y la normatividad vigente de la Facultad de Ciencias Fisicas y Matemáticas y la Universidad Nacional Pedro Ruiz Gallo.

Siendo las 10:15 am se dio por concluido el presente acto académico, dándose conformidad al presente acto con la firma de los miembros del jurado.

Dr. Leandro Agapito Aznarán Castillo
Presidente

M.Sc. Elmer Lluén Cumpa
Secretario

Lic. Mat. Juan Antonio Cornetero Capitán
Vocal

M.Sc. Marco Antonio Martin Peralta Lui
Asesor

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Un concepto clave en el campo del reconocimiento de patrones es el de la incertidumbre. Surge tanto a través del ruido en las mediciones, como a través del tamaño finito de los conjuntos de datos. La teoría de la probabilidad proporciona un marco coherente para la cuantificación y manipulación de la incertidumbre y constituye uno de los cimientos centrales para el reconocimiento de patrones.

Cuando se combina con la teoría de la decisión, nos permite hacer predicciones óptimas teniendo en cuenta toda la información disponible, aunque esa información sea incompleta o ambigua.

El tema de investigación de este trabajo, es la segmentación de imágenes, para lo cuál hacemos uso del algoritmo *Expectation-Maximization*, para la estimación de parámetros probabilísticos en modelos de mezclas Gaussianas. La segmentación de imágenes es el proceso de dividir una imagen digital en varios objetos, o grupos de píxeles. Esto se realiza con el fin de simplificar o cambiar la representación de una imagen en otras más significativas o más fácil de analizar. El proceso de segmentar consiste en asignar una etiqueta, clase o cluster (*label*) a cada píxel de la imagen, de manera que los píxeles que tienen una misma etiqueta, tendrán características visuales similares, es decir pertenecen al mismo grupo.

El algoritmo EM es un método de clasificación no-supervisada, es un método iterativo, usado frecuentemente para estimar los valores de los parámetros en modelos probabilísticos en problemas de datos con variables no observables o datos incompletos. La

introducción de datos ocultos es una construcción artificial que favorece a la estimación del conjunto de parámetros no observables Θ.

El algoritmo EM fue originalmente introducida por Arthur Dempster, Nan Laird y Donald Rubin en 1977. El método es usado para encontrar los parámetros Θ de la función de máxima verosimilitud del modelo. La estimación de Máxima Verosimilitud es un problema de optimización difícil, pues no existe una forma cerrada para $\hat{\Theta}_{ML}$. Es por ello que se hace la introducción de datos ocultos. Ahora la función de verosimilitud es una función de los datos completos (observables y ocultos), y por tanto es mucho más fáctible encontrar los parámetros del vector Θ.

El EM es un método estadístico de *Clustering* similar al *K-means*, pero con un enfoque probabilístico. El algoritmo EM supone que todos los datos han sido generados a partir de K distribuciones de probabilidad de las cuáles no conocemos a priori sus parámetros.

Después de procesar las imágenes por el algoritmo EM, las imágenes segmentadas pueden ser utilizadas en diferentes aplicaciones, como por ejemplo, la reconstrucción de imágenes.

CAPÍTULO 1

PRELIMINARES

En este capítulo introduciremos algunos conceptos básicos de la teoría de la probabilidad (James, 1981), (Rolla, 2012).

1.1. Espacios de Probabilidad

Supongamos que vamos a realizar un experimento cuyo resultado no puede ser predicho de antemano. Sin embargo, suponga que sabemos todos los posibles resultados de tal experimento. A este conjunto de todos los posibles resultados, denotaremos por Ω, y lo llamaremos *espacio muestral*.

Definición 1. Espacio muestral.
Un espacio muestral es el conjunto no vacío Ω de todos los posibles resultados de un determinado experimento.

Ejemplo 1. Considere el experimento de lanzar una moneda equilibrada y observar la cara superior de la moneda, entonces $\Omega = \{C, S\}$ donde C es cara y S es sello; pues esos dos resultados son los únicos posibles.

Ejemplo 2. Considere el experimento de lanzar un dado y observar la cara superior del dado, entonces $\Omega = \{1, 2, 3, 4, 5, 6\}$, pues esos seis resultados son los únicos posibles.

Definición 2. Evento aleatório.
Sea Ω el espacio muestral. Todo subconjunto A de Ω ($A \subset \Omega$) será llamado evento. Un evento A al cual atribuimos una probabilidad será llamado evento aleatorio.

Observación 1.

1. Como $\emptyset, \Omega \subset \Omega$, entonces $\emptyset$ y Ω son eventos aleatórios.
   - El conjunto vacío es llamado evento imposible.
   - El comjunto Ω es llamado evento cierto.
2. Si $\omega \in \Omega$, el evento $\{\omega\}$ es llamado *elemental* (o simple).

Observación 2. Dos eventos A y B son llamados mutuamente exclusivos o incompatibles si $A \cap B = \emptyset$.

Definición 3. σ-álgebra
Una clase $\mathcal{A}$ de subconjuntos de un conjunto no vacío Ω es llamada σ-álgebra si satisface las siguientes propiedades:

(i) $\Omega \in \mathcal{A}$.

(ii) Si $A \in \mathcal{A}$, entonces $A^c \in \mathcal{A}$.

(iii) Si $A_n \in \mathcal{A}$, para $n = 1, 2, 3, \ldots$, entonces $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Observación 3. Los elementos de $\mathcal{A}$ son llamados *eventos*, o subconjuntos de Ω $\mathcal{A}$-medibles o simplemente subconjuntos medibles de Ω si no hubiera confusión con la σ-álgebra referente.

Ejemplo 3. Álgebra de Borel
$\mathcal{B}$ indica la σ-álgebra de Borel en la recta, i.e., la menor σ-álgebra que contiene todos los intervalos. Los elementos de esta σ-álgebra son los borelianos de la recta. Se puede decir que un boreliano es un conjunto que puede ser obtenido de una cantidad numerable de intervalos aplicándoles las operaciones $\cup$, $\cap$, ^{c} una cantidad numerable de veces.

Definición 4. Espacio medible
El par $(\Omega, \mathcal{A})$ es llamado espacio medible, donde Ω es un conjunto no vacío y $\mathcal{A}$ es una σ-álgebra de subconjuntos de Ω.

Definición 5. Medida de probabilidad

Sea Ω un espacio muestral y $\mathcal{A}$ una σ-álgebra para un experimento dado. Una medida de probabilidad (o simplemente probabilidad) P es una aplicación $P : \mathcal{A} \to [0, 1]$ que satisface las siguientes propiedades:

(i) $P(A) \geq 0$, $A \in \mathcal{A}$.

(ii) $P(\Omega) = 1$.

(iii) Si $A_1, A_2, \ldots \in \mathcal{A}$, y $A_i \cap A_j = \emptyset \; \forall i \neq j$ (los eventos son disjuntos dos a dos) entonces $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Axioma 1. Continuidad en el vacío

Si la sucesión $(A_n)_{n\geq 1}$, donde $A_n \in \mathcal{A} \; \forall n$, decrece al vacío, entonces $P(A_n) \to 0$ cuando $n \to \infty$.

Teorema 1. ***Propiedades de una probabilidad.***

1. *$P(\emptyset) = 0$.*
2. *Para todo $A \in \mathcal{A}$, tenemos $P(A^c) = 1 - P(A)$.*
3. *Para todo $A \in \mathcal{A}$, tenemos $0 \leq P(A) \leq 1$.*
4. *Sean A, $B \in \mathcal{A}$. Si $A \subset B$, entonces:*
   - *$P(B - A) = P(B) - P(A)$.*
   - *$P(A) \leq P(B)$.*
5. *Sean A, $B \in \mathcal{A}$. Entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*
6. *Para cualquier sucesión de eventos $A_1, A_2, \ldots, A_n \in \mathcal{A}$, $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$.*
7. *Continuidad de una probabilidad: Si $A_n \uparrow A$, entonces $P(A_n) \uparrow P(A)$. Si $A_n \downarrow A$, entonces $P(A_n) \downarrow P(A)$.*

Demostración.

1. Sea $\Omega = \Omega \cup \emptyset$, además $\Omega \cap \emptyset = \emptyset$, entonces:

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset) \\ &= P(\Omega) + P(\emptyset) \end{aligned}$$

   De ahí, $1 = 1 + P(\emptyset)$, entonces, $P(\emptyset) = 0$.

2. Sea $A \in \mathcal{A}$, $\Omega = A \cup A^c$, además $A \cap A^c = \emptyset$, por lo tanto:

$$\begin{aligned} P(\Omega) &= P(A \cup A^c) \\ &= P(A) + P(A^c) \end{aligned}$$

De ahí, $1 = P(A) + P(A^c)$, entonces, $P(A^c) = 1 - P(A)$.

3. Sea $A \in \mathcal{A}$, y de lo anterior se tiene que $P(A) \leq 1$, además de la primera propiedad de medida probabildad se cumple que $P(A) \geq 0$. Por lo tanto: $0 \leq P(A) \leq 1$.

4. Sean A, $B \in \mathcal{A}$. Si $A \subset B$, entonces $B = A \cup (B - A)$, y $A \cap (B - A) = \emptyset$, de donde:

   - $P(B) = P(A) + P(B - A)$, por lo tanto: $P(B - A) = P(B) - P(A)$.
   - De lo anterior, se tiene que: $P(A) = P(B) - P(B - A) \leq P(B)$, pues $P(B - A) \geq 0$; por lo tanto: $P(A) \leq P(B)$.

5. Sean A, $B \in \mathcal{A}$, tal que $A \cup B = (A - B) \cup (A \cap B) \cup (B - A)$, donde $(A - B)$, $(A \cap B)$ y $(B - A)$ son disjuntos dos a dos; entonces,

$$\begin{aligned} P(A \cup B) &= P(A - B) + P(A \cap B) + P(B - A) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Pues $A = (A - B) \cup (A \cap B)$, $(A - B) \cap (A \cap B) = \emptyset$ y $B = (B - A) \cup (A \cap B)$, $(B - A) \cap (A \cap B) = \emptyset$.

6. Para cualquier sucesión de eventos $A_1, A_2, \ldots, A_n \in \mathcal{A}$, se tiene que

$$\bigcup_{i=1}^{n} A_i = \underbrace{A_1}_{B_1} \cup \underbrace{(A_2 - A_1)}_{B_2} \cup \underbrace{(A_3 - (A_1 \cup A_2))}_{B_3} \cup \ldots \cup \underbrace{(A_n - (\bigcup_{i=1}^{n-1} A_i))}_{B_n}$$

donde los eventos $\{B_i\}$, $i = 1, \ldots, n$ son disjuntos dos a dos:

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= P(A_1 \cup (A_2 - A_1) \cup (A_3 - (A_1 \cup A_2)) \cup \ldots \cup (A_n - (\bigcup_{i=1}^{n-1} A_i))) \\ &= P(A_1) + P(A_2 - A_1) + P(A_3 - (A_1 \cup A_2)) + \ldots + P(A_n - (\bigcup_{i=1}^{n-1} A_i)) \\ &\leq P(A_1) + P(A_2) + \ldots + P(A_n) \end{aligned}$$

Pues $A_j - (\bigcup_{i=1}^{j-1} A_i) \subset A_j$ para $j = 2, \ldots, n$.

7. Si $A_n \downarrow A$, (i.e. $A_n \supset A_{n+1}\ \forall\ n$ y $\bigcap_{n\geq 1} A_n = A$). Entonces, $P(A_n) \geq P(A_{n+1})$, y $(A_n - A) \downarrow \emptyset$ por lo tanto $P(A_n - A) \to \emptyset$ (por la continuidad en el vacío). Además $P(A_n - A) = P(A_n) - P(A)$, pues $A \subset A_n$. Por lo que se tiene, $P(A_n) - P(A) \to 0$, entonces $P(A_n) \downarrow P(A)$.
   Si $A_n \uparrow A$, (i.e. $A_n \subset A_{n+1}\ \forall\ n$ y $\bigcup_{n\geq 1} A_n = A$). Entonces $A_n^c \downarrow A^c$, por lo tanto $P(A_n^c) \downarrow P(A^c)$, o sea, $1 - P(A_n) \uparrow 1 - P(A)$; por lo tanto $P(A_n) \uparrow P(A)$.

□

Definición 6. Espacio de probabilidad
Un espacio de probabilidad es una terna $(\Omega, \mathcal{A}, P)$, donde:

(i) Ω es un conjunto no vacío.

(ii) $\mathcal{A}$ es una σ-álgebra de subconjuntos de Ω.

(iii) P es una probabilidad definida en $\mathcal{A}$.

1.2. Probabilidad Condicional

Definición 7. Probabilidad Condicional
Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad. Si $B \in \mathcal{A}$ y $P(B) > 0$, la probabilidad condicional de A dado B es definida por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ A \in \mathcal{A} \tag{1.1}$$

donde $P(A \cap B)$ es conocida como la probabilidad conjunta.

Teorema 2. ***Regla del Producto***
Sean $A, B \in \mathcal{A}$, con $P(A) > 0$ y $P(B) > 0$, entonces:

$$\begin{aligned} P(A \cap B) &= P(B).P(A|B) \\ &= P(A).P(B|A) \end{aligned} \tag{1.2}$$

Demostración.
Por la Ecuación 1.1, $P(A \cap B) = P(B).P(A|B)$.

Además $A \cap B = B \cap A$, entonces, $P(B \cap A) = P(A).P(B|A)$.
Por lo tanto:

$$P(A \cap B) = P(B).P(A|B) = P(A).P(B|A)$$

□

Teorema 3. ***Regla del Producto generalizado***
Sean $A_1, A_2, \ldots A_n \in \mathcal{A}$*, y para todo* $n = 2, 3, \ldots$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1).P(A_2|A_1).P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \tag{1.3}$$

Demostración.

$$\begin{aligned}
P(A_1 \cap A_2 \cap \ldots \cap A_n) &= P((A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \cap A_n) \\
&= P(A_1 \cap A_2 \cap \ldots \cap A_{n-1}).P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \\
&= P(A_1 \cap A_2 \cap \ldots \cap A_{n-2} \cap A_{n-1}).P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \\
&= P(A_1 \cap A_2 \cap \ldots \cap A_{n-2}).P(A_{n-1}|A_1 \cap A_2 \cap \ldots \cap A_{n-2}). \\
&\quad .P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \\
&\vdots \\
&= P(A_1 \cap A_2 \cap A_3).P(A_4|A_1 \cap A_2 \cap A_3) \ldots P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \\
&= P(A_1 \cap A_2)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \\
&= P(A_1).P(A_2|A_1).P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1})
\end{aligned}$$

□

Definición 8. Sea $(\Omega, \mathcal{A})$ un espacio medible. Una partición de Ω es una familia de conjuntos $A_1, A_2, \ldots$, tales que:

(i) $A_i \in \mathcal{A}$ para todo i.

(ii) $\bigcup_i A_i = \Omega$.

(iii) $A_i \cap A_j = \emptyset$, para todo $i \neq j$.

Es decir, los conjuntos $A_1, A_2, \ldots$ son disjuntos dos a dos y su unión es el conjunto Ω.

Vamos a admitir que la sucesión $A_1, A_2, \ldots$ sea finita o enumerable.

Observación 4. Para todo evento $B \in \mathcal{A}$ tenemos:

$$B = \bigcup_i (A_i \cap B)$$

Teorema 4. ***Teorema de la Probabilidad Total***

Sea $A_1, A_2, \ldots$ una partición de $(\Omega, \mathcal{A})$. Para todo $B \in \mathcal{A}$ se cumple que:

$$P(B) = \sum_i P(A_i).P(B|A_i) \tag{1.4}$$

Demostración.

Por la Observación 4, sea $B = \bigcup_i (A_i \cap B)$ donde $C_i = A_i \cap B$ son disjuntos, pues los A_i lo son. Entonces:

$$\begin{aligned} P(B) &= P\left(\bigcup_i (A_i \cap B)\right) \\ &= \sum_i P(A_i \cap B) \\ &= \sum_i P(A_i).P(B|A_i) \end{aligned}$$

□

Teorema 5. ***Fórmula de Bayes***

Si la sucesión de eventos aleatorios $A_1, A_2, \ldots$ forman una partición de Ω, entonces:

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{\sum_j P(A_j)P(B|A_j)} \tag{1.5}$$

Demostración.

Por la Equación 1.1 tenemos:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$$

Por la Equación 1.2 y la Equación 1.4 tenemos:

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$$

□

Observación 5. A la fórmula de Bayes a veces se le llama fórmula de probabilidades "posteriores". En efecto, las probabilidades $P(A_i)$ pueden ser llamadas probabilidades "a priori", y las $P(A_i|B)$ probabilidades "a posteriori". Esta fórmula de Bayes es útil cuando conocemos las probabilidades de los A_i, y las probabilidades condicionales de B dado A_i, pero no conocemos directamente la probabilidad de B.

1.3. Independencia

Definición 9. Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad. Los eventos aleatorios A y B son independientes si

$$P(A \cap B) = P(A).P(B) \tag{1.6}$$

Definición 10. Los eventos aleatorios A_i, $i \in I$ (I un conjunto de índices) son independientes 2 a 2 (o de a pares) si:

$$P(A_i \cap A_j) = P(A_i).P(A_j) \ \forall \ i, j \in I, \ i \neq j$$

Definición 11.

(a) Los eventos $A_1, \ldots, A_n$, $(n \geq 2)$ son llamados independientes (colectivamente) si:

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1}).P(A_{i_2}) \ldots P(A_{i_m})$$

$\forall \ 1 \leq i_1 < i_2 < \ldots < i_m \leq n$, $\forall m = 2, 3, \ldots, n$ (i.e., si todas las combinaciones satisfacen la propiedad de independencia).

