ADVANCED MACHINE LEARNING

ÍNDICE

1 MÉTODOS DE INFERENCIA

La estadística inferencial ofrece dos principales paradigmas para estimar parámetros desconocidos a partir de datos: el enfoque frecuentista y el enfoque bayesiano. Ambos persiguen el mismo objetivo fundamental pero difieren en sus fundamentos filosóficos y metodológicos.

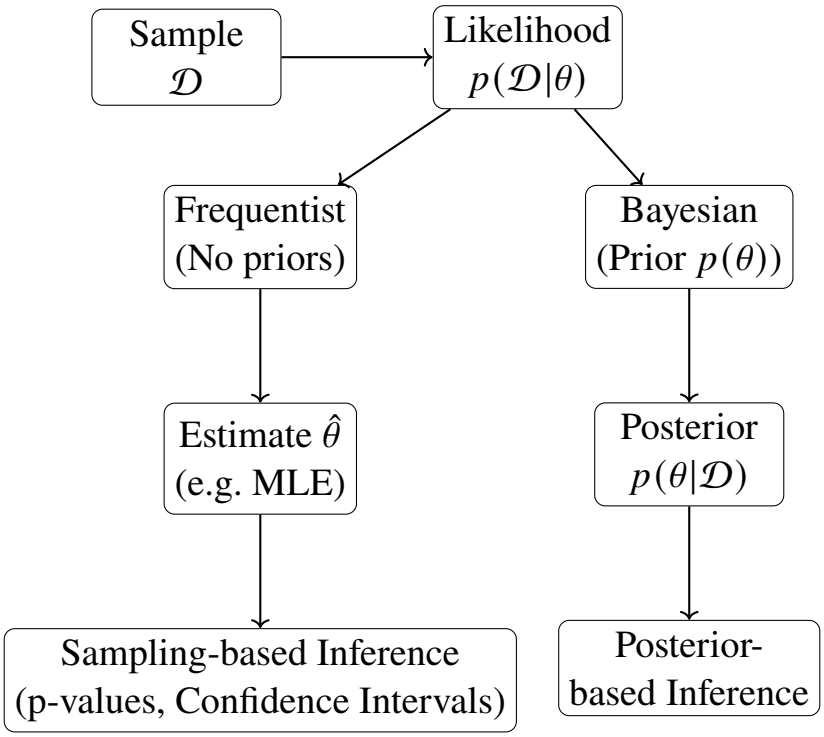


Figura 1: Comparison between Frequentist and Bayesian inference approaches.

La Figura 1 ilustra las diferencias clave entre estos enfoques. El método frecuentista considera los parámetros como cantidades fijas pero desconocidas, mientras que el enfoque bayesiano los trata como variables aleatorias con distribuciones de probabilidad.

Máxima verosimilitud (MLE) es la herramienta principal en la inferencia frecuentista, que busca el valor del parámetro que maximiza la función de verosimilitud dada la muestra observada:

$$x_1, x_2, ..., x_n \rightarrow \theta$$

$$p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

En contraste, la inferencia bayesiana utiliza el teorema de Bayes para actualizar la probabilidad de una hipótesis conforme se dispone de más evidencia. Comienza con una distribución a priori que representa nuestras creencias sobre los parámetros antes de observar los datos. Después de observar los datos, actualizamos estas creencias para obtener una distribución posterior, que combina la información previa y la evidencia observada.

La principal diferencia entre ambos enfoques radica en el tratamiento de los parámetros y la incorporación de información previa:

- Frecuentista (sin priors): parámetros fijos y desconocidos, estimaciones puntuales.

- Bayesiano (con priors): parámetros como variables aleatorias con distribuciones de probabilidad completas.

1.1 Inferencia Bayesiana

En estadística bayesiana, los parámetros θ se tratan como variables aleatorias con sus propias distribuciones de probabilidad. Se comienza con una **distribución prior** $p(\theta)$ que representa las creencias sobre θ antes de ver los datos. Después de observar los datos x, actualizamos nuestras creencias utilizando el teorema de Bayes (1).

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{1}$$
$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

El posterior $p(\theta|x)$ combina el conocimiento previo y la evidencia observada.

$p(x|\theta)$ $\rightarrow$ La **verosimilitud** mide la probabilidad de los datos dados los parámetros, permite cuantificar el nivel de relación entre los datos y los parámetros; cuantifica que tan buenas o malas son las hipótesis.