(b) Los eventos $A_1, A_2, \ldots$, son independientes si $\forall \ n \geq 2$, $A_1, A_2, \ldots, A_n$ son independientes.

(c) Los eventos A_i, $i \in I$ (donde I es un conjunto de índices tal que $\#I \geq 2$, donde $\#I$ denota el número de elementos de I) son independientes si toda subfamilia finita de ellos es una familia de eventos independientes, i.e., si $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$ son independientes para toda combinación $\{i_1, \ldots, i_m\}$ de elementos de I, $\forall m = 2, 3, \ldots$

Observación 6.

- A los eventos de la definición 10, a veces son llamados estadísticamente o mutuamente independientes.
- En el item (c) de la definición anterior, vemos que toda subfamilia de una familia de eventos independientes es una familia de eventos independientes.

1.4. Variables aleatorias

Cuando realizamos un experimento aleatorio, muchas veces estamos interesados en una o más cantidades, las cuáles son dadas en función del resultado del experimento. A esas cantidades se les dá el nombre de Variables Aleatorias. Informalmente, una variable aleatoria es una función que asocia una característica numérica a cada resultado del experimento dado.

Ejemplo 4. Escoger 10 cartas de la baraja y contar cuántas de esas cartas son de corazones.

Ejemplo 5. Lanzar un dado y observar la cara superior. En este caso, tenemos $\Omega = \{1, 2, 3, 4, 5, 6\}$ y:

$$X(\omega) = \omega, \text{ donde } \omega \in \Omega$$

Ejemplo 6. Lanzar una moneda n veces y observar la sucesión de caras (c) y sellos (s) obtenidos. Los posibles resultados aquí obtenidos, son sucesiones de tamaño n de caras y sellos, y podemos definir:

$$\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i = c \text{ ó } s; i = 1. \ldots, n\}$$

El número de caras observadas en los n lanzamientos es una característica numérica de la sucesión de caras y sellos. En efecto, si definimos $X =$ número de caras observadas, vemos que el valor de X depende del resultado del experimento y podemos definir:

$$\begin{aligned} X(\omega) &= \text{número de caras en } \omega = (\omega_1, \ldots, \omega_n) \\ &= \#\{i : \omega_i = c, 1 \leq i \leq n\} \end{aligned}$$

Definición 12. Variable Aleatoria
Una variable aleatoria X en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ es una función real definida en el espacio Ω tal que el conjunto $\{\omega \in \Omega : X(w) \leq x\} = [X \leq x]$ es un evento aleatorio para todo $x \in \mathbb{R}$.
Es decir:

$$X : \Omega \to \mathbb{R}$$

es una variable aleatoria si $[X \leq x] \in \mathcal{A}$ para todo $x \in \mathbb{R}$.

Observación 7. Pero no toda función $X : \Omega \to \mathbb{R}$ es una variable aleatoria. Para que X sea una variable aleatoria, se necesita garantizar que todo evento relacionado a la variable aleatoria pueda ser medible.

Definición 13. Función de Distribución
La función de distribución (acumulada) de la variable aleatoria X, representada por F_X, o simplemente por F cuando no hubiera confusión, es definida por:

$$F_X(x) = P(X \leq x),\ x \in \mathbb{R} \tag{1.7}$$

Proposición 1. ***Propiedades de la Función de Distribución***
Si X es una variable aleatoria, su función de distribución F satisface las siguientes condiciones:

1. *Si $x_1 \leq x_2$ entonces $F(x_1) \leq F(x_2)$; i.e. F es no-decreciente.*
2. *Si $x_n \downarrow y$ entonces $F(x_n) \downarrow F(y)$; i.e. F es continua por derecha.*
3. *Si $x_n \downarrow -\infty$, entonces $F(x_n) \downarrow 0$. Si $x_n \uparrow \infty$, entonces $F(x_n) \uparrow 1$ (también se puede escribir como $F(-\infty) = 0$ y $F(\infty) = 1$).*

Demostración.

1. $x_1 \leq x_2 \Rightarrow [X \leq x_1] \subset [X \leq x_2]$. Entonces

$$F(x_1) = P(X \leq x_1) \leq P(X \leq x_2) = F(x_2)$$

2. Si $x_n \downarrow y$, entonces $[X \leq x_n]$ es una sucesión decreciente de eventos aleatorios y $\bigcap_{n\geq 1} [X \leq x_n] = [X \leq y]$ (pues $X \leq y$ si, y solamente si, $X \leq x_n\ \forall n$). En otras palabras, $[X \leq x_n] \downarrow [X \leq y]$, y por la continuidad de toda probabilidad, se tiene que $F(x_n) = P(X \leq x_n) \downarrow P(X \leq y) = F(y)$.

3. Si $x_n \downarrow -\infty$, la sucesión de eventos $[X \leq x_n]$ es una sucesión decreciente y además $\bigcap_{n=1}^{\infty} [X \leq x_n] = \emptyset$. Por lo tanto $[X \leq x_n] \downarrow \emptyset$ entonces $F(x_n) = P(X \leq x_n) \downarrow 0$.
Si $x_n \uparrow +\infty$, la sucesión de eventos $[X \leq x_n]$ es una sucesión creciente y además $\bigcup_{n=1}^{\infty} [X \leq x_n] = \Omega$. Por lo tanto $[X \leq x_n] \uparrow \Omega$ entonces $F(x_n) = P(X \leq x_n) \uparrow 1$.

□

1.4.1. Tipos de Variables Aleatorias

Definición 14. Variable Aleatoria Discreta

Una variable aleatoria X (así como su función de distribución F_X) es llamada variable aleatoria discreta si toma un número finito o numerable de valores, i.e., si existe un conjunto finito o numerable $\{x_1, x_2, x_3, \ldots\} \subseteq \mathbb{R}$ tal que $X(w) \in \{x_1, x_2, x_3, \ldots\}\ \forall \omega \in \Omega$. En este caso, la función $p(x_n)$, $n = 1, 2, \ldots$, se llama función de probabilidad (o función de frecuencia):

$$p_X(x_n) = P(X = x_n)$$

La función de distribución de una variable aleatoria discreta es dada por:

$$F_X(x) = \sum_{n:x_n \leq x} P(X = x_n) = \sum_{n:x_n \leq x} p_X(x_n)$$

Ejemplo 7. Sea X una variable aleatoria discreta con función de probabilidad dada por:

x_i	3	5	7	9	11	13
p_i	2/15	4/15	1/15	2/15	3/15	3/15

La gráfica correspondiente a la función de probabilidad es:

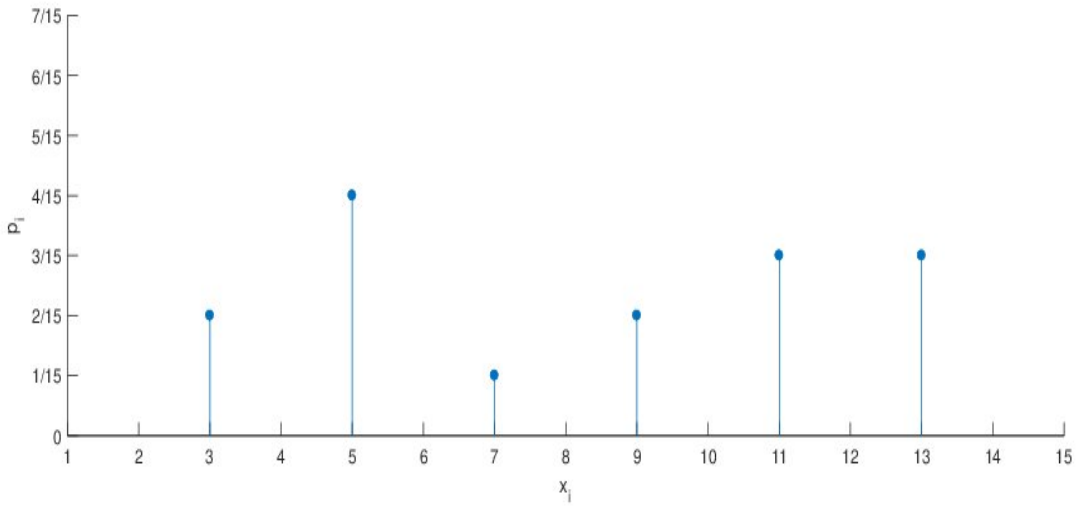


Figura 1.1: Función de probabilidad de la variable X discreta

Y su correspondiente función de distribución es:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < 3 \\ 2/15 & \text{si } 3 \leq x < 5 \\ 6/15 & \text{si } 5 \leq x < 7 \\ 7/15 & \text{si } 7 \leq x < 9 \\ 9/15 & \text{si } 9 \leq x < 11 \\ 12/15 & \text{si } 11 \leq x < 13 \\ 1 & \text{si } x \geq 13 \end{cases}$$

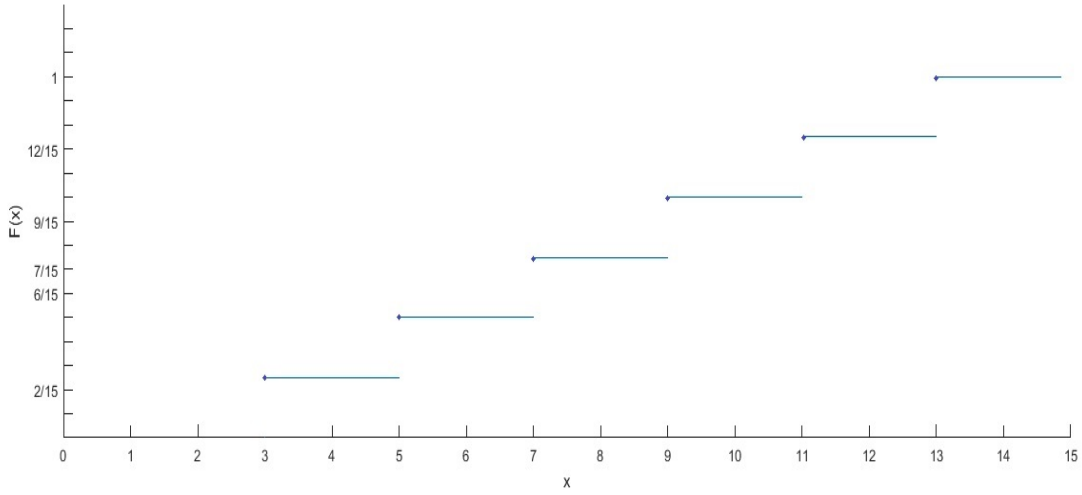


Figura 1.2: Función de distribución de la variable X discreta

Definición 15. Variable Aleatoria Continua

Una variable aleatoria X (así como su función de distribución F_X) es llamada variable aleatoria absolutamente continua si existe $f_X(x) \geq 0$ tal que:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \ \forall x \in \mathbb{R}$$

En este caso, decimos que f_X es la función de densidad de probabilidad de X, o simplemente la densidad de X.

Observación 8. Por el Teorema Fundamental del Cálculo, tenemos que:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Entonces $dF_X(x) = f_X(x)dx$.

Ejemplo 8. Sea X una variable aleatoria continua con función de distribución dada por:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

La gráfica correspondiente a la función de distribución es:

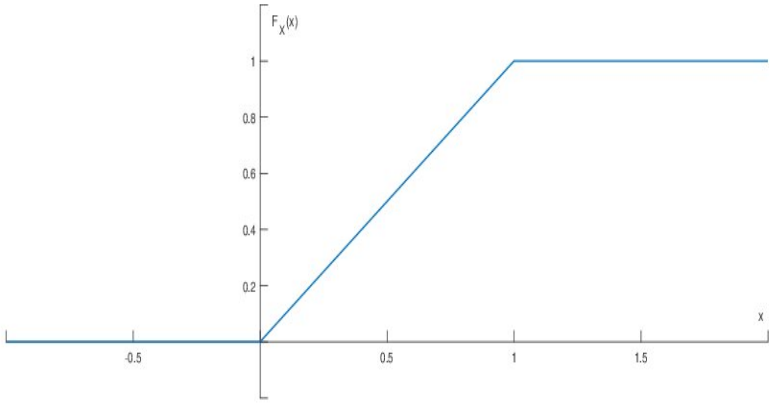


Figura 1.3: Función de distribución de la variable X continua

La densidad de X está dada por:

$$f(x) = F'_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & x < 0 \quad \text{ó} \quad x > 1 \end{cases}$$

El valor de f en los puntos 0 y 1 es arbitrario, pues la integral $\int_{-\infty}^{x} f(t)dt$ sigue siendo igual a $F_X(x)$. Se puede definir como $f(0) = f(1) = 0$ ó $f(0) = f(1) = 1$.

Proposición 2. *Si X es una variable aleatoria en $(\Omega, \mathcal{A}, P)$, entonces el evento*

$$[X \in B] = \{\omega \in \Omega : X(\omega) \in B\}$$

es un evento aleatorio para todo boreliao B, i.e.,

$$[X \in B] \in \mathcal{A}, \ \forall B \in \mathcal{B} \ \ (\textit{donde } \mathcal{B} \textit{ es la } \sigma\textit{-álgebra de Borel})$$

Demostración. Como $\mathcal{B}$ es la σ-álgebra de los borelianos (la menor σ-álgebra que contiene a los intervalos):

(i) Si $B = (-\infty, b]$, entonces $[X \in B] \in \mathcal{A}$ por definición.

(ii) Si $B = (a, \infty)$, entonces $B = (-\infty, a]^c$ y $[X \in B] = [X \leq a]^c \in \mathcal{A}$.

(iii) Si $B = (a, b]$, entonces $[X \in B] = [a < X \leq b] = [X \leq b] - [X \leq a] \in \mathcal{A}$.

(iv) Si $B = (a, b)$, entonces

$$B = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right]$$

y

$$[X \in B] = \bigcup_{n=1}^{\infty} \left(a < X \leq b - \frac{1}{n}\right] \in \mathcal{A}$$

De forma análoga, se verifica que para todo intervalo B, $[X \in B] \in \mathcal{A}$. Además vale para $B = \bigcup_{i=1}^{n} B_i$, donde los B_i son intervalos disjuntos, ya que $[X \in B] = \bigcup_{i=1}^{n} [X \in B_i]$.

Definición 16. La probabilidad P_X, definida en la σ-álgebra de Borel por $P_X(B) = P(X \in B)$, se llama la distribución de X.

1.5. Vectores Aleatorios

Definición 17. Un vector $\widetilde{X} = (X_1, \ldots, X_n)$, donde X_i es una variable aleatoria para todo i definida en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, P)$, es llamado vector aleatorio si:

$$\widetilde{X}^{-1}(B) \in \mathcal{A} \text{ para todo } B \in \mathcal{B}^n$$

Definición 18. La función de distribución conjunta $F = F_{\widetilde{X}} = F_{X_1,\ldots,X_n}$ de un vector aleatorio $\widetilde{X}$ es definido por:

$$F_{\widetilde{X}}(\widetilde{x}) = F_{\widetilde{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n), \ \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

F también se llama función de distribución conjunta de las variables aleatorias $X_1, \ldots, X_n$.

Observación 9. El evento $[X_1 \leq x_1, \ldots, X_n \leq x_n] = \bigcap_{i=1}^{n} [X_i \leq x_i]$ es aleatorio, ya que las X_i son variables aleatorias y por lo tanto $[X_i \leq x_i] \in \mathcal{A}\ \forall i$. El vector aleatorio $\widetilde{X}$ es una función definida en el espacio muestral Ω a valores en $\mathbb{R}^n$, i.e., $\widetilde{X} : \Omega \to \mathbb{R}^n$.

Proposición 3. ***Propiedades de la Función de Distribución*** F ***de un vector aleatorio*** $(X_1, \ldots, X_n)$
Si X es un vector aleatorio, su función de distribución F satisface las siguientes distribuciones:

(F1) $F(x_1, \ldots, x_n)$ es no decreciente en cada una de las variables.

(F2) $F(x_1, \ldots, x_n)$ es continua por derecha en cada una de las variables.

(F3) Para todo i,

$$\lim_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0.$$

También,

$$\lim_{\forall i, x_i \to +\infty} F(x_1, \ldots, x_n) = 1.$$

Observación 10. Para $I = (a, b]$ y $g : \mathbb{R}^k \to \mathbb{R}$ definamos:

$$\triangle_{k,I} g(x_1, \ldots, x_k) = g(x_1, \ldots, x_{k-1}, b) - g(x_1, \ldots, x_k) = g(x_1, \ldots, x_{k-1}, a)$$

Como F es la función de distribución del vector aleatorio (X, Y): si $I_1 = (a_1, b_1]$ y $I_2 = (a_2, b_2]$, entonces

$$\begin{aligned} \triangle_{1,I_1} \triangle_{2,I_2} F(x, y) &= \triangle_{1,I_1} [F(x, b_2) - F(x, a_2)] \\ &= F(b_1, b_2) - F(b_1, a_2) - [F(a_1, b_2) - F(a_1, a_2)] \geq 0 \end{aligned}$$

Para n general, la propiedad será:

(F4) $\triangle_{1,I_1} \ldots \triangle_{n,I_n} F(x_1, \ldots, x_n) \geq 0$, $\forall I_k = (a_k, b_k], a_k < b_k, k = 1, \ldots, n$.

Definición 19. Una función $F : \mathbb{R}^n \to \mathbb{R}$ que satisface las propiedades F1, F2, F3 y F4 se llama *función de distribución n-dimensional* (o de n variables).

Definición 20.

(a) Si el vector aleatorio $(X_1, \ldots, X_n)$ toma solamente un número finito o numerable de valores, se llama discreto.

(b) Sea el vector aleatorio $(X_1, \ldots, X_n)$ y su función de distribución F. Si existe una función $f(x_1, \ldots, x_n) \geq 0$ tal que

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \ldots dt_n, \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

entonces f se llama densidad del vector aleatorio $(X_1, \ldots, X_n)$, o densidad conjunta de las variables aleatorias $X_1, \ldots, X_n$, y en este caso, decimos que $(X_1, \ldots, X_n)$ es continuo.

Observación 11. La σ-álgebra de Borel en $\mathbb{R}^n$ ($\mathcal{B}^n$) es la menor σ-álgebra que contiene a todo rectángulo n-dimensional, o sea, la σ-álgebra generada por los rectángulos.

Definición 21. La probabilidad definida en $\mathcal{B}^n$ por $P(\widetilde{X} \in B)$ se llama distribución de $\widetilde{X}$ o distribución conjunta de $X_1, \ldots, X_n$.
Notación: $P_{\widetilde{X}}(B) = P(\widetilde{X} \in B)$, $P_{\widetilde{X}}$ es la distribución de $\widetilde{X}$.

Proposición 4.

(a) Si el vector aleatorio $\widetilde{X}$ es discreto, entonces:

$$P_{\widetilde{X}}(B) = \sum_{i, \widetilde{x}_i \in B} P(\widetilde{X} = \widetilde{x}_i) \ \forall B \in \mathcal{B}^n$$

(b) Si $\widetilde{X}$ es continuo con densidad $f(x_!, \ldots, x_n)$, entonces

$$P_{\widetilde{X}}(B) = P(\widetilde{X} \in B) = \int_B \ldots \int f(x_1, \ldots, x_n) dx_1 \ldots x_n.$$

1.5.1. Independencia

Definición 22. Se dice que las variables aleatorias $X_1, X_2, \ldots, X_n$ definidas en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, son colectivamente independientes, o simplemente independientes, si:

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i)$$

para todo $B_i \in \mathcal{A}$, $i = 1, 2, \ldots, n$.

Proposición 5. *Si $X_1, X_2, \ldots, X_n$ son independientes, entonces*

$$F_{X_1,X_2,\ldots,X_n}(x_1 \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i), \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

Demostración. Supongamos que $X_1, X_2, \ldots, X_n$ son independientes, entonces:

$$\begin{aligned} F_{X_1,X_2,\ldots,X_n}(x_1 \ldots, x_n) &= P(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= P(X_1 \in (-\infty, x_1], \ldots, X_n \in (-\infty, x_1]) \\ &= \prod_{i=1}^{n} P(X_i \in (-\infty, x_i]) \\ &= \prod_{i=1}^{n} P(X_i \leq x_i) \\ &= \prod_{i=1}^{n} F_{X_i}(x_i), \forall (x_1, \ldots, x_n) \end{aligned}$$

Proposición 6. *Si $X_1, X_2, \ldots, X_n$ son independientes y poseen densidades $f_{X_1}, \ldots, f_{X_n}$, entonces la función*

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), (x_1, \ldots, x_n) \in \mathbb{R}^n$$

es la densidad conjunta de las variables aleatorias $X_1, X_2, \ldots, X_n$, i.e., $f = f_{X_1,X_2,\ldots,X_n}$.

Demostración. Si $X_1, X_2, \ldots, X_n$ son independientes, entonces

$$\begin{aligned} F_{X_1\ldots,X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} F_{X_i}(x_i) \\ &= \prod_{i=1}^{n} \int_{-\infty}^{x_i} f_{X_i}(t_i) dt_i \\ &= \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f_{X_1}(t_1) \ldots f_{X_n}(t_n) dt_1 \ldots dt_n. \end{aligned}$$

Luego, $\prod_{i=1}^{n} f_{X_i}$ es la densidad conjunta de $X_1, X_2, \ldots, X_n$.

Definición 23.

(a) Si $F(x, y)$ es la función de distribución conjunta de X e Y, entonces la función de distribución de X es

$$F_X(x) = \lim_{y \to +\infty} F(x, y) = F(x, +\infty) = P(X \leq x) = P(X \leq x, Y \leq +\infty).$$

F_X obtenida así se llama función de distribución marginal de X.

(b) Si $f(x,y)$ es la densidad conjunta de X e Y, entonces X tiene densidad dada por

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

f_X así obtenida se llama densidad marginal de X.

Observación 12.

- **Caso discreto**: Dada una variable aleatoria (X,Y) con función masa de probabilidad conjunta denotada por $p_{ij} = P(X = x_i, Y = y_j)$, se define la función masa de probabilidad marginal de X como:

$$P_X(x_i) = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} = p_i.$$

Análogamente, la función masa de probabilidad marginal de Y es:

$$P_Y(y_i) = P(Y = y_i) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} = p_j.$$

Entonces, las funciones de distribución marginales resultan:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i, Y = y_j) = \sum_{x_i \leq x} p_i$$

$$F_Y(y) = P(Y \leq y) = \sum_{y_j \leq y} P(X = x_i, Y = y_j) = \sum_{y_i \leq y} p_j$$

- **Caso continuo**: Sea una variable aleatoria (X,Y) con función de densidad conjunta $f(x,y)$. Las funciones de densidad marginales están dadas por:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

Las correspondientes funciones de distribución marginales resultan:

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \in \mathbb{R}) = \int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(u,v)dvdu$$

$$F_Y(y) = P(Y \leq y) = P(X \in \mathbb{R}, Y \leq y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{y} f(u,v)dvdu$$

Ejemplo 9. Sea (X, Y) una variable aleatoria bidimensional discreta con función de masa de probabilidad dada por:

$X \setminus Y$	0	1	2	3	p_i
1	0	1/14	1/14	2/14	4/14
2	1/14	3/14	0	1/14	5/14
3	2/14	1/14	2/14	0	5/14
p_j	3/14	5/14	3/14	3/14	1

La función de masa de probabilidad de X es:

$$P_X(x_1) = P_X(1) = P(X = 1) = \sum_{j=1}^{4} P(X = 1, Y = y_j) = p_1 = 4/14$$

$$P_X(x_2) = P_X(2) = P(X = 2) = \sum_{j=1}^{4} P(X = 2, Y = y_j) = p_2 = 5/14$$

$$P_X(x_3) = P_X(3) = P(X = 3) = \sum_{j=1}^{4} P(X = 3, Y = y_j) = p_3 = 5/14$$

La función de masa de probabilidad de Y es:

$$P_Y(y_1) = P_Y(0) = P(Y = 0) = \sum_{i=1}^{3} P(X = x_i, Y = 0) = p_1 = 3/14$$

$$P_Y(y_2) = P_Y(1) = P(Y = 1) = \sum_{i=1}^{3} P(X = x_i, Y = 1) = p_2 = 5/14$$

$$P_Y(y_3) = P_Y(2) = P(Y = 2) = \sum_{i=1}^{3} P(X = x_i, Y = 2) = p_3 = 3/14$$

$$P_Y(y_4) = P_Y(3) = P(Y = 3) = \sum_{i=1}^{3} P(X = x_i, Y = 3) = p_4 = 3/14$$

La función de distribución marginal de X es:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < 1 \\ 4/14 & \text{si } 1 \leq x < 2 \\ 9/14 & \text{si } 2 \leq x < 3 \\ 1 & \text{si } x \geq 3 \end{cases}$$

La función de distribución marginal de Y es:

$$F_Y(x) = P(Y \leq y) = \begin{cases} 0 & \text{si } y < 0 \\ 3/14 & \text{si } 0 \leq y < 1 \\ 8/14 & \text{si } 1 \leq y < 2 \\ 11/14 & \text{si } 2 \leq y < 3 \\ 1 & \text{si } y \geq 3 \end{cases}$$

Ejemplo 10. Sea (X, Y) una variable aleatoria bidimensional con función de densidad conjunta dada por:

$$f(x,y) = \begin{cases} 2x(4-xy) & \text{si } 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{caso contrario} \end{cases}$$

Las funciones de densidades marginales son:

$$f_X(x) = \int_0^1 2x(4-xy)dy = 2\left(4xy - \frac{x^2y^2}{2}\right)\Big|_0^1 = 2\left(4x - \frac{x^2}{2}\right) = 8x - x^2, \quad 0 \leq x \leq 1$$

$$f_Y(y) = \int_0^1 2x(4-xy)dx = 2\left(\frac{4x^2}{2} - \frac{x^3y}{3}\right)\Big|_0^1 = 2\left(2 - \frac{y}{3}\right) = 4 - \frac{2y}{3}, \quad 0 \leq y \leq 1$$

La función de distribución conjunta es:

$$\begin{aligned} F(x,y) &= \int_0^x \int_0^y 2u(4-uv)dvdu = 2\int_0^x \left(4uv - \tfrac{u^2v^2}{2}\right)\Big|_0^y du = 2\int_0^x \left(4uy - \tfrac{u^2y^2}{2}\right) du \\ &= 2\left(\tfrac{4u^2y}{2} - \tfrac{u^3y^2}{6}\right)\Big|_0^x = x^2y\left(4 - \tfrac{xy}{3}\right), \quad 0 \leq x \leq 1, 0 \leq y \leq 1 \end{aligned}$$

Y sus funciones de distribución marginales son:

$$F_X(x) = \int_{-\infty}^x \int_{-\infty}^{+\infty} f(u,v)dvdu = \int_{-\infty}^x f_X(u)du = \int_0^x (8u - u^2)du = 4x^2 - \frac{x^3}{3} = F(x,1), \quad 0 \leq x \leq 1$$

$$F_Y(y) = \int_{-\infty}^{+\infty} \int_{-\infty}^y f(u,v)dvdu = \int_{-\infty}^y f_Y(v)dv = \int_0^y \left(4 - \frac{2v}{3}\right) du = 4y - \frac{y^2}{3} = F(1,y), \quad 0 \leq y \leq 1$$

1.6. Esperanza Matemática

Definición 24. Sea X una variable aleatoria discreta con función de probabilidad $p(x_i)$. La esperanza matemática para el caso discreto, se define por:

$$EX = \sum_i x_i p(x_i) = \sum_i x_i P(X = x_i) \tag{1.8}$$

La esperanza de X se llama también *media de* X, o *valor esperado de* X. En efecto, EX es un promedio ponderado, donde los pesos son las probabilidades $p(x_i)$. i.e., EX es un promedio de los valores posibles de X, ponderada conforme a la distribución de X.

Definición 25. Sea X una variable aleatoria cualquiera y F su función de distribución. La esperanza de X está definida por:

$$EX = \int_{-\infty}^{\infty} x dF(x) \tag{1.9}$$

Observación 13.

- Se usan varias integrales para representar la esperanza, siendo las más comunes:

$$EX = \int x dF_X(x) = \int x P_X(dx) = \int X dP$$

- Si X tiene densidad $f(x)$, entonces:

$$EX = \int x dF_X(x) = \int x f(x) dx$$

Propiedades de la esperanza

1. Si $X = c$ (i.e., $X(\omega) = c$, $\forall \omega \in \Omega$) entonces $EX = c$.
2. Si $X \leq Y$ entonces $EX \leq EY$, si las esperanzas están bien definidas.
3. Si EX está bien definida,entonces $E(aX + b) = aEX + b$ para todo $a, b \in \mathbb{R}$ (convención: $0 \cdot \infty = 0$).
4. $E(aX + bY) = aEX + bEY$, cuando el término de la derecha de la igualdad tiene sentido.

Definición 26. Momentos
Sea X una variable aleatoria. El valor de $E(X-b)^k$, si existe, es llamado k-ésimo momento de X alrededor de b, para $b \in \mathbb{R}$, $k = 1, 2, 3, \ldots$.
El k-ésimo momento alrededor de cero, EX^k, es llamado simplemente k-ésimo momento de X o momento de orden k de X.
Si X es integrable, entonces el k-ésimo momento alrededor de la media, $E(X-EX)^k$, es llamado k-ésimo momento central de X.

Observación 14. El primer momento es la esperanza y el primer momento central es nulo: $E(X-EX)=0$.

Definición 27. Varianza
La varianza de X es el segundo momento central:

$$VarX = E(X-EX)^2 = EX^2 - (EX)^2 \tag{1.10}$$

Notación:
$VarX = V(X) = \sigma_X^2 = \sigma^2(X)$.
$\sigma_X = \sqrt{VarX}$ es la desviación estándar de X.

Definición 28. Covarianza
La covarianza entre X y Y está definida por:

$$Cov(X,Y) = E[(X-EX)(Y-EY)] = E[XY] - EX.EY \tag{1.11}$$

Observación 15.

1. Si $Cov(X,Y)=0$, se dice que X e Y son no correlacionadas.
2. $Cov(X,Y) = Cov(Y,X)$. Es decir la matriz de covarianza, es una matriz simétrica.

Observación 16. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias l-dimensional, donde $X_i = (x_{i1}, x_{i2}, \ldots, x_{il})^T$, $i = 1, 2, \ldots, n$, con media $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$ y matriz de covarianza Σ de dimensión $l \times l$ es una matriz definida positiva.

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$$

Demostración. Una matriz es definida positiva si, para todo vector $y \in \mathbb{R}^l$, se cumple que $y^T \Sigma y > 0$.
En efecto, sea w cualquier vector de dimensión l, vamos a definir la variable unidimensional (Martínez, 2006):

$$v_i = w^T(X_i - \mu)$$

donde μ es el vector l-dimensional de la media, $i = 1, 2, \ldots, n$.
La media de los valores de v_i es:

$$\begin{aligned}
\bar{v} &= \tfrac{1}{n}\sum_{i=1}^{n} v_i \\
&= \tfrac{1}{n} w^T \sum_{i=1}^{n}(X_i - \mu) \\
&= \tfrac{1}{n} w^T[(X_1 - \mu) + (X_2 - \mu) + \ldots + (X_n - \mu)] \\
&= \tfrac{1}{n} w^T[(X_1 + X_2 + \ldots, X_n) - n\mu] \\
&= \tfrac{1}{n} w^T[n\mu - n\mu] \\
&= 0
\end{aligned}$$

Ahora calculemos la varianza de v:

$$\begin{aligned}
Var(v) &= E[v^2] - (Ev)^2 \\
&= E[v^2] - \bar{v}^2 \\
&= E[v^2] \\
&= \tfrac{1}{n}\sum_{i=1}^{n} v_i^2 \\
&= \tfrac{1}{n}\sum_{i=1}^{n}[w^T(X_i - \mu)][(X_i - \mu)^T w] > 0 \\
&= w^T \Sigma w > 0
\end{aligned}$$

Por lo que w es cualquier vector, entonces Σ es definida positiva.

CAPÍTULO 2

RECONOCIMIENTO DE PATRONES

En este capítulo introduciremos algunos conceptos básicos del Reconocimiento de Patrones, Teoría e Inferencia Bayesiana (Bishop, 2006), (Theodoridis & Koutroumbas, 2009).

2.1. Introducción

El reconocimiento de patrones es la disciplina científica cuyo objetivo es la clasificación de *objetos* en un número de categorias o *clases*. Nos referiremos a esos objetos usando el término genérico *patrones*. El reconocimiento de patrones actúa en diferentes áreas, tales como: Visión Artificial, Reconocimiento de Carácteres, Diagnóstico asistido por computadora, Reconocimiento de voz, Reconocimiento facial, Biometría, Recuperación Minería de Datos, Bioinformática, etc.

El campo del reconocimiento de patrones está relacionado con el descubrimiento automático de regularidades en los datos, mediante el uso de algoritmos computacionales; y con el uso de estas regularidades, tomar acciones tales como: la clasificación de los datos en diferentes categorias.

Es decir, asignar objetos desconocidos (patrones) en la clase correcta, se conoce como clasificación. Las características de estos objetos, son cantidades medibles obtenidas de los patrones, y la tarea de clasificación se basa en estos respectivos valores. Un vector de características viene a ser una serie de características $x_1, x_2, \ldots, x_l$. Es decir:

$$\mathbf{x} = [x_1, x_2, \ldots, x_l]^T \in \mathbb{R}^l$$

Los vectores de características son tratados como vectores aleatorios.

El clasificador consiste en un conjunto de funciones, cuyos valores son calculados en $\mathbf{x}$, que determinan la clase a la cuál los correspondientes patrones pertenecen.

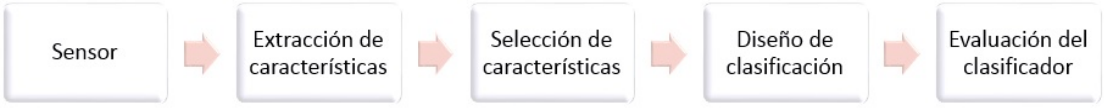


Figura 2.1: Etapas para un Sistema de clasificación

La Figura 2.1 muestra las diversas etapas seguidas para el diseño de un sistema de clasificación.
Hay dos formas para el reconocimiento de patrones:

- Supervisado: Se usa patrones cuyas clases son conocidas a-priori y son usados para la fase de entrenamiento.
- No-supervisado: El número de clases en general, es desconocido y no hay patrones de entrenamiento.

2.2. Clasificadores basados en la Teoría de Decisión de Bayes

Sea el vector de características:

$$\mathbf{x} = [x_1, x_2, \ldots, x_l]^T$$

El método de clasificación consiste en asignar un objeto, el cuál está representado por el vector de características, a la clase disponible más probable: $\omega_1, \omega_2, \ldots, \omega_M$. Es decir, $\mathbf{x}$ es asignado a la clase ω_i, si $P(\omega_i|\mathbf{x})$ es máxima.