$p(\theta)$ $\rightarrow$ El **prior** refleja el conocimiento previo sobre los parámetros antes de observar los datos.

$p(x)$ $\rightarrow$ La **evidencia** es una constante de normalización que asegura que la distribución posterior sea válida.

$p(\theta|x)$ $\rightarrow$ El **posterior** es la distribución actualizada de los parámetros después de observar los datos.

Para que el posterior sea una solución cerrada, el prior y la verosimilitud deben cumplir ciertas condiciones. En muy pocos casos se encuentra que el prior y el likelihood sean gaussianas, por lo que se utilizan métodos numéricos para aproximar el posterior.

Ejercicio 1. Dada una muestra $x_1, x_2, ..., x_n$ que proviene de una distribución Gaussiana con media μ y varianza σ^2:

A. Calcular el estimador de máxima verosimilitud para μ.

B. Suponiendo σ^2 conocido, calcular la distribución posterior para μ dado un prior Gaussiano $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$.

Solución:

A. Para calcular el estimador de máxima verosimilitud, primero planteamos la verosimilitud $p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\mu, \sigma^2)$ dada la muestra observada. Asumimos que las observaciones son independientes e idénticamente distribuidas (i.i.d), por lo que la verosimilitud se puede expresar como el producto de las verosimilitudes individuales:

$$p(AB) = p(A)p(B)$$

Sabiendo que $\log(AB) = \log(A) + \log(B)$, y dado que es una función monótonamente creciente, podemos maximizar la log-verosimilitud en lugar de la verosimilitud directamente, lo que simplifica los cálculos:

$$\log p(x_1, x_2, ..., x_n|\mu) = \sum_{i=1}^{N} \log p(x_i|\mu)$$

$$\log p(x_1, x_2, ..., x_n|\mu) = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right]$$

$$\log p(x_1, x_2, ..., x_n|\mu) = \sum_{i=1}^{N} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

Derivando e igualando a cero para encontrar el valor de μ que maximiza la log-verosimilitud:

$$\frac{d}{d\mu} \log p(x_1, x_2, ..., x_n|\mu) = \sum_{i=1}^{N} \left[0 + \frac{1}{\sigma^2}(x_i - \mu) \right] = 0$$

$$\sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\sum_{i=1}^{N} x_i - N\mu = 0$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Por lo tanto, el estimador de máxima verosimilitud para μ es la media muestral. El promedio es el estimador de máxima verosimilitud para la media de una distribución Gaussiana.

B. Para calcular la distribución posterior $p(\mu|x)$ usando el teorema de Bayes (1), necesitamos la verosimilitud $p(x|\mu)$ y el prior $p(\mu)$.

La verosimilitud para una muestra i.i.d de una distribución Gaussiana es:

$$p(x|\mu) = \prod_{i=1}^{N} p(x_i|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$p(x|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right)$$

El prior Gaussiano es:

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

Sabiendo que μ es una variable aleatoria, calculamos la distribución posterior $p(\mu|x)$ combinando la verosimilitud y el prior. La expresión para la distribución posterior paso a paso es:

$$p(\mu|x) \propto p(x|\mu)p(\mu)$$

$$p(\mu|x) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

$$p(\mu|x) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Expandiendo los términos en el exponente:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} = -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}x_i^2 - 2\mu\sum_{i=1}^{N}x_i + N\mu^2\right) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

Agrupando los términos en μ^2, μ y los términos constantes:

$$-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 + \left(\frac{\sum_{i=1}^{N}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu + \text{constantes}$$

Completando el cuadrado para expresar el exponente en la forma estándar de una distribución Gaussiana:

$$-\frac{1}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu - \frac{\frac{\sum_{i=1}^{N}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2 + \text{constantes}$$

De esta forma, identificamos que la distribución posterior $p(\mu|x)$ es una distribución Gaussiana con media y varianza dadas por:

$$\mu_n = \frac{\frac{\mu_0}{\sigma_0^2}+\frac{N\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}}$$

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

Donde $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$ es la media muestral de los datos observados.