Primero, consideremos el caso para dos clases. Sean ω_1, ω_2 dos clases a las que pertenecen los patrones. Asumamos, que las *a priori probabilities* $P(\omega_1)$ y $P(\omega_2)$ son conocidas. Otra cantidad estadística que vamos a asumir que conocemos son las *class-conditional probability density function* $p(\mathbf{x}|\omega_i)$, $i = 1, 2$, las cuáles describen la distribución de los vectores característicos en cada una de las clases. La pdf $p(\mathbf{x}|\omega_i)$ es llamada la función de verosimilitud (en inglés *likelihood function*) de ω_i con respecto a $\mathbf{x}$. Si los vectores característicos sólo toman valores discretos, la función de densidad $p(\mathbf{x}|\omega_i)$ se convierte en probabilidad y la denotaremos por $P(\mathbf{x}|\omega_i)$.

De la regla de Bayes, se tiene que:

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{2.1}$$

donde $p(x)$ es la pdf de x:

$$p(\mathbf{x}) = \sum_{i=1}^{2} p(\mathbf{x}|\omega_i)P(\omega_i) \tag{2.2}$$

La regla de clasificación de Bayes puede ahora expresarse como:

- Si $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, $\mathbf{x}$ es clasificado a ω_1.
- Si $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$, $\mathbf{x}$ es clasificado a ω_2.

Usando la Ecuación 2.1 la regla de decisión puede basarse en las inecuaciones:

$$p(\mathbf{x}|\omega_1)P(\omega_1) \gtrless p(\mathbf{x}|\omega_2)P(\omega_2) \tag{2.3}$$

$p(\mathbf{x})$ no es tomado en cuenta, porque es la suma de todas las clases y no afecta en la decisión.

Sea x_0 el umbral que particiona el espacio de características en dos regiones R_1 y R_2. Por la regla de decisión de Bayes, para todos los valores de $\mathbf{x}$ en R_1, el clasificador decide que

$\mathbf{x}$ es clasificado a ω_1; y para los valores en R_2, el clasificador decide que $\mathbf{x}$ es clasificado a ω_2.

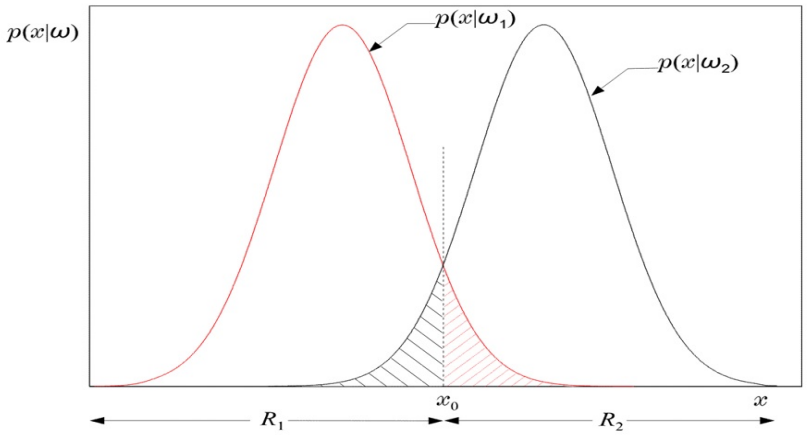


Figura 2.2: Ejemplo de dos regiones R_1 y R_2 formadas por el clasificador Bayesiano para el caso de dos clases equiprobables (Theodoridis & Koutroumbas, 2009)

De la Figura 2.2, se observa que la probabilidad del error es el área sombreada total. En efecto, existe una probabilidad finita para un $\mathbf{x}$ que se encuentra en la región R_2 y al mismo tiempo pertenezca a la clase ω_1. Entonces nuestra decisión es errónea. Lo mismo ocurre, si para los puntos que se originan en la clase ω_2, se encuentren en la región R_1. Por lo tanto, la probabilidad total, P_e, de cometer un error de decisión para el caso de dos clases equiprobables ($P(\omega_1) = P(\omega_2)$), está dada por

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(\mathbf{x}|\omega_2)d\mathbf{x} + \frac{1}{2}\int_{x_0}^{+\infty} p(\mathbf{x}|\omega_1)d\mathbf{x}$$

El clasificador Bayesiano es óptimo cuando la probabilidad del error de clasificación es minimizado.

Los errores de decisión son inevitables. Un error ocurre si $\mathbf{x} \in R_1$ a pesar de pertenecer a ω_2; o si $\mathbf{x} \in R_2$ a pesar de pertenecer a ω_1. El error de decisión para dos clases, es dada por:

$$P_e = P(\mathbf{x} \in R_2, \omega_1) + P(\mathbf{x} \in R_1, \omega_2) \tag{2.4}$$

donde $P(\cdot,\cdot)$ es la probabilidad conjunta de dos eventos.

$$\begin{aligned} P_e &= P(\mathbf{x} \in R_2|\omega_1)P(\omega_1) + P(\mathbf{x} \in R_1|\omega_2)P(\omega_2) \\ &= P(\omega_1) \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} + P(\omega_2) \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} \end{aligned} \tag{2.5}$$

O usando la regla de Bayes:

$$P_e = \int_{R_2} P(\omega_1|\mathbf{x})p(\mathbf{x})d\mathbf{x} + \int_{R_1} P(\omega_2|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{2.6}$$

Ahora consideremos, el caso general, sea $\underline{x} = [x_1, x_2, \ldots, x_l]^T$ el vector de características $l-$dimensional. El objetivo es clasificar este patrón con respecto a las M clases disponibles $\omega_1, \omega_2, \ldots, \omega_M$. El vector de características $\underline{x}$ es clasificado a la clase ω_i, si:

$$P(\omega_i|\underline{x}) > P(\omega_j|\underline{x}), \quad \forall j \neq i \tag{2.7}$$

2.2.1. Funciones discriminantes y Superficies de Decisión

Ahora está claro que minimizar la probabilidad del error es equivalente a particionar el espacio de características en 2 regiones, para el caso de 2 clases. Sea R_1 y R_2 regiones contiguas, entonces tales regiones son separadas por una superficie de decisión en el espacio de características bidimensional. Para el caso de la probabilidad del error mínimo, este es descrito por:

$$P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) = 0 \tag{2.8}$$

Desde un lado de la superficie, esta diferencia es positiva, y del otro lado es negativa. A veces, en lugar de trabajar directamente con probabilidades, puede ser más conveniente trabajar con una función equivalente de ellas, por ejemplo, $g_i(\mathbf{x}) \equiv f(P(\omega_i|\mathbf{x})), i = 1, 2$, donde $f(\cdot)$ es una función monótonamente creciente. $g_i(\mathbf{x})$ es conocida como una *función discriminante*.

La regla de decisión es ahora considerada como:

$$\mathbf{x} \text{ es clasificado a } \omega_i \text{ si } g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i \tag{2.9}$$

Las superficies de desición, que separan regiones contiguas, son descritas por:

$$g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0, \quad i, j = 1, 2 \quad i \neq j \tag{2.10}$$

2.2.2. Clasificación Bayesiana para distribuciones normales

Definición 29. La función de densidad de probabilidad Gausiana
La función de densidad de probabilidad Gausiana 1-dimensional o univariable, es definida por:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.11}$$

Los parámetros μ y σ^2 denotan la media y la varianza respectivamente. El valor medio de la variable aleatoria x es igual a μ, dado por:

$$\mu = E[x] = \int_{-\infty}^{+\infty} xp(x)dx \tag{2.12}$$

donde $E[\cdot]$ denota la media (o esperanza) de una variable aleatoria. El parámetro σ^2 es definido por:

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx \tag{2.13}$$

La función de densidad de probabilidad Gausiana multivariable en el espacio l-dimensional, es dada por:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{2.14}$$

donde $|\Sigma|$ denota el determinante de Σ, $\mu = E[\mathbf{x}] = E[(x_1, x_2, \ldots, x_l)^T] = (\mu_1, \mu_2, \ldots, \mu_l)^T$ es el valor medio y Σ es la matriz de covarianza de dimensión $l \times l$, definida como:

$$\Sigma = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1l} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{l1} & \sigma_{l2} & \ldots & \sigma_l^2 \end{bmatrix} \tag{2.15}$$

donde $\sigma_i^2 = E[(x_i - \mu_i)^2]$ es la varianza de x_i, $\sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)]$ es la covarianza entre las variables aleatorias x_i y x_j. Por lo tanto, la diagonal principal de la matriz Σ consiste de las varianzas respectivas de los elementos del vector aleatorio, y los elementos fuera de la diagonal principal son las covarianzas respectivas entre los elementos del vector aleatorio.

Observación 17.

- Es claro que si $l = 1$, la Gausiana multivariable coincide con la Gausiana univariable.
- $\mathcal{N}(\mu, \Sigma)$ denota la pdf Gaussiana con media μ y covarianza Σ.
- Para el caso de $l = 2$, sea $\mathbf{x} = (x_1, x_2)^T$, $\mu = (\mu_1, \mu_2)^T$, entonces:

$$\begin{aligned} \Sigma &= E\left[\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} [x_1 - \mu_1, x_2 - \mu_2]\right] \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix} \end{aligned}$$

  donde $E[x_i] = \mu_i$, para $i = 1, 2$ y además $\sigma_{12} = E[(x_1 - \mu_1)(x_2 - \mu_2)]$ es la covarianza entre las variables aleatorias x_1 y x_2, la cuál mide la correlación estadística entre las variables. Si las variables son independientes estadísticamente su covarianza es cero. Además, los elementos de la diagonal principal de Σ, son las varianzas de los elementos respectivos del vector aleatorio.

- Si las variables aleatorias x_i son independientes estadísticamente, entonces la media del producto es igual al producto de las medias, es decir, $E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_i - \mu_i)]E[(x_j - \mu_j)] = 0$. En este caso, la matriz de covarianza es diagonal.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_l^2 \end{bmatrix}$$

2.2.2.1. El clasificador bayesiano para clases normalmente distribuidas

Considere las pdfs, $p(x|\omega_i)$, $i = 1, 2, \dots, M$ (funciones likelihood de ω_i con respecto a x), que describen la distribución de datos en cada una de las clases, son distribuciones normales multivariadas, esto es, $\mathcal{N}(\mu_i, \Sigma_i)$, $i = 1, 2, \dots, M$. Debido a la forma exponencial de las densidades involucradas, es preferible trabajar con las siguientes funciones discriminantes, la función logarítmica (monótona) $\ln(\cdot)$:

$$g_i(x) = \ln(p(x|\omega_i)P(\omega_i)) = \ln p(x|\omega_i) + \ln P(\omega_i) \tag{2.16}$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \ln P(\omega_i) + c_i \tag{2.17}$$

donde $c_i = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$

$$g_i(x) = -\frac{1}{2}x^T\Sigma_i^{-1}x + \frac{1}{2}x^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i + \frac{1}{2}\mu_i^T\Sigma_i^{-1}x + \ln P(\omega_i) + c_i \tag{2.18}$$

En general, esta es una forma cuadrática nolineal.

Ejemplo 11. Considere el caso particular de $l = 2$ y dos clases ω_1, ω_2, entonces $x = [x_1, x_2]^T$, $\mu_i = [\mu_{i1}, \mu_{i2}]^T$ y

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}$$

para $i = 1, 2$.

Entonces la Ecuación 2.18 se convierte en:

$$g_i(x) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(\omega_i) + c_i \tag{2.19}$$

Las curvas de decisión asociadas $g_i(x) - g_j(x) = 0$ son cuadráticas (i.e., elipsoides, parábolas, hipérbolas). En tales casos, el clasificador bayesiano es un clasificador cuadrático, en el sentido que la partición del espacio característico es realizado por medio de superfîcies de decisión cuadrática.

2.3. Estimación de funciones de densidad de probabilidad desconocidas

Hasta ahora, hemos asumido que las funciones de densidad de probabilidad son conocidas. Sin embargo, este no es el caso más común. En muchos problemas, las pdf tienen que ser estimadas a partir de los datos disponibles. A veces, podemos saber el tipo de la pdf (por ejemplo: Gusiana, binormal, etc.), pero no conocemos ciertos parámetros, tales como las medias o varianzas. Por otro lado, en otros casos no podemos tener información acerca del tipo de la pdf pero podemos conocer ciertos parámetros, tales como las medias o varianzas. Por lo tanto, dependiendo de la información disponible, diferentes enfoques podemos adoptar.

2.3.1. Estimación de parámetros de Máxima Verosimilitud (en inglés Maximum Likelihood (ML))

Consideremos un problema de M clases con vectores de características distribuidos de acuerdo a $p(x|\omega_i)$, $i = 1, 2, \ldots, M$. Supongamos que estas funciones de verosimilitud son dadas en una forma paramétrica y que los parámetros correspondientes forman los vectores θ_i los cuáles son desconocidos. Para mostrar la dependencia de θ_i, escribimos $p(x|\omega_i; \theta_i)$. El objetivo es estimar los parámetros desconocidos usando un conjunto de vectores de características conocidos en cada clase.

Sea $X = \{x_1, x_2, \ldots, x_N\}$ el conjunto de las muestras aleatorias, asumimos independencia estadística entre las diferentes muestras, entonces la pdf conjunta $p(X; \theta)$ es dada por:

$$p(X; \theta) = p(x_1, x_2, \ldots, x_N; \theta) = \prod_{k=1}^{N} p(x_k; \theta) \tag{2.20}$$

Esta es una función de θ, y también es conocida como la función de verosimilitud de θ con respecto a X. El método de máxima verosimilitud (o en inglés *maximum likelihood* (ML)) estima θ cuando la función de verosimilitud tome su valor máximo, es decir:

$$\hat{\theta}_{ML} = \arg \max_{\theta} \prod_{k=1}^{N} p(x_k; \theta) \tag{2.21}$$

Una condición necesaria que $\hat{\theta}_{ML}$ debe cumplir para alcanzar un máximo es que el gradiente de la función de verosimilitud con respecto a θ sea cero, es decir:

$$\frac{\partial \prod_{k=1}^{N} p(x_k; \theta)}{\partial \theta} = 0 \tag{2.22}$$

Debido a que la función logaritmo es una función monótona, definimos la función log-verosimilitud como:

$$L(\theta) = \ln \prod_{k=1}^{N} p(x_k; \theta) \tag{2.23}$$

y la Ecuación 2.22 es equivalente a:

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial \ln p(x_k; \theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{1}{p(x_k; \theta)} \frac{\partial p(x_k; \theta)}{\partial \theta} = 0 \tag{2.24}$$

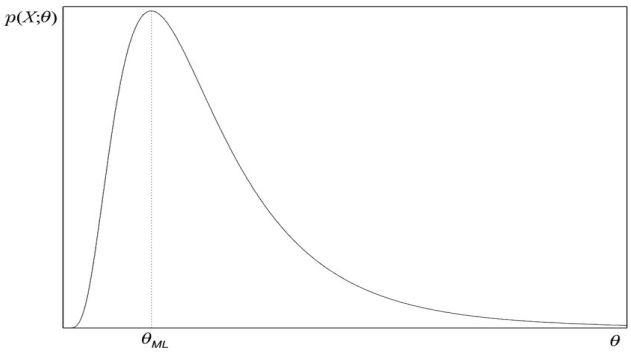


Figura 2.3: El estimador de máxima verosimilitud corresponde al pico de $p(X;\theta)$ Theodoridis y Koutroumbas (2009)

La Figura 2.3 ilustra el método para el caso de parámetro desconocido único. La estimación de ML corresponde al pico de la función log-verosimilitud.

Ejemplo 12. Suponga que los N puntos de datos $x_1, x_2, \ldots, x_N$, han sido generados por una pdf Gaussiana unidimensional, donde μ es conocida y σ^2 es desconocida.
La función de log-verosimilitud para este caso viene dada por:

$$\begin{aligned}
L(\sigma^2) &= \ln \prod_{k=1}^{N} p(x_k;\sigma^2) \\
&= \ln \prod_{k=1}^{N} \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{(x_k-\mu)^2}{2\sigma^2}\right) \\
&= \sum_{k=1}^{N} \ln\left(\tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{(x_k-\mu)^2}{2\sigma^2}\right)\right) \\
&= \sum_{k=1}^{N} \ln\left((2\pi\sigma^2)^{\frac{-1}{2}} \exp\left(-\tfrac{(x_k-\mu)^2}{2\sigma^2}\right)\right) \\
&= \sum_{k=1}^{N} \left(\ln(2\pi\sigma^2)^{\frac{-1}{2}} + \ln\exp\left(-\tfrac{(x_k-\mu)^2}{2\sigma^2}\right)\right) \\
&= \sum_{k=1}^{N} \ln(2\pi\sigma^2)^{\frac{-1}{2}} + \sum_{k=1}^{N} -\tfrac{1}{2\sigma^2}(x_k-\mu)^2 \\
&= -\tfrac{N}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2
\end{aligned}$$

Tomando la derivada de $L(\sigma^2)$ con respecto a σ^2 e igualando a cero, se obtiene:

$$\frac{\partial L(\sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{k=1}^{N}(x_k-\mu)^2 = 0$$

y finalmente la estimación de ML de σ^2 resulta de la solución de la ecuación anterior:

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 \tag{2.25}$$

Para N finito, $\hat{\sigma}^2_{ML}$ en la Ecuación 2.25 es una estimación sesgada de la varianza. Es decir:

$$E[\hat{\sigma}^2_{ML}] = \frac{N-1}{N}\sigma^2$$

donde σ^2 es la varianza de la pdf Gaussiana.