Alternativamente, podemos calcular la distribución posterior utilizando la forma logarítmica:

$$\log p(\mu|x) = \log p(x|\mu) + \log p(\mu) - \log p(x)$$

Donde $p(x)$ es la constante de normalización:

$$p(x) = \int p(x|\mu)p(\mu)d\mu$$

Desarrollando el logaritmo del posterior:

$$\log p(\mu|x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 + \text{cte.}$$

Expandiendo los términos cuadráticos:

$$\log p(\mu|x) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu + \mu^2) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) + \text{cte.}$$

Agrupando los términos en μ^2 y μ:

$$\log p(\mu|x) = -\frac{1}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 + \left(\frac{\sum_{i=1}^{N}x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu + \text{cte.}$$

Para identificar la distribución posterior como una Gaussiana, comparamos con la forma logarítmica de una distribución normal:

$$\log \mathcal{N}(\bar{\mu}, \bar{\sigma^2}) = -\frac{1}{2}\log(2\pi\bar{\sigma^2}) - \frac{(\mu-\bar{\mu})^2}{2\bar{\sigma^2}}$$

Expandiendo:

$$\log \mathcal{N}(\bar{\mu}, \bar{\sigma^2}) = -\frac{1}{2}\log(2\pi\bar{\sigma^2}) - \frac{1}{2\bar{\sigma^2}}(\mu^2 - 2\mu\bar{\mu} + \bar{\mu}^2)$$

$$\log \mathcal{N}(\bar{\mu}, \bar{\sigma^2}) = -\frac{1}{2\bar{\sigma^2}}\mu^2 + \frac{\bar{\mu}}{\bar{\sigma^2}}\mu + \text{cte.}$$

Comparando los coeficientes con nuestra expresión del posterior, obtenemos:

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

$$\bar{\mu} = \bar{\sigma^2}\left(\frac{\sum_{i=1}^{N}x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) = \frac{\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Por lo tanto, la distribución posterior $p(\mu|x)$ es una distribución Gaussiana:

$$p(\mu|x) = \mathcal{N}(\bar{\mu}, \bar{\sigma^2})$$

Esta expresión muestra cómo el posterior combina la información del prior con la de los datos observados, ponderando cada fuente según su precisión relativa. A mayor cantidad de datos (N más grande) o menor varianza en los datos (σ^2 más pequeña), mayor será el peso asignado a la evidencia empírica. Por otro lado, a menor incertidumbre en el prior (σ_0^2 más pequeña), mayor será la influencia del conocimiento previo en la estimación final.

La Figura 2 ilustra este proceso de actualización bayesiana para un caso específico. En ella podemos observar cómo la distribución prior (línea azul discontinua) con media $\mu_0 = 2$ y varianza $\sigma_0^2 = 1.5^2$ se combina con la verosimilitud de los datos (línea roja punteada) que tienen media muestral $\bar{x} = 4$ (con N=10 y $\sigma^2 = 1$) para producir la distribución posterior (línea verde continua) con media $\mu_n = 3.7$ y varianza $\sigma_n^2 = 0.095$.

Este ejemplo ilustra principios clave de la inferencia bayesiana. La media posterior ($\mu_n = 3.7$) se encuentra entre la media del prior ($\mu_0 = 2$) y la media muestral ($\bar{x} = 4$), reflejando la influencia combinada del conocimiento previo y los datos. Además, la varianza posterior ($\sigma_n^2 = 0.095$) es menor que las varianzas del prior y la verosimilitud, mostrando cómo el enfoque bayesiano reduce la incertidumbre al integrar múltiples fuentes de información.

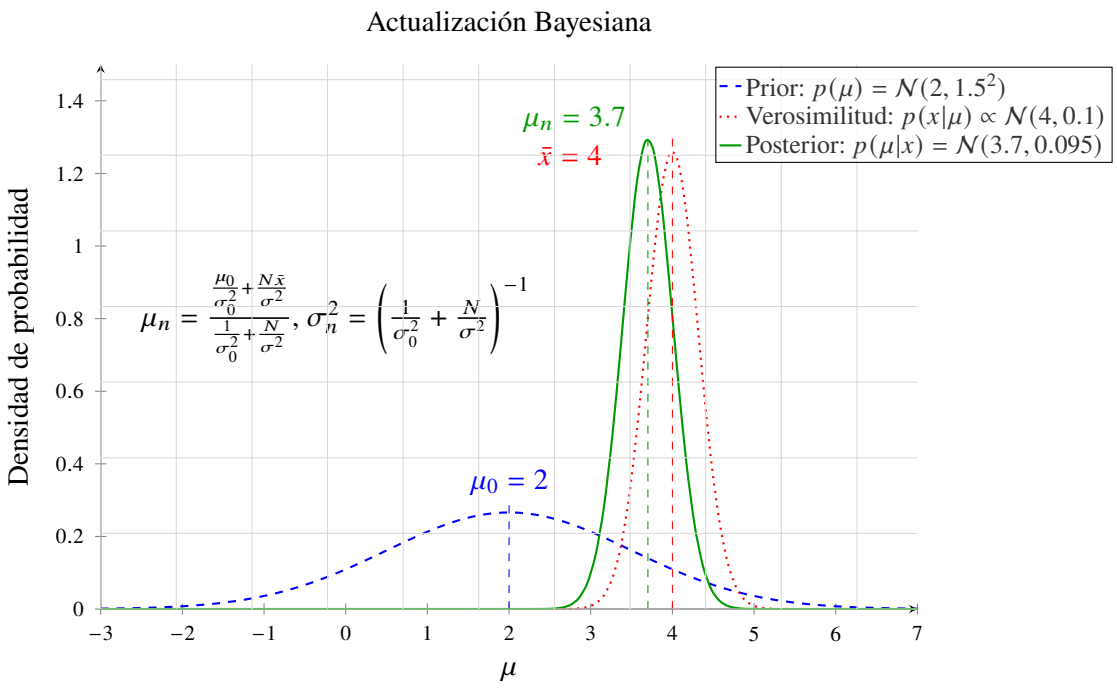
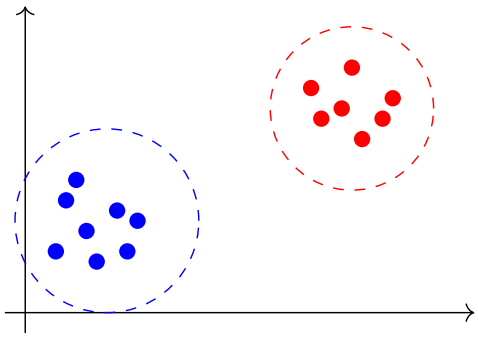


Figura 2: Actualización bayesiana de la distribución de μ. Ilustra cómo el conocimiento previo (prior) se combina con la evidencia (verosimilitud) para obtener la distribución posterior. En este ejemplo, asumimos un prior $\mathcal{N}(2, 1.5^2)$, datos con media muestral $\bar{x} = 4$ ($N = 10$, $\sigma^2 = 1$), resultando en un posterior $\mathcal{N}(3.7, 0.095)$ que se desplaza hacia la evidencia pero con menor varianza que ambas distribuciones originales.

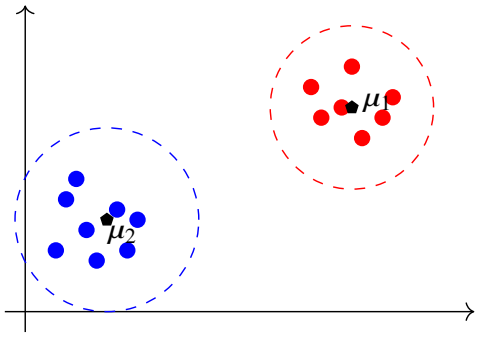
2 CLUSTERING

En este módulo, estudiamos modelos de aprendizaje no supervisado. Esto significa que nuestros conjuntos de datos no contienen etiquetas. Por lo tanto, tenemos un conjunto de datos que consiste únicamente en atributos de entrada $[x_1, x_2, ..., x_N]$, donde cada $x_i \in \mathbb{R}^P$. El objetivo es asignar cada muestra x a uno de los K clústeres.



2.1 Fundamentos de K-means

Los grupos se eligen de manera que la distancia entre puntos dentro del mismo grupo sea menor que la distancia entre puntos en diferentes grupos. Para lograr esto, es útil definir el vector $\boldsymbol{\mu}_k \in \mathbb{R}^P$, que representa el centro del k-ésimo grupo. De esta manera, fijamos un conjunto de vectores $\boldsymbol{\mu}_k{}_{k=1}^{K}$ tal que la suma de las distancias desde cada punto a su μ_k más cercano sea lo más pequeña posible.



Para cada muestra x_n, definimos un vector binario r_n, compuesto por K valores r_{nk}, donde $r_{nk} = 1$ si la instancia x_n pertenece al grupo k, y $r_{nk} = 0$ en caso contrario. Cabe destacar que cada muestra x_n solo puede pertenecer a uno de los K grupos. Por ejemplo, si $K = 4$ grupos y la instancia x_n pertenece al grupo 3, el vector r_n será $r_n = [0, 0, 1, 0]$.