En efecto:

$$\begin{aligned}
E[\hat{\sigma}^2_{ML}] &= E\left[\tfrac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2\right] \\
&= \tfrac{1}{N}E\left[\sum_{k=1}^{N}(x_k-\mu)^2\right] \\
&= \tfrac{1}{N}E\left[\sum_{k=1}^{N}(x_k^2 - 2x_k\mu + \mu^2)\right] \\
&= \tfrac{1}{N}\left(E\left[\sum_{k=1}^{N}x_k - 2\mu\sum_{k=1}^{N}x_k^2 + \sum_{k=1}^{N}\mu^2\right]\right)
\end{aligned}$$

Sabemos que $\mu = \frac{1}{N}\sum_{k=1}^{N}x_k$, entonces:

$$\begin{aligned}
E[\hat{\sigma}^2_{ML}] &= \tfrac{1}{N}\left(E\left[\sum_{k=1}^{N}x_k^2 - 2N\mu^2 + N\mu^2\right]\right) \\
&= \tfrac{1}{N}\left(E\left[\sum_{k=1}^{N}x_k^2 - N\mu^2\right]\right) \\
&= \tfrac{1}{N}\left(\sum_{k=1}^{N}E[x_k^2] - NE[\mu^2]\right) \\
&= \tfrac{1}{N}\sum_{k=1}^{N}E[x_k^2] - E[\mu^2] \\
&= \tfrac{1}{N}\sum_{k=1}^{N}E[x_k^2] - E\left[\left(\tfrac{1}{N}\sum_{k=1}^{N}x_k\right)^2\right] \\
&= \tfrac{1}{N}\sum_{k=1}^{N}E[x_k^2] - \tfrac{1}{N^2}E\left[\left(\sum_{k=1}^{N}x_k\right)^2\right]
\end{aligned}$$

Además, $E[x_k^2] = \sigma^2 + \mu^2$ y $\left(\sum_{k=1}^{N}x_k\right)^2 = \sum_{k=1}^{N}x_k^2 + \sum_{i\neq j}x_ix_j$ (el segunto sumando tiene

$(N^2 - N)$-elementos), entonces:

$$\begin{aligned}\sum_{k=1}^{N} E[x_k^2] &= NE[x_k^2] \\ &= N(\sigma^2 + \mu^2)\end{aligned}$$

$$\begin{aligned}E\left[\left(\sum_{k=1}^{N} x_k\right)^2\right] &= \sum_{k=1}^{N} E[x_k^2] + \sum_{i\neq j} E[x_i x_j] \\ &= N(E[x_k^2]) + (N^2 - N)E[x_i]E[x_j] \\ &= N(\sigma^2 + \mu^2) + (N^2 - N)\mu\mu \\ &= N(\sigma^2 + \mu^2) + (N^2 - N)\mu^2 \\ &= N(\sigma^2 + N\mu^2)\end{aligned}$$

Reemplazando, tenemos:

$$\begin{aligned}E[\hat{\sigma}_{ML}^2] &= \tfrac{1}{N}N(\sigma^2 + \mu^2) - \tfrac{1}{N^2}\left(N(\sigma^2 + N\mu^2)\right) \\ &= \sigma^2 + \mu^2 - \tfrac{1}{N}(\sigma^2 + N\mu^2) \\ &= \sigma^2 + \mu^2 - \tfrac{1}{N}\sigma^2 - \mu^2 \\ &= \tfrac{N-1}{N}\sigma^2\end{aligned}$$

Para valores grandes de N, tenemos:

$$E[\hat{\sigma}_{ML}^2] = \left(1 - \frac{1}{N}\right)\sigma^2 \approx \sigma^2$$

Ejemplo 13. Sean $x_1, x_2, \ldots, x_N$ vectores $l-$dimensional, que han sido generados por una distribución Normal, con matriz de covarianza conocida y media desconocida, esto es:

$$p(x_k; \mu) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k - \mu)^T\Sigma^{-1}(x_k - \mu)\right)$$

Obtenga la estimación de ML del vector de la media desconocida.

La función de log-verosimilitud para este caso viene dada por:

$$\begin{aligned}
L(\mu) &= \ln \prod_{k=1}^{N} p(x_k;\mu) \\
&= \ln \prod_{k=1}^{N} \tfrac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) \\
&= \sum_{k=1}^{N} \ln \tfrac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) \\
&= \sum_{k=1}^{N} \ln((2\pi)^l|\Sigma|)^{\frac{-1}{2}} \exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) \\
&= \sum_{k=1}^{N} \left(\ln((2\pi)^l|\Sigma|)^{\frac{-1}{2}} + \ln\exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)\right) \\
&= \sum_{k=1}^{N} \ln((2\pi)^l|\Sigma|)^{\frac{-1}{2}} + \sum_{k=1}^{N} -\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) \\
&= -\tfrac{N}{2}\ln((2\pi)^l|\Sigma|) - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) \\
&= -\tfrac{N}{2}\ln((2\pi)^l) - \tfrac{N}{2}\ln(|\Sigma|) - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) \\
&= -\tfrac{Nl}{2}\ln(2\pi) - \tfrac{N}{2}\ln(|\Sigma|) - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)
\end{aligned}$$

Tomando el gradiente con respecto a μ, obtenemos:

$$\begin{aligned}
\tfrac{\partial L(\mu)}{\partial \mu} &= \tfrac{\partial}{\partial \mu}\left(-\tfrac{Nl}{2}\ln(2\pi) - \tfrac{N}{2}\ln(|\Sigma|) - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) \\
&= -\tfrac{1}{2}\tfrac{\partial}{\partial \mu}\left(\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) \\
&= -\tfrac{1}{2}\left(\sum_{k=1}^{N}(-2)\Sigma^{-1}(x_k-\mu)\right) \text{ por la Ecuación A.8} \\
&= \sum_{k=1}^{N}\Sigma^{-1}(x_k-\mu)
\end{aligned}$$

Igualando el gradiente con respecto a μ a cero, obtenemos:

$$\sum_{k=1}^{N}\Sigma^{-1}(x_k-\mu) = 0$$

Por lo tanto:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

Es decir, la estimación ML de la media, para las densidades gaussianas, es la media muestral.

Ejemplo 14. Sean $x_1, x_2, \ldots, x_N$ vectores $l-$dimensional, que han sido generados por una distribución Normal, con media conocida y matriz de covarianza desconocida, esto es:

$$p(x_k;\Sigma) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)$$

La función de log-verosimilitud para este caso viene dada por (por el ejemplo anterior):

$$L(\Sigma) = -\frac{Nl}{2}\ln(2\pi) - \frac{N}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)$$

Vamos a reescribir la función de log-verosimilitud en función de Σ^{-1}.

$$\begin{aligned}
L(\Sigma) &= -\tfrac{Nl}{2}\ln(2\pi) - \tfrac{N}{2}\ln(\tfrac{1}{|\Sigma|})^{-1} - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) \\
&= -\tfrac{Nl}{2}\ln(2\pi) + \tfrac{N}{2}\ln(|\Sigma^{-1}|) - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) \text{ por la Ecuación A.4} \\
&= -\tfrac{Nl}{2}\ln(2\pi) + \tfrac{N}{2}\ln(|\Sigma^{-1}|) - \tfrac{1}{2}\sum_{k=1}^{N} tr(\Sigma^{-1}(x_k-\mu)(x_k-\mu)^T) \text{ por la Ecuación A.9}
\end{aligned}$$

Tomando el gradiente con respecto a Σ^{-1}, obtenemos:

$$\begin{aligned}
\tfrac{\partial L(\Sigma)}{\partial \Sigma^{-1}} &= \tfrac{\partial}{\partial \Sigma^{-1}}\left(-\tfrac{Nl}{2}\ln(2\pi) + \tfrac{N}{2}\ln(|\Sigma^{-1}|) - \tfrac{1}{2}\sum_{k=1}^{N} tr(\Sigma^{-1}(x_k-\mu)(x_k-\mu)^T)\right) \\
&= \tfrac{N}{2}\tfrac{\partial}{\partial \Sigma^{-1}}\left(\ln(|\Sigma^{-1}|)\right) - \tfrac{1}{2}\sum_{k=1}^{N}\tfrac{\partial}{\partial \Sigma^{-1}}\left(tr(\Sigma^{-1}(x_k-\mu)(x_k-\mu)^T)\right) \\
&= \tfrac{N}{2}((\Sigma^{-1})^{-1})^T - \tfrac{1}{2}\sum_{k=1}^{N}((x_k-\mu)(x_k-\mu)^T))^T \text{ por la Ecuación A.7 y A.6} \\
&= \tfrac{N}{2}(\Sigma)^T - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)(x_k-\mu)^T \text{ por la Ecuación A.2} \\
&= \tfrac{N}{2}\Sigma - \tfrac{1}{2}\sum_{k=1}^{N}(x_k-\mu)(x_k-\mu)^T \text{ pues } \Sigma \text{ es simétrica}
\end{aligned}$$

Igualando la Ecuación anterior a cero, obtenemos:

$$\frac{N}{2}\Sigma - \frac{1}{2}\sum_{k=1}^{N}(x_k-\mu)(x_k-\mu)^T = 0$$

Por lo tanto:

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)(x_k-\mu)^T$$

2.3.2. Estimación de la Probabilidad Máxima a Posteriori (en inglés Maximum a Posteriori (MAP))

Para la derivación de la estimación de máxima verosimilitud, consideramos θ como un parámetro desconocido. En esta subsección consideraremos θ como un vector aleatorio descrito por una pdf $p(\theta|X)$, que supondremos conocida. Sea $X = \{x_1, x_2, \ldots, x_N\}$. Del teorema de Bayes, tenemos:

$$p(\theta)p(X|\theta) = p(X)p(\theta|X) \tag{2.26}$$

o lo que es equivalente:

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)} \tag{2.27}$$

La estimación de la Probabilidad Máxima a posteriori (MAP) $\hat{\theta}_{MAP}$ se define en el punto donde $p(\theta|X)$ se convierte en máximo:

$$\hat{\theta}_{MAP} = \arg\max p(\theta|X)$$

Es decir:

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial\theta}p(\theta|X) = 0 \quad \text{ó} \quad \frac{\partial}{\partial\theta}(p(\theta)p(X|\theta)) = 0 \tag{2.28}$$

Note que $p(X)$ no está involucrado en la Ecuación 2.28 ya que es independiente de θ. La diferencia entre las estimaciones de ML y MAP radica en la participación de $p(\theta)$ en este último caso. Si suponemos que $p(\theta)$ obedece a la distribución uniforme, es decir, es constante para todo θ, ambas estimaciones arrojan resultados idénticos. Esto también es aproximadamente cierto si $p(\theta)$ muestra una pequeña variación. Sin embargo, en el caso general, los dos métodos producen resultados diferentes.

Debido a que la función logaritmo es una función monótona, definimos:

$$\hat{\theta}_{MAP} = \arg\max \ln p(\theta|X)$$

Por lo tanto la Ecuación 2.28, es equivalente a:

$$\frac{\partial \ln p(\theta)}{\partial\theta} + \frac{\partial \ln p(X|\theta)}{\partial\theta} = 0$$

Ejemplo 15. Sean $x_1, x_2, \ldots, x_N$ vectores $l-$dimensional, que han sido generados por una distribución Normal, con matriz de covarianza conocida $\Sigma = \sigma^2 I$ y media desconocida, esto es:

$$p(x_k;\mu) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)$$

Además el vector medio desconocido se distribuye normalmente como:

$$p(\mu) = \frac{1}{(2\pi)^{\frac{l}{2}}\sigma_\mu^l} \exp\left(-\frac{1}{2}\frac{\|\mu-\mu_0\|^2}{\sigma_\mu^2}\right)$$

Obtenga la estimación de MAP del vector de la media desconocida.

$$\frac{\partial}{\partial\mu}\ln p(\mu) + \frac{\partial}{\partial\mu}\ln\prod_{k=1}^{N} p(x_k|\mu) = 0$$

$$\frac{\partial}{\partial\mu}\ln\frac{1}{(2\pi)^{\frac{l}{2}}\sigma_\mu^l}\exp\left(-\tfrac{1}{2}\tfrac{\|\mu-\mu_0\|^2}{\sigma_\mu^2}\right) + \frac{\partial}{\partial\mu}\sum_{k=1}^{N}\ln\frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right) = 0$$

$$\frac{\partial}{\partial\mu}\left[\ln\frac{1}{(2\pi)^{\frac{l}{2}}\sigma_\mu^l} + \ln\exp\left(-\tfrac{1}{2}\tfrac{\|\mu-\mu_0\|^2}{\sigma_\mu^2}\right)\right] + \sum_{k=1}^{N}\frac{\partial}{\partial\mu}\left[\ln\frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} + \ln\exp\left(-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)\right] = 0$$

$$\frac{\partial}{\partial\mu}\left[-\tfrac{1}{2}\tfrac{\|\mu-\mu_0\|^2}{\sigma_\mu^2}\right] + \sum_{k=1}^{N}\frac{\partial}{\partial\mu}\left[-\tfrac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right] = 0$$

$$-\tfrac{1}{2\sigma_\mu^2}2(\mu-\mu_0) - \tfrac{1}{2}\sum_{k=1}^{N}\frac{\partial}{\partial\mu}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu) = 0 \text{ por la Ecuación A.8}$$

$$-\tfrac{(\mu-\mu_0)}{\sigma_\mu^2} - \tfrac{1}{2}\sum_{k=1}^{N}(-2)\Sigma^{-1}(x_k-\mu) = 0$$

$$-\tfrac{(\mu-\mu_0)}{\sigma_\mu^2} + \sum_{k=1}^{N}\tfrac{1}{\sigma^2}(x_k-\mu) = 0$$

$$-\tfrac{\mu}{\sigma_\mu^2} + \tfrac{\mu_0}{\sigma_\mu^2} + \tfrac{1}{\sigma^2}\sum_{k=1}^{N}x_k - \tfrac{N}{\sigma^2}\mu = 0$$

$$\tfrac{\mu_0}{\sigma_\mu^2} + \tfrac{1}{\sigma^2}\sum_{k=1}^{N}x_k = \left(\tfrac{1}{\sigma_\mu^2} + \tfrac{N}{\sigma^2}\right)\mu$$

$$\tfrac{\mu_0\sigma^2+\sigma_\mu^2\sum_{k=1}^{N}x_k}{\sigma_\mu^2+\sigma^2} = \left(\tfrac{\sigma^2+N\sigma_\mu^2}{\sigma^2+\sigma_\mu^2}\right)\mu$$

Por lo tanto:

$$\hat{\mu}_{MAP} = \frac{\mu_0\sigma^2 + \sigma_\mu^2\sum_{k=1}^{N}x_k}{\sigma^2 + N\sigma_\mu^2}$$

Dividimos a la ecuación anterior por σ^2, entonces tenemos:

$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{N}x_k}{1 + N\frac{\sigma_\mu^2}{\sigma^2}}$$

Si $\frac{\sigma_\mu^2}{\sigma^2} \gg 1$ (es decir la varianza σ_μ^2 es mucho mayor que la varianza σ^2), entonces, la ecuación anterior puede ser escrita como:

$$\begin{aligned}\hat{\mu}_{MAP} &= \frac{\mu_0}{1+N\frac{\sigma_\mu^2}{\sigma^2}} + \frac{\frac{\sigma_\mu^2}{\sigma^2}}{1+N\frac{\sigma_\mu^2}{\sigma^2}} \sum_{k=1}^{N} x_k \\ &= \frac{\mu_0}{1+N\frac{\sigma_\mu^2}{\sigma^2}} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_\mu^2}+N\frac{1}{\sigma^2}} \sum_{k=1}^{N} x_k\end{aligned}$$

Entonces:

$$\hat{\mu}_{MAP} \approx \hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

2.3.3. Inferencia Bayesiana

Ambos métodos considerados en las subsecciones anteriores, calculan una estimación específica del vector de parámetros desconocidos. En el método que vamos a describir a continuación, se adopta una ruta diferente. Dado el conjunto X de los N vectores de entrenamiento y la información *a priori* sobre la pdf $p(\theta)$, el objetivo es calcular la pdf condicional $p(x|X)$. A partir de las identidades conocidas de los conceptos básicos de estadística, sabemos que:

$$p(x|X) = \int p(x|\theta)p(\theta|X)d\theta \tag{2.29}$$

donde:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta} \tag{2.30}$$

$$p(X|\theta) = \prod_{k=1}^{N} p(x_k|\theta) \tag{2.31}$$

La densidad condicional $p(\theta|X)$ también se conoce como la estimación de la pdf a posteriori, ya que se actualiza lo que "sabemos" de las propiedades estadísticas de θ, después de haber observado el conjunto de datos X. Una vez más, la Ecuación 2.31 supone independencia estadística entre las muestras de entrenamiento.

En general, el cálculo de $p(x|X)$ requiere la integración del lado derecho de la Ecuación 2.29. Sin embargo, las soluciones analíticas son factibles sólo para formas muy especiales de las funciones involucradas. Para la mayoría de los casos, las soluciones analíticas para la Ecuación 2.29, así como para el denominador en la Ecuación 2.30, no son posibles, y

tenemos que recurrir a aproximaciones numéricas.

Mirando más cuidadosamente la Ecuación 2.29 y suponiendo que $p(\theta|X)$ es conocido, entonces $p(x|X)$ no es más que el promedio de $p(x|\theta)$ con respecto a θ, es decir:

$$p(x|X) = E_\theta[p(x|\theta)]$$

Ejemplo 16. Sea $p(x|\mu)$ una pdf Gaussiana unidimensional $\mathcal{N}(\mu, \sigma^2)$ con un parámetro desconocido, la media, que también se supone que sigue una pdf Gaussiana $\mathcal{N}(\mu_0, \sigma_0^2)$ (Jurcicek, 2014), (Murphy, 2007). Entonces:

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} = \frac{1}{\alpha}\prod_{k=1}^{N} p(x_k|\mu)p(\mu)$$

donde para un conjunto de datos de entrenamiento dado X, $p(X)$ es una constante denotada como α, es decir:

$$\begin{aligned}
p(\mu|X) &= \tfrac{1}{\alpha}\prod_{k=1}^{N} \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{(x_k-\mu)^2}{2\sigma^2}\right)\tfrac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\\
&= \tfrac{1}{\alpha}\left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^N\exp\left(-\tfrac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2\right)\tfrac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\\
&= \tfrac{1}{\alpha}\tfrac{1}{(\sqrt{2\pi}\sigma)^N\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{1}{2}\left[\tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2+\tfrac{1}{\sigma_0^2}(\mu-\mu_0)^2\right]\right)\\
&= \tfrac{1}{\alpha}\tfrac{1}{(\sqrt{2\pi}\sigma)^N\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{1}{2}A\right)
\end{aligned}$$

Sea $\bar{x}_N = \frac{1}{N}\sum_{k=1}^{N} x_k$, entonces:

$$\begin{aligned}
\sum_{k=1}^{N}(x_k-\mu)^2 &= \sum_{k=1}^{N}(x_k-\bar{x}_N+\bar{x}_N-\mu)^2\\
&= \sum_{k=1}^{N}\left[(x_k-\bar{x}_N)^2+2(x_k-\bar{x}_N)(\bar{x}_N-\mu)+(\bar{x}_N-\mu)^2\right]\\
&= \sum_{k=1}^{N}(x_k-\bar{x}_N)^2+2(\bar{x}_N-\mu)\sum_{k=1}^{N}(x_k-\bar{x}_N)+N(\bar{x}_N-\mu)^2\\
&= \sum_{k=1}^{N}(x_k-\bar{x}_N)^2+2(\bar{x}_N-\mu)\left(\sum_{k=1}^{N}x_k-N\bar{x}_N\right)+N(\bar{x}_N-\mu)^2\\
&= \sum_{k=1}^{N}(x_k-\bar{x}_N)^2+2(\bar{x}_N-\mu)\left(\sum_{k=1}^{N}x_k-\sum_{k=1}^{N}x_k\right)+N(\bar{x}_N-\mu)^2\\
&= \sum_{k=1}^{N}(x_k-\bar{x}_N)^2+N(\bar{x}_N-\mu)^2
\end{aligned}$$

Por lo tanto:

$$\begin{aligned}
A &= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2+\tfrac{1}{\sigma_0^2}(\mu-\mu_0)^2 \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N}{\sigma^2}(\bar{x}_N-\mu)^2+\tfrac{1}{\sigma_0^2}(\mu-\mu_0)^2 \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N}{\sigma^2}(\bar{x}_N^2-2\bar{x}_N\mu+\mu^2)+\tfrac{1}{\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2) \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}+\left(\tfrac{N}{\sigma^2}+\tfrac{1}{\sigma_0^2}\right)\mu^2-2\left(\tfrac{N\bar{x}_N}{\sigma^2}+\tfrac{\mu_0}{\sigma_0^2}\right)\mu
\end{aligned}$$

Haciendo:

$$\sigma_N^2=\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}=\left(\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)^{-1}=\frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}$$

$$\mu_N=\frac{N\sigma_0^2\bar{x}_N+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}=\sigma_N^2\left(\frac{N\sigma_0^2\bar{x}_N+\sigma^2\mu_0}{\sigma^2\sigma_0^2}\right)$$

Reemplazando entonces, tenemos:

$$\begin{aligned}
A &= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}+\tfrac{1}{\sigma_N^2}\mu^2-2\left(\tfrac{N\sigma_0^2\bar{x}_N+\mu_0\sigma^2}{\sigma^2\sigma_0^2}\right)\mu \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}+\tfrac{1}{\sigma_N^2}\mu^2-2\left(\tfrac{\mu_N}{\sigma_N^2}\right)\mu \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}+\tfrac{1}{\sigma_N^2}(\mu^2-2\mu_N\mu) \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}-\tfrac{1}{\sigma_N^2}\mu_N^2+\tfrac{1}{\sigma_N^2}(\mu^2-2\mu_N\mu+\mu_N^2) \\
&= \tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}-\tfrac{1}{\sigma_N^2}\mu_N^2+\tfrac{1}{\sigma_N^2}(\mu-\mu_N)^2
\end{aligned}$$

Por lo tanto:

$$\begin{aligned}
p(\mu|X) &= \tfrac{1}{\alpha}\tfrac{1}{(\sqrt{2\pi}\sigma)^N\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{1}{2}\left(\tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}-\tfrac{1}{\sigma_N^2}\mu_N^2+\tfrac{1}{\sigma_N^2}(\mu-\mu_N)^2\right)\right) \\
&= \tfrac{1}{\alpha}\tfrac{1}{(\sqrt{2\pi}\sigma)^N\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{1}{2}\left(\tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}-\tfrac{1}{\sigma_N^2}\mu_N^2\right)-\tfrac{1}{2}\left(\tfrac{1}{\sigma_N^2}(\mu-\mu_N)^2\right)\right) \\
&= \tfrac{1}{\alpha}\tfrac{1}{(\sqrt{2\pi}\sigma)^N\sqrt{2\pi}\sigma_0}\exp\left(-\tfrac{1}{2}\left(\tfrac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\bar{x}_N)^2+\tfrac{N\bar{x}_N^2}{\sigma^2}+\tfrac{\mu_0^2}{\sigma_0^2}-\tfrac{1}{\sigma_N^2}\mu_N^2\right)\right)\exp\left(-\tfrac{(\mu-\mu_N)^2}{2\sigma_N^2}\right) \\
&\propto \tfrac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\tfrac{(\mu-\mu_N)^2}{2\sigma_N^2}\right)
\end{aligned}$$

Es decir:

$$p(\mu|X)\propto\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right)\sim\mathcal{N}(\mu|\mu_N,\sigma_N^2)$$

donde

$$\mu_N = \frac{N\sigma_0^2 \bar{x}_N + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} = \frac{N\sigma_0^2 \bar{x}_N}{N\sigma_0^2 + \sigma^2} + \frac{\sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

$$\bar{x}_N = \frac{1}{N} \sum_{k=1}^{N} x_k = \hat{\mu}_{ML}$$

Observamos que si $N = 0$, entonces la pdf posterior $p(\mu|X)$ es la pdf gaussiana con media μ_0. Cuando $N \to \infty$, la pdf posterior $p(\mu|X)$ tiende a la pdf gaussiana con media $\hat{\mu}_{ML}$. Además su varianza disminuye cuando N crece. Por lo tanto, para valores grandes de N, $p(\mu|X)$ llega a su punto máximo alrededor de la media muestral $\hat{\mu}_{ML}$.

Por lo tanto, el predictivo posterior viene dado por (Murphy, 2007):

$$\begin{aligned} p(x|X) &= \int p(x|\mu) p(\mu|X) d\mu \\ &= \int \mathcal{N}(x|\mu, \sigma^2) \mathcal{N}(\mu|\mu_N, \sigma_N^2) d\mu \\ &= \mathcal{N}(x|\mu_N, \sigma^2 + \sigma_N^2) \end{aligned}$$

Pues:

$$\begin{aligned} x &= (x - \mu) + \mu \\ x - \mu &\sim \mathcal{N}(0, \sigma^2) \\ \mu &\sim \mathcal{N}(\mu_N, \sigma_N^2) \end{aligned}$$

Entonces $x \sim \mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$

$$p(x|X) = \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_N^2)}} \exp\left(-\frac{1}{2} \frac{(x - \mu_N)^2}{\sigma^2 + \sigma_N^2} \right)$$

Observe que a medida que N tiende al infinito, el valor de la media desconocida del gaussiano tiende a la estimación de $\hat{\mu}_{ML}$ y la varianza al valor σ^2.

CAPÍTULO 3

MODELOS DE MEZCLA

En este capítulo estudiaremos la teoría de Modelos Mixtos, variables ocultas, variables observables, el Algoritmo de Esperanza-Maximización (EM).

3.1. Introducción

Un modelo de mezcla es un modelo probabilístico para representar la presencia de subpoblaciones dentro de una misma población general, sin requerir que un conjunto de datos observados identifique la subpoblación a la que pertenece una observación individual (McLachlan & Peel, 2004), (Mahjoub *et al.*., 2012) (Chen, 2008), (Wilson, 2015), (Haas, 2002).

El modelo de mezcla se construye a partir de una combinación ponderada de distribuciones de probabilidad de componentes (Baxter, 2017).

McLachlan y Basford (1988) enfatiza las distribuciones de mezclas finitas en el modelado de datos heterogéneos, en el campo del análisis de agrupamientos. Los autores se centran en la adaptación de modelos de mezcla, mediante un enfoque basado en la probabilidad, utilizando la máxima verosimilitud. El algoritmo EM proporciona una forma conveniente para el cálculo iterativo de soluciones de la ecuación de *likelihood*. También hacen

hincapié en las mezclas de distribuciones normales, ya que estos modelos se emplean más ampliamente en la práctica y se estudian con mayor frecuencia.

Las mezclas de distribuciones, en particular las normales, se han utilizado ampliamente como modelos en una amplia variedad de situaciones prácticas importantes, en las que se puede ver que los datos surgen de dos o más poblaciones mezcladas en proporciones variables (Tukey, 1960).

3.2. Modelos de Mezcla Finita

Una manera para modelar una pdf $p(x)$ desconocida es vía una combinación lineal de funciones de densidades. Sea $X \in \Omega$ una variable aleatoria tal que $\Omega \in \mathbb{R}^d$ para algún $d \in \mathbb{N}$ y sea $p_i(x)$ una función de densidad de probabilidad para cada $i = 1, \ldots, g$, donde $g \in \mathbb{N}$. La variable aleatoria X surge de un modelo de mezcla finita, digamos g poblaciones $G_1, \ldots, G_g$ en algunas proporciones $\pi_1, \ldots, \pi_g$ respectivamente (McLachlan & Peel, 2004) donde:

$$\sum_{i=1}^{g} \pi_i = 1 \text{ y } \pi_i \geq 0 \quad (i = 1, \ldots, g)$$

La función de densidad de probabilidad de una observación x puede ser representada en forma de mezcla finita:

$$p(x; \Theta) = \sum_{i=1}^{g} \pi_i . p_i(x; \theta_i) \tag{3.1}$$

donde $p_i(x; \theta_i)$ es la función de densidad de probabilidad corresponde a G_i, y Θ denota el vector de todos los parámetros desconocidos asociados con las formas paramétricas adoptadas para estas g componentes de densidades. Por ejemplo, en el caso particular de las densidades de componentes normales multivariantes, θ_i consiste del vector de la media μ_i y la matriz de covarianza Σ_i, es decir $\theta_i = (\mu_i, \Sigma_i)$. El vector:

$$\Theta = (\pi, \theta)$$

consiste de todos los parámetros desconocidos del modelo de la mezcla, donde $\pi = (\pi_1, \ldots, \pi_g)$, y $\theta = (\theta_1, \ldots, \theta_g)$

3.2.1. Modelo de Mezcla Finita de Gaussianas

Si bien la distribución gaussiana tiene algunas propiedades analíticas importantes, sufre de limitaciones significativas cuando se trata de modelar conjuntos de datos reales. En muchos casos una distribución gaussiana simple no puede modelar adecuadamente algunos datos, mientras que una superposición lineal de un número finito de gaussianas da una mejor caracterización del conjunto de datos.

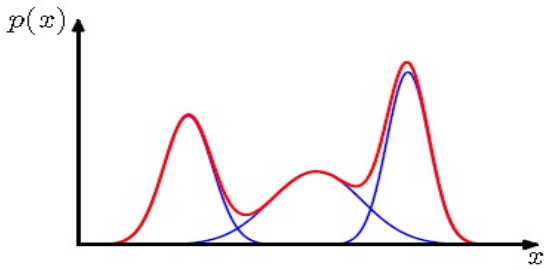


Figura 3.1: Ejemplo de una distribución de mezcla gaussiana en una dimensión que muestra tres gaussianos en azul y su suma en rojo. (Bishop, 2006)

Dichas superposiciones, se pueden formular como modelos probabilísticos conocidos como distribuciones de mezclas (McLachlan & Basford, 1988), (McLachlan & Peel, 2004).

En la Figura 3.1 vemos que una combinación lineal de gaussianos puede dar lugar a densidades muy complejas. Al utilizar un número suficiente de gaussianos, y al ajustar sus medias y covarianzas, así como los coeficientes en la combinación lineal, casi cualquier densidad continua se puede aproximar a una precisión arbitraria.

Por lo tanto, consideramos una superposición de K densidades Gaussianas de la forma (Góngora, 2010):

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3.2}$$

la cuál se llama una mezcla de gaussianos. Cada densidad gaussiana $\mathcal{N}(x|\mu_k, \Sigma_k)$ se denomina componente de la mezcla y tiene su propia media μ_k y matriz de covarianza Σ_k. Los parámetros π_k se denominan coeficientes de mezcla (proporción o peso de la

mezcla), que satisfacen $0 \leq \pi_k \leq 1$ y:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.3}$$

De las reglas de suma y producto de probabilidades, la densidad marginal es dada por:

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) \tag{3.4}$$

la cuál es equivalente a la Ecuación 3.2, en la cuál podemos ver $\pi_k = p(k)$ como la probabilidad *prior* de elegir la $k-$ésima componente, y la densidad $\mathcal{N}(x|\mu_k, \Sigma_k) = p(x|k)$ como la probabilidad de x condicionado en k.

Sea $p(k|x)$ la probabilidad posterior, las cuáles también se conocen como responsabilidades. Del teorema de Bayes estas son dados por:

$$\begin{aligned} \gamma_k(x) &\equiv p(k|x) \\ &= \frac{p(k)p(x|k)}{p(x)} \\ &= \frac{p(k)p(x|k)}{\sum_l p(l)p(x|l)} \\ &= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(x|\mu_l, \Sigma_l)} \end{aligned} \tag{3.5}$$

La forma de la distribución de la mezcla gaussiana se rige por los parámetros π, μ y Σ, donde $\pi = \{\pi_1, \ldots, \pi_k\}$, $\mu = \{\mu_1, \ldots, \mu_k\}$ y $\Sigma = \{\Sigma_1, \ldots, \Sigma_k\}$. Una forma de encontrar los valores de estos parámetros es usar la máxima verosimilitud. De la Ecuación 3.2 el logaritmo de la función de verosimilitud para un conjunto de datos $X = \{x_1, \ldots, x_n\}$ (asumimos independencia) viene dado por:

$$\begin{aligned} \ln p(X|\pi, \mu, \Sigma) &= \ln \prod_{n=1}^{N} p(x_n|\pi, \mu, \Sigma) \\ &= \ln \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \\ &= \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \end{aligned} \tag{3.6}$$

Vemos que la situación ahora es mucho más compleja que con una sóla distribución gaussiana, debido a la presencia de la suma sobre k dentro del logaritmo.

3.2.1.1. Máxima verosimilitud

Supongamos que tenemos un conjunto de N datos de observaciones $\{x_1, \ldots, x_N\}$, donde cada $x_n \in \mathbb{R}^D$, $n = 1, \ldots, N$ y deseamos modelar estos datos usando una mezcla de K gaussianas. Podemos representar este conjunto de datos como una matriz X de dimensión $N \times D$ en la que $n-$ésima fila es dada por x_n^T (T denota transpuesta). Del mismo modo, las variables latentes correspondientes se denotarán mediante una matriz Z de orden $N \times K$ con filas denotadas por z_n^T. Si suponemos que los puntos de datos se dibujan independientemente de la distribución, entonces podemos expresar el modelo de mezcla gaussiana para este conjunto de datos i.i.d.. El logaritmo de la función de verosimilitud es dado por la Ecuación 3.6: (Jiménez, 2006)

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

Maximizar la función log-verisimilitud para un modelo de mezcla gaussiana resulta ser un problema más complejo que para el caso de un solo gaussiano. La dificultad surge por la presencia de la suma sobre k que aparece dentro del logaritmo, de modo que la función de logaritmo ya no actúa directamente sobre el gaussiano.