El objetivo es encontrar los valores de r_{nk} y $\boldsymbol{\mu}_k$ que minimicen la distancia entre cada punto x_n y su centro correspondiente μ_k. Matemáticamente, esto se expresa como la minimización de la siguiente función objetivo:

$$J(\boldsymbol{\mu}, \boldsymbol{r}) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$$

Los valores de r_{nk} y $\boldsymbol{\mu}_k$ se pueden estimar utilizando un enfoque iterativo. El algoritmo K-means sigue los siguientes pasos:

Algorithm 1: K-means

Seleccionar K instancias como centroides iniciales $\boldsymbol{\mu}_k$;
repeat
  for $i \leftarrow 1, N$ **do**
    Para cada $\boldsymbol{x}_n$, definir su vector $\boldsymbol{r}_n$ con los elementos r_{nk}:
    $$r_{nk} = \begin{cases} 1, & \text{si la distancia entre } \boldsymbol{x}_n \text{ y } \boldsymbol{\mu}_k \text{ es la menor} \\ 0, & \text{en caso contrario.} \end{cases}$$
  end
  Actualizar el valor de los centroides: $\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$;
until *Se cumpla algún criterio de convergencia*;

La función objetivo $J(\boldsymbol{\mu}, \boldsymbol{r})$ actúa como una **función de distorsión** que cuantifica la suma total de distancias cuadradas entre los puntos de datos y sus centroides asignados. Es importante destacar que el algoritmo K-means puede interpretarse como un método de **descenso coordinado**, donde se alternan dos pasos: la actualización de las asignaciones $\boldsymbol{r}$ mientras se mantienen fijos los centroides $\boldsymbol{\mu}$, y la actualización de los centroides $\boldsymbol{\mu}$ mientras se mantienen fijas las asignaciones $\boldsymbol{r}$. Durante la ejecución del algoritmo, el valor de J decrece monótonamente hasta alcanzar la convergencia; sin embargo, se debe tener en cuenta que J es una **función no convexa**, lo que implica que K-means puede converger a un mínimo local en lugar del óptimo global. Para mitigar este problema, una estrategia práctica comúnmente empleada consiste en ejecutar el algoritmo varias veces con diferentes inicializaciones aleatorias de centroides y seleccionar la agrupación que presente el menor valor de distorsión $J(\boldsymbol{\mu}, \boldsymbol{r})$.

Una importante limitación del algoritmo K-means es que solo puede capturar grupos con formas circulares (o hiperesféricas en dimensiones superiores). Esto se debe a que la función objetivo se basa únicamente en la distancia euclidiana entre los puntos y los centroides, lo que favorece agrupaciones con forma circular. Sin embargo, en aplicaciones reales, los datos pueden presentar estructuras más complejas con formas elongadas, no lineales o con densidades variables, donde K-means no resulta adecuado. Esta limitación es precisamente una de las motivaciones para explorar modelos más flexibles como los GMM. Otro inconveniente es que la estimación que realiza K-means es **determinista**, lo que significa que cada punto se asigna de manera definitiva a un solo clúster, constituyendo lo que se conoce como **clasificación dura**. En situaciones donde los datos presentan solapamientos entre grupos o incertidumbre en las asignaciones, este enfoque puede no ser el más adecuado. En ciertos contextos analíticos es necesaria una **estimación suave** de las asignaciones, donde cada punto tenga una probabilidad de pertenecer a cada uno de los K clústeres predefinidos en lugar de una asignación binaria y definitiva. Este enfoque probabilístico permite capturar la incertidumbre inherente en la asignación de puntos que se encuentran en las fronteras entre clústeres o en regiones de solapamiento. Este es otro aspecto que los GMM abordan de manera más efectiva, como veremos a continuación.

2.2 Modelos de Mezcla Gaussiana (GMM)

Los Modelos de Mezcla Gaussiana (GMM) son una extensión del algoritmo K-means. En lugar de asignar cada punto a un solo clúster, GMM permite que cada punto tenga una probabilidad de pertenecer a cada clúster. Esto se logra modelando los datos como una mezcla de varias distribuciones gaussianas.

Cada clúster k se representa mediante una distribución gaussiana con media $\boldsymbol{\mu}_k$ y covarianza $\boldsymbol{\Sigma}_k$. La probabilidad de que un punto x_n pertenezca al clúster k se calcula utilizando la función de densidad de la distribución gaussiana:

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^P |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right),$$