Al derivar $\ln p(X|\pi, \mu, \Sigma)$ (Ecuación 3.6) con respecto a las medias μ_k de las componentes Gaussianas, e igualando a cero, obtenemos:

$$\begin{aligned} \frac{\partial}{\partial \mu_k} \ln p(X|\pi, \mu, \Sigma) &= \frac{\partial}{\partial \mu_k} \left(\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \right) \\ &= \sum_{n=1}^{N} \frac{\partial}{\partial \mu_k} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \mu_k} \ln p(X|\pi,\mu,\Sigma) &= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \mu_k}\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \mu_k}\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \mu_k}\left(\pi_k \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} \frac{\partial}{\partial \mu_k}\left(\exp\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right) \frac{\partial}{\partial \mu_k}\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k) \frac{\partial}{\partial \mu_k}\left(-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)(-\Sigma_k^{-1}(x_n-\mu_k))}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}
\end{aligned}
$$

Por lo tanto, tenemos:

$$
0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \Sigma_k^{-1}(x_n-\mu_k) \tag{3.7}
$$

denotemos

$$
\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \tag{3.8}
$$

donde $\gamma(z_{nk})$ representan las probabilidades posteriores o responsabilidades. Multiplicando por Σ_k (que suponemos que no es singular) a la Ecuación 3.7 obtenemos:

$$\begin{aligned} 0 &= -\sum_{n=1}^{N} \gamma(z_{nk})\Sigma_k^{-1}(x_n - \mu_k) \\ &= -\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k) \\ &= -\sum_{n=1}^{N} \gamma(z_{nk})x_n + \sum_{n=1}^{N} \gamma(z_{nk})\mu_k \end{aligned}$$

Entonces:

$$\sum_{n=1}^{N} \gamma(z_{nk})\mu_k = \sum_{n=1}^{N} \gamma(z_{nk})x_n$$

Despejando de la ecuación anterior μ_k, se tiene:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})x_n \tag{3.9}$$

donde definimos:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{3.10}$$

N_k puede ser interpretado como el número efectivo de puntos asignados al grupo k. Vemos que la media μ_k para la $k-$ésima componente gaussiana se obtiene tomando una media ponderada de todos los puntos en el conjunto de datos, en el que el factor de ponderación para el punto de datos x_n viene dado por la probabilidad posterior $\gamma(z_{nk})$ que la k componente fue responsable de generar x_n.

La maximización de $\ln p(X|\pi, \mu, \Sigma)$ (Ecuación 3.6) con respecto a la matriz de covarianza Σ_k de las componentes Gaussianas, es más complicado (Bishop, 2006). Al efectuar dichos cálculos, que no serán realizados en esta tesis, se obtiene:

$$\Sigma_k = \frac{1}{N_k} \sum_{k=1}^{K} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \tag{3.11}$$

la cúal tiene la misma forma que el resultado correspondiente para un solo Gaussiano ajustado al conjunto de datos, pero nuevamente cada punto de los datos ponderado por la probabilidad posterior correspondiente y con el denominador dado por el número

efectivo de puntos asociados con la componente correspondiente.

Finalmente, maximizamos $\ln p(X|\pi, \mu, \Sigma)$ (Ecuación 3.6) con respecto a los coeficientes de mezcla π_k. Aquí debemos tener en cuenta la restricción de la Ecuación 3.3, que requiere que los coeficientes de mezcla sumen uno. Esto se puede lograr usando un multiplicador de Lagrange (Ecuación A.10) y maximizando la siguiente cantidad:

$$L(\pi_k, \lambda) = \ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{3.12}$$

Derivando la Ecuación 3.12 con respecto a π_k, se tiene:

$$\begin{aligned}
\frac{\partial}{\partial \pi_k} L(\pi_k, \lambda) &= \frac{\partial}{\partial \pi_k} \left(\ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \right) \\
&= \frac{\partial}{\partial \pi_k} \ln p(X|\pi, \mu, \Sigma) + \frac{\partial}{\partial \pi_k} \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \\
&= \frac{\partial}{\partial \pi_k} \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} + \lambda \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} + \lambda \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \pi_k} \left(\sum\limits_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} + \lambda
\end{aligned}$$

lo cual dá:

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum\limits_{j} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} + \lambda \tag{3.13}$$

donde de nuevo vemos que las responsabilidades aparecen en la ecuación. Si a ambos lados multiplicamos por π_k y sumamos sobre k y hacemos uso de la Ecuación 3.3, encontramos que

$$\begin{aligned}
0 &= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum\limits_{j} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} + \lambda \pi_k \\
0 &= \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum\limits_{j} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} + \lambda \sum_{k=1}^{K} \pi_k \\
0 &= N + \lambda
\end{aligned}$$

Es decir $\lambda = -N$. Usando esto para eliminar λ y reorganizando, obtenemos:

$$\pi_k = \frac{N_k}{N} \tag{3.14}$$

tal que el coeficiente de mezcla de la $k-$ésima componente es dado por la responsabilidad promedio que esa componente asume para explicar los puntos de datos.

Vale la pena enfatizar que los resultados dadas en las Ecuaciones 3.9, 3.11 y 3.14 no constituyen una solución de forma cerrada para los parámetros del modelo de mezcla porque las responsabilidades $\gamma(z_{nk})$ dependen de esos parámetros de manera compleja a través de la Ecuación 3.8.

Es por eso que para resolver este problema, el logaritmo de la función de verosimilitud dado por la Ecuación 3.6, consideraremos un enfoque alternativo conocido como algoritmo EM que tiene una amplia aplicabilidad. El cual será descrito a continuación.

3.2.1.2. EM para mezclas Gaussiana

Para encontrar soluciones de máxima verosimilitud para modelos con variables latentes aplicamos un método elegante y poderoso llamado el algoritmo de maximización de expectativas o algoritmo EM (McLachlan & Peel, 2004), (Bishop, 2006), (Dempster, Laird, & Rubin, 1977).

Como se ha visto anteriormente, la solución de la máxima verosimilitud para los parámetros ya no tiene una solución analítica. Ahora para resolver este problema, podemos emplear un método poderoso llamado el Algoritmo Esperanza-Maximización (EM).

Ahora pasamos a una formulación de mezclas gaussianas en términos de variables latentes (o variables ocultas). Esto nos proporcionará una visión más profunda de esta importante distribución, y también servirá para motivar el algoritmo EM.

Consideramos una mezcla de K gaussianos, la pdf es dada por:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

tal que $0 \leq \pi_k \leq 1$ y $\sum_{k=1}^{K} \pi_k = 1$.

Introduciremos una variable aleatoria binaria $K-$dimensional z que tiene una representación 1-de-K en la que un elemento particular z_k es igual a 1 y todos los demás elementos son iguales a 0. Es decir:

$$z = (z_1, z_2, \ldots, z_k, \ldots, z_K)$$

donde $z_k \in \{0, 1\}$, $k = 1, \ldots, K$ y $\sum_{k=1}^{K} z_k = 1$. Vemos que hay K posibles estados para el vector z de acuerdo a qué elemento es distinto de cero.

Definimos la distribución conjunta $p(x, z)$ en términos de una distribución marginal $p(z)$ y una distribución condicional $p(x|z)$, es decir:

$$p(x, z) = p(z)p(x|z) \tag{3.15}$$

La distribución marginal sobre z es dada en términos de los coeficientes de mezcla π_k, de modo que:

$$p(z_k = 1) = \pi_k$$

donde los parámetros $\{\pi_k\}$ deben satisfacer:

$$0 \leq \pi_k \leq 1 \tag{3.16}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.17}$$

con el fin de ser probabilidades válidas.

Debido a que z usa una representación 1-de-K, también podemos escribir esta distribución en la forma:

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{3.18}$$

Del mismo modo, la distribución condicional de x dado un valor particular para z es un gaussiano:

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3.19}$$

Además:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \tag{3.20}$$

La distribución marginal de x es obtenida sumando la distribución conjunta $p(x, z)$ sobre todos los posibles estados de z:

$$\begin{aligned} p(x) &= \sum_z p(x, z) \\ &= \sum_z p(z)p(x|z) \\ &= p(z_1 = 1)p(x|z_1 = 1) + p(z_2 = 1)p(x|z_2 = 1) + \ldots + p(z_K = 1)p(x|z_K = 1) \\ &= \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \ldots + \pi_K \mathcal{N}(x|\mu_K, \Sigma_K) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned} \tag{3.21}$$

Por tanto, la distribución marginal de x es una mezcla gaussiana de la forma de la Ecuación 3.2. Si tenemos varias observaciones $x_1, \ldots, x_N$, entonces, ya que hemos representado la distribución marginal en la forma $p(x) = \sum_z p(x, z)$, se deduce que para cada punto de los datos observado x_n hay una variable latente correspondiente z_n.

Por lo tanto, hemos encontrado una formulación equivalente de la mezcla gaussiana que involucra una variable latente explícita. Puede parecer que no hemos ganado mucho al hacerlo. Sin embargo, ahora podemos trabajar con la distribución conjunta $p(x, z)$ en lugar de la distribución marginal $p(x)$, y esto conducirá a simplificaciones significativas, especialmente a través de la introducción del algoritmo EM.

Otra cantidad que desempeñará un papel importante es la probabilidad condicional de z dada x. Usaremos $\gamma(z_k)$ para denotar $p(z_k = 1|x)$, cuyo valor se puede encontrar usando el teorema de Bayes.

$$\begin{aligned} \gamma(z_k) &\equiv p(z_k = 1|x) \\ &= \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \end{aligned} \tag{3.22}$$

Veremos π_k como la probabilidad prior de $z_k = 1$, y la cantidad $\gamma(z_k)$ como la probabilidad posterior correspondiente una vez que hayamos observado x. Como veremos más adelante, $\gamma(z_k)$ también puede verse como la responsabilidad que la componente k asume para "explicar" la observación x.

3.2.2. Una vista alternativa del algoritmo EM

En esta parte presentamos una vista complementaria del algoritmo EM, que reconoce el papel clave desempeñado por las variables latentes. Discutiremos este enfoque en primer lugar en un contexto general, y luego, vamos a considerar el caso particular para las mezclas gaussianas.

El objetivo del algoritmo EM es encontrar soluciones de *maximum likelihood* para modelos con variables latentes. Denotamos el conjunto de todos los datos observados por X, en la que n-ésima fila es dada por x_n^T, y de manera similar denotamos el conjunto de todas las variables latentes por Z, donde la n-ésima fila es denotada por z_n^T. El conjunto de todos los parámetros del modelo se denota por Θ, por lo que la función de log $-$verosimilitud es dada por:

$$\ln p(X|\Theta) = \ln \left\{ \sum_Z p(X, Z|\Theta) \right\} \tag{3.23}$$

Una observación clave es que la suma de las variables latentes aparece dentro del logaritmo. La presencia de la suma evita que el logaritmo actúe directamente sobre la distribución conjunta, lo que resulta en expresiones complicadas para la solución de máxima verosimilitud.

Ahora suponga que, para cada observación en X, tenemos su valor correspondiente en la variable latente o variable oculta Z. Llamaremos a $\{X, Z\}$ el conjunto de datos completo, y nos referiremos a los datos observados reales X como incompletos. La función de verosimilitud para el conjunto de datos completo simplemente toma la forma $\ln p(X, Z|\Theta)$, y supondremos que la maximización de esta función de log $-$verosimilitud de datos completos es sencilla.

En la práctica, sin embargo, no se nos da el conjunto completo de datos $\{X, Z\}$, sino sólo los datos incompletos X. Nuestro conocimiento de los valores de las variables latentes Z viene dado únicamente por la distribución posterior $p(Z|X, \Theta)$. Como no podemos usar la función log $-$verosimilitud de datos completos, consideramos en su lugar su valor esperado bajo la distribución posterior de la variable latente, esto corresponde al paso E del algoritmo EM. En el siguiente paso M, maximizamos esta esperanza. Si la estimación actual para los parámetros se denota Θ^{old}, entonces un par de pasos sucesivos E y

M dan lugar a una estimación revisada Θ^{new}. El algoritmo se inicializa eligiendo algún valor inicial para los parámetros Θ_0.

En el paso E, usamos los valores de los parámetros actuales Θ^{old} para encontrar la distribución posterior de las variables latentes dadas por $p(Z|X, \Theta^{old})$. Luego usamos esta distribución posterior para encontrar la esperanza de la función log −verosimilitud de datos completos evaluada para algún valor de parámetro general Θ. Esta esperanza, denotada por $\mathcal{Q}(\Theta, \Theta^{old})$, es dada por:

$$\begin{aligned} \mathcal{Q}(\Theta, \Theta^{old}) &= E_{Z|X,\Theta^{old}}\left[\sum_z \ln p(X, Z|\Theta)|X, \Theta^{old}\right] \\ &= \sum_Z p(Z|X, \Theta^{old}) \ln p(X, Z|\Theta) \end{aligned} \tag{3.24}$$

En el paso M, determinamos la estimación revisada del parámetro Θ^{new} maximizando la función anterior:

$$\Theta^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \tag{3.25}$$

Tenga en cuenta que en la definición de $\mathcal{Q}(\Theta, \Theta^{old})$, el logaritmo actúa directamente sobre la distribución conjunta $p(X, Z|\Theta)$, por lo tanto la maximización del paso M correspondiente será, por suposición, tratable.

Ahora consideramos la aplicación del algoritmo EM utilizando variable latente para el caso específico de un modelo de mezcla gaussiana. Recuerde que nuestro objetivo es maximizar la función log −verosimilitud (Ecuación 3.14), que se calcula utilizando el conjunto de datos observados X, y como vimos anteriormente, esto era más difícil que para el caso de una sóla distribución gaussiana debido a la presencia de la suma sobre k que ocurre dentro del logaritmo. Supongamos, entonces, que además del conjunto de datos observado X, también se nos dieron los valores de las variables ocultas discretas correspondientes Z.

Ahora considere el problema de maximizar la probabilidad del conjunto de datos completo $\{X, Z\}$. De las ecuaciones 3.9 y 3.11, esta función de probabilidad toma la forma:

$$p(X, Z|\mu, \Sigma, \pi) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}} \tag{3.26}$$

donde z_{nk} denota la $k-$ésima componente de z_n. Tomando el logaritmo, obtenemos:

$$\ln p(X, Z|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\} \tag{3.27}$$

La comparación con la función log $-$verosimilitud (Ecuación 3.14) para los datos incompletos muestra que la suma de k y el logaritmo se han intercambiado. El logaritmo ahora actúa directamente sobre la distribución gaussiana, que es un miembro de la familia exponencial. No es sorprendente que esto conduzca a una solución mucho más simple para el problema de máxima verosimilitud. Considere primero la maximización con respecto a las medias y las covarianzas. Como z_n es un vector $K-$dimensional con todos los elementos iguales a 0, excepto un único elemento que tiene el valor 1, la función log $-$verosimilitud de datos completos es simplemente una suma de K contribuciones independientes, una para cada componente de la mezcla. Por lo tanto, la maximización con respecto a una media o una covarianza es exactamente como para un solo Gaussiano, excepto que involucra sólo el subconjunto de puntos de datos que están "asignados" a esa componente. Para la maximización con respecto a los coeficientes de mezcla, observamos que estos se acoplan para diferentes valores de k en virtud de la restricción de suma (Ecuación 3.9). Una vez más, esto puede hacerse cumplir usando un multiplicador de Lagrange como antes, y conduce al resultado

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk} \tag{3.28}$$

tal que los coeficientes de mezcla sean iguales a las fracciones de los puntos de datos asignados a las componentes correspondientes. Por lo tanto, vemos que la función log $-$verosimilitud de datos completos se puede maximizar trivialmente en forma cerrada.

Usando las ecuaciones 3.9 y 3.11 junto con el Teorema de Bayes, vemos que esta distribución posterior toma la forma de:

$$p(Z|X, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)]^{z_{nk}} \tag{3.29}$$

y, por lo tanto, factoriza sobre n para que bajo la distribución posterior, los $\{z_n\}$ sean independientes. El valor esperado de la variable indicadora z_{nk} bajo esta distribución

posterior viene dado por

$$\begin{aligned}\mathbb{E}[z_{nk}] &= \frac{\sum\limits_{z_{nk}} z_{nk}[\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)]^{z_{nk}}}{\sum\limits_{z_{nj}}[\pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)]^{z_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} = \gamma(z_{nk}) \end{aligned} \tag{3.30}$$

que es solo la responsabilidad del componente k para el punto de datos x_n. Por lo tanto, el valor esperado de la función log −verosimilitud de datos completos es por tanto, dado por:

$$\mathbb{E}_Z[\ln p(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\} \tag{3.31}$$

Ahora podemos proceder de la siguiente manera. Primero elegimos algunos valores iniciales para los parámetros μ^{old}, Σ_{old} y π^{old}, y los usamos para evaluar las responsabilidades (el paso E). Luego mantenemos las responsabilidades fijas y maximizamos (Ecuación 3.31) con respecto a μ_k, Σ_k y π_k (el paso M). Esto conduce a soluciones de forma cerrada para μ^{nuevo}, Σ^{nuevo} y π^{nuevo}. Este es precisamente el algoritmo EM para mezclas gaussianas derivadas anteriormente.

CAPÍTULO 4

APLICACIÓN DEL ALGORITMO EM

En este capítulo aplicaremos la teoría del Algoritmo Esperanza-Maximización (EM) para el caso particular de Mezclas Gaussianas, aplicado a la segmentación de imágenes.

Para este fin, usaremos imágenes sintéticas y reales que presentaremos a continuación. El programa usado para implementar estos algoritmos, fue usado el Matlab 2019b (con licencia de estudiante). En Matlab fue adoptado y programado el algoritmo EM, para visualizar la segmentación de imágenes.

Considere la imagen $X = \{x_1, x_2, \ldots, x_N\}$, la cuál se ha vectorizado de longitud N, asumimos que N es el número de píxeles dentro de la imagen. La parte oculta Z, puede ser representada como un conjunto de N etiquetas $Z = \{z_1, z_2, \ldots, z_N\}$ asociados a los N píxeles, indicando a cuál grupo pertenece (sea K la cantidad de grupos presentes en la imagen $\Omega = \{\omega_1, \ldots, \omega_K\}$). Cada etiqueta es un vector binario

$$z_j = [z_j^1, z_j^2, \ldots, z_j^K]$$

tal que $z_j^i = 1$, $j = 1, \ldots, N$, $i = 1, \ldots, K$ si el píxel x_j de X pertenece al $i-$ésimo grupo y $z_j^i = 0$ caso contrario.

Se aplica el algoritmo EM para encontrar el vector de parámetros desconocidos $\Theta = [\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_k]$ de la función de log $-$verosimilitud. En este trabajo, los valores del vector de parámetros Θ usado para inicializar el algoritmo EM, fueron

adoptados por un simple procedimiento que consiste en dividir el histograma de X en K regiones de igual longitud (Melgani, 2006):

$$\triangle = \frac{\max\limits_{j=1,\ldots,N}\{x_j\} - \min\limits_{j=1,\ldots,N}\{x_j\}}{K} \tag{4.1}$$

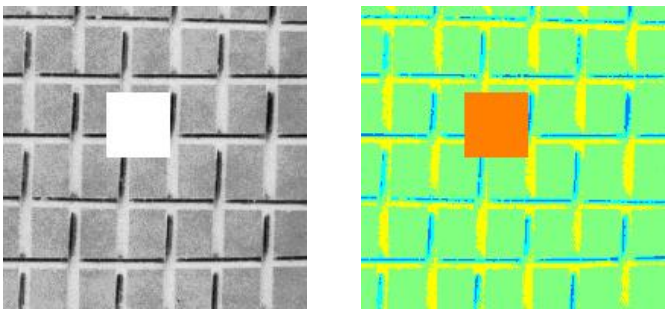
Después, las muestras comprendidas en la $i-$ésima región del histograma delimitado por los valores $x^i_{left} = \min\limits_{j=1,\ldots,N}\{x_j\} + (i-1).\triangle$ y $x^i_{right} = x^i_{left} + \triangle$ son utilizados para calcular los valores iniciales de los tres parámetros asociados con el $i-$ésimo grupo.

Después de la convergencia del algoritmo EM, las estimaciones finales de los parámetros definirán completamente las clases de datos gaussianos disponibles en X. Por último se transforma en un mapa de clasificación con un error mínimo mediante la conocida regla de decisión *maximum a posteriori probability (MAP)*. Dado que las estimaciones finales de z^i_j representan las estimaciones de las probabilidades posteriores $P(\omega_i|x_j)$, $i = 1, \ldots, K$, $j = 1, \ldots, N$, se puede asignar a cada píxel x_j de X la etiqueta del grupo óptimo, de modo que:

$$\hat{\omega} = \arg\max_{\omega_i \in \Omega} P(\omega_i|x_j) \tag{4.2}$$

4.1. Algunos resultados obtenidos

El primer ejemplo que presentamos para la segmentación de imágenes es una imagen de textura en escala de grises que se encuentra en la base de datos del albúm de Brodatz. La imagen tiene una longitud de 200×200 píxeles.



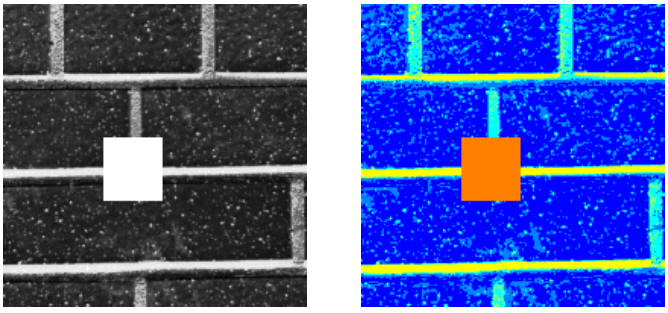
(a) Imagen original

(b) Segmentación de la Imagen

Figura 4.1: Segmentación de la Imagen 1 de Textura

En la Figura 4.1(a) observamos la imagen original en escala de grises, en la parte 4.1(b) hacemos uso del algoritmo EM con 5 grupos distintos para realizar la segmentación.

El segundo ejemplo que presentamos para la segmentación de imágenes también es una imagen de textura en escala de grises que se encuentra en la base de datos del albúm de Brodatz. La imagen tiene una longitud de 200×200 píxeles.



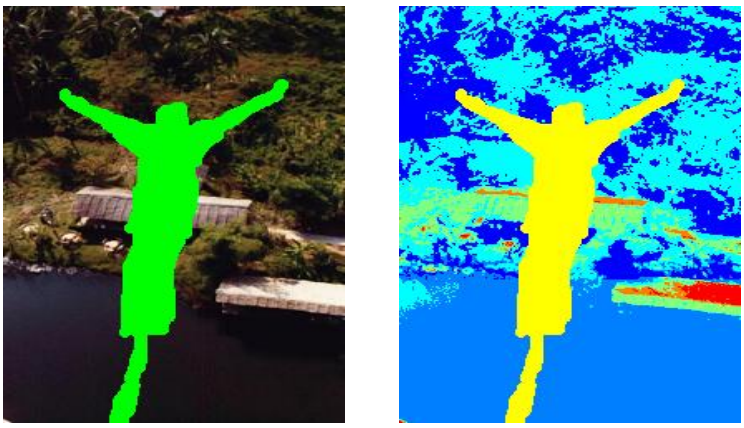
(a) Imagen original

(b) Segmentación de la Imagen

Figura 4.2: Segmentación de la Imagen 2 de Textura

En la Figura 4.2(a) observamos la imagen original en escala de grises, en la parte 4.2(b) hacemos uso del algoritmo EM con 6 grupos distintos para realizar la segmentación.

El tercer ejemplo que presentamos para la segmentación de imágenes es una imagen real en escala de RGB. La imagen tiene una longitud de 308×206 píxeles.



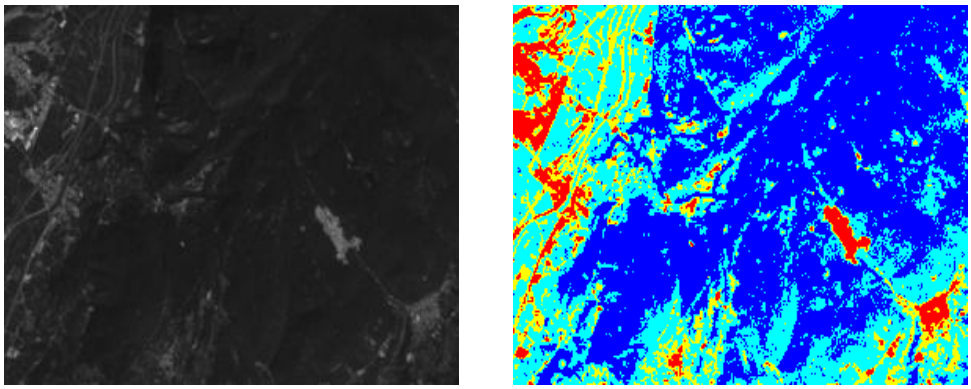
(a) Imagen original

(b) Segmentación de la Imagen

Figura 4.3: Segmentación de una Imagen real

En la Figura 4.3(a) observamos la imagen original en escala RGB, en la parte 4.3(b) hacemos uso del algoritmo EM con 7 grupos distintos para realizar la segmentación.

El último ejemplo que presentamos para la segmentación de imágenes es una imagen satelital, tomada sobre la región de Trentino, en el norte de Italia. La imagen usada en este experimento representa una sección de una escena tomada por el sensor Landsat-7 ETM+. La imagen es la banda 3 y tiene una longitud de 220×290 píxeles, adquirida en septiembre del 2000. Los principales tipos de grupos presente en esta imagen son: área urbana, agua, bosque, tierra, parto y viñedos



(a) Imagen original

(b) Segmentación de la Imagen

Figura 4.4: Segmentación de una Imagen satelital

En la Figura 4.4(a) observamos la imagen original en escala de grises, en la parte 4.4(b) hacemos uso del algoritmo EM con 4grupos distintos para realizar la segmentación.

CAPÍTULO 5

CONCLUSIONES

En este trabajo se ha presentado un estudio sobre el algoritmo *Expectation-Maximization*, se hice uso del algoritmo para el caso de mezclas Gaussianas, para la estimación de los valores de los parámetros del vector de características Θ en particular aplicado para la segmentación de imágenes.

El algoritmo EM fue utilizado para encontrar los estimadores de la función de máxima verosimilitud del modelo. Se observó que la introducción de las variables ocultas, proporcionan una ayuda significativa a la hora de la estimación de los parámetros. Además se utilizó la información de las variables ocultas para realizar un mapa de clasificación y así asignar a cada observación (datos) a un grupo determinado. En nuestro caso, utilizamos como conjunto de datos una imagen determinada, donde cada píxel es asignado a un grupo específico, lo que nos permite la segmetación de la imagen, que fue un objetivo de este trabajo.

Para evaluar el desenvolvimiento del algoritmo, fueron procesadas diferentes imágenes, las imágenes segmentadas posteriormente pueden ser utilizadas en diferentes aplicaciones, como por ejemplo, la reconstrucción de imágenes.

Se implementó el algoritmo *Expectation-Maximization* en el Software Matlab 2019b, con licencia de estudiante gratuita otorgada por un mes, para el procesamiento de las imágenes.

5.1. Sugerencias para trabajos futuros

Algunas propuestas para la continuación de la investigación sobre el algoritmo EM son:

- Estudiar la convergencia del algoritmo EM.
- Estudiar métodos alternativos para la elección de los parámetros iniciales.
- Comparación del algoritmo *K-means*, con la segmentación propuesta en este trabajo.
- Implementar el algoritmo en softwares libres, como por ejemplo R o Python.

REFERENCIAS

Baxter, R. A. (2017). Mixture model. En C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 841–844). Boston, MA: Springer US.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Borman, S. (2004). The expectation maximization algorithm-a short tutorial. *Submitted for publication*, *41*.

Chen, D. (2008). Expectation-maximization algorithm and image segmentation.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Escobar, S. L. (2007). *Algoritmos de agrupamiento global para datos mezclados*. Tesis de Master no publicada, Instituto Nacional de Astrofísica, Óptica y Electrónica, México.

Góngora, R. E. P. (2010). *Mezcla infinita de gaussianas*. Tesis de Master no publicada, Centro de Investigación en matemáticas (CIMAT), México.

Gupta, M. R., Chen, Y., y cols. (2011). Theory and use of the em algorithm. *Foundations and Trends® in Signal Processing*, *4*(3), 223–296.

Haas, S. M. (2002). *The expectation-maximization and alternating minimization algorithms.* September.

Herrero, A. G. (2015). *Algoritmos para la estimación de modelos de mezclas gaussianas.*

James, B. R. (1981). *Probabilidade: um curso em nivel intermediário*. Projeto Euclides, IMPA.

Jiménez, J. J. A. (2006). *Combinación del aprendizaje multitarea y del algoritmo EM en problemas de clasificación con datos incompletos.*

Jurcicek, F. (2014). Bayesian inference.

Mahjoub, M. A., y cols. (2012). Image segmentation by adaptive distance based on em algorithm. *arXiv preprint arXiv:1204.1629*.

Martínez, E. H. (2006). Tratamiento matricial de los datos multivariantes.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 84). M. Dekker New York.

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Melgani, F. (2006). Contextual reconstruction of cloud-contaminated multitemporal multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(2), 442–455.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, *13*(6), 47–60.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution.

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. *Método Informáticos Avanzados*, 164–174.

Rolla, L. T. (2012). *Inttrodução à probabilidade. notas de aula*.

Theodoridis, S., & Koutroumbas, C. (2009). Pattern recognition: Elsevier inc.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 448–485.

Wilson, S. E. (2015). *Methods for clustering data witn missing values*. Tesis de Master no publicada, Universiteit Leiden, The Netherlands.

Zhai, C. (2007). A note on the expectation-maximization (em) algorithm. *Course note of CS410*.

Anexos

ANEXO A

FÓRMULAS ADICIONALES

En este anexo, vamos a incluir algunas fórmulas adicionales de matrices, su traza y sus derivadas:

1. Si c es un escalar, entonces es igual a su traza, tr denota la traza:

$$tr(c) = c \tag{A.1}$$

2. Sean A y B dos matrices, tal que AB este bien definido, entonces:

$$(AB)^T = B^T A^T \tag{A.2}$$

3. Si dos matrices A y B son tales que, los productos AB y BA son ambos bien definidos, entonces:

$$tr(AB) = tr(BA) \tag{A.3}$$

4. Sea A una matriz cuadrada, $|A|$ denota el determinante de la matriz A, entonces:

$$|A^{-1}| = \frac{1}{|A|} \tag{A.4}$$

5. La traza es un operador linear: Si A y B son dos matrices; y α y β son dos escalares, entonces:

$$tr(\alpha A + \beta B) = \alpha tr(A) + \beta tr(B) \tag{A.5}$$

6. El gradiente de la traza del producto de dos matrices A y B con respecto a A es:

$$\frac{\partial}{\partial A} tr(AB) = B^T \tag{A.6}$$

7. El gradiente del logaritmo natural del determinante de A ($|A|$) es:

$$\frac{\partial}{\partial A}\ln|A| = (A^{-1})^T \tag{A.7}$$

8. Si x es un vector $l \times 1$ que no depende de A; y A es una matriz simétrica $l \times l$, entonces:

$$\frac{\partial}{\partial x}(x^T A x) = 2Ax \tag{A.8}$$

9. Sea x un vector $l \times 1$, A una matriz $l \times l$ dimensional. Puesto que $x^T Ax$ es un escalar, podemos tomar su traza, entonces se obtiene:

$$\begin{aligned} x^T Ax &= tr(x^T Ax) \text{ por la Ecuación A.1} \\ &= tr(xx^T A) \text{ por la Ecuación A.3} \\ &= tr(Axx^T) \text{ por la Ecuación A.3} \end{aligned} \tag{A.9}$$

10. **Multiplicador de Lagrange**

Los multiplicadores de Lagrange se utilizan para encontrar los puntos estacionarios de una función de varias variables sujeta a una o más restricciones. Por ejemplo, considere el problema de encontrar el máximo de una función $f(x)$, sujeta a una restricción $g(x) = 0$. Un enfoque elegante y simple para solucionar este problema, es introducir un parámetro λ llamado multiplicador de Lagrange.

Definimos la función Lagrangiana, dada por:

$$L(x, \lambda) = f(x) + \lambda g(x) \tag{A.10}$$

Para encontrar el máximo de la función $f(x)$ sujeto a la restricción $g(x) = 0$, definimos la función Lagrangiana dada por la Ecuación A.10 y luego encontramos el punto estacionario de $L(x, \lambda)$ con respecto a x y λ.

ANEXO B

ALGORITMO *K-MEANS*

El algoritmo *k-means* es uno de los algoritmos de agrupamiento *batch* más usados para tratar problemas de *clustering*. La idea principal es definir K centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada grupo y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos de un paso al siguiente. El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas (Pascual, Pla, & Sánchez, 2007).

Un *cluster* es un conjunto o grupo de datos (observaciones) que presentan propiedades comunes entre sí y diferentes a las de otros clusters. *Clustering* se define como el procedimiento de agrupar los datos en grupos de acuerdo a algunos criterios (Herrero, 2015).

K-means es un algoritmo de clasificación no supervisada, el cuál agrupa los objetos en K grupos, teniendo en cuenta sus características. El algoritmo consiste en minimizar la suma de distancias entre cada objeto y el centroide de cada grupo. Para minimizar esta distancia, se utiliza con frecuencia la distancia euclidiana.

El primer paso, en el algoritmo *k-means* es fijar el número K a priori (cantidad de grupos). Para cada grupo, definimos el centroide $\mu_i \in \mathbb{R}^D$, donde $i \in \{1, 2, \ldots, K\}$.

Dado un conjunto de observaciones $Y = \{y_1, y_2, \ldots, y_N\}$, donde $y_j \in \mathbb{R}^D$, $j \in \{1, 2, \ldots, N\}$. El algoritmo *k-means* tiene como objetivo agrupar las N observaciones en K grupos, donde cada observación pertenece al centroide del grupo más cercano. Denotemos este conjunto de grupos por:

$$S = \{S_1, S_2, \ldots, S_K\}$$

donde S_i, $i \in \{1, 2, \ldots, K\}$ es el conjunto de observaciones pertenecientes al $i-$ésimo grupo. El problema se formula de la siguiente forma:

$$J = \arg\min_{S} \sum_{i=1}^{K} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{B.1}$$

B.1. Pasos del algoritmo *k-means*

1. **Inicialización:** Se eligen aleatoriamente K observaciones y se utilizan como centroides iniciales $\mu^{(0)} = \{\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_K^{(0)}\}$ de los grupos.

2. **Asignación:** Se asigna cada observación a un grupo, donde la suma de distanca euclidiana sea la mínima

   $$S_i^{(n)} = \{x_j : \|x_j - \mu_i^{(n)}\|^2 \leq \|x_j - \mu_z^{(n)}\|^2, \forall z, 1 \leq z \leq K\} \tag{B.2}$$

   En cada iteración n, cada observación x_j se asigna a un único grupo $S_i^{(n)}$. Es decir, se calcula la distancia euclidiana de cada observación a cada centroide y cada observación será asignada al grupo S_i más cercano.

3. **Actualización:** Se recalculan los centroides (las medias) de los grupos, teniendo en cuenta las observación asignadas a cada grupo, dado por:

   $$\mu_i^{(n+1)} = \frac{1}{|S_i^{(n)}|} \sum_{x_j \in S_i^{(n)}} x_j \tag{B.3}$$

   donde $|S_i^{(n)}|$ es el número de observaciones en el $i-$ésimo grupo.

Una vez calculados estos nuevos centroides, se realiza de nuevo la asignación de las observaciones a los centroides más cercanos. Los pasos de asignación y actualización se repiten iterativamente, hasta que no se produzcan cambios de posición de los centroides ($\mu_i^{(n+1)} = \mu_i^{(n)}$).