Un modelo de agrupamiento muy útil

Gaussian Mixture Models
EM Algorithm

Edwin Santiago Alférez B.

Departamento de Ingeniería de Sistemas
Facultad de Ingeniería
Pontificia Universidad Javeriana

Contenido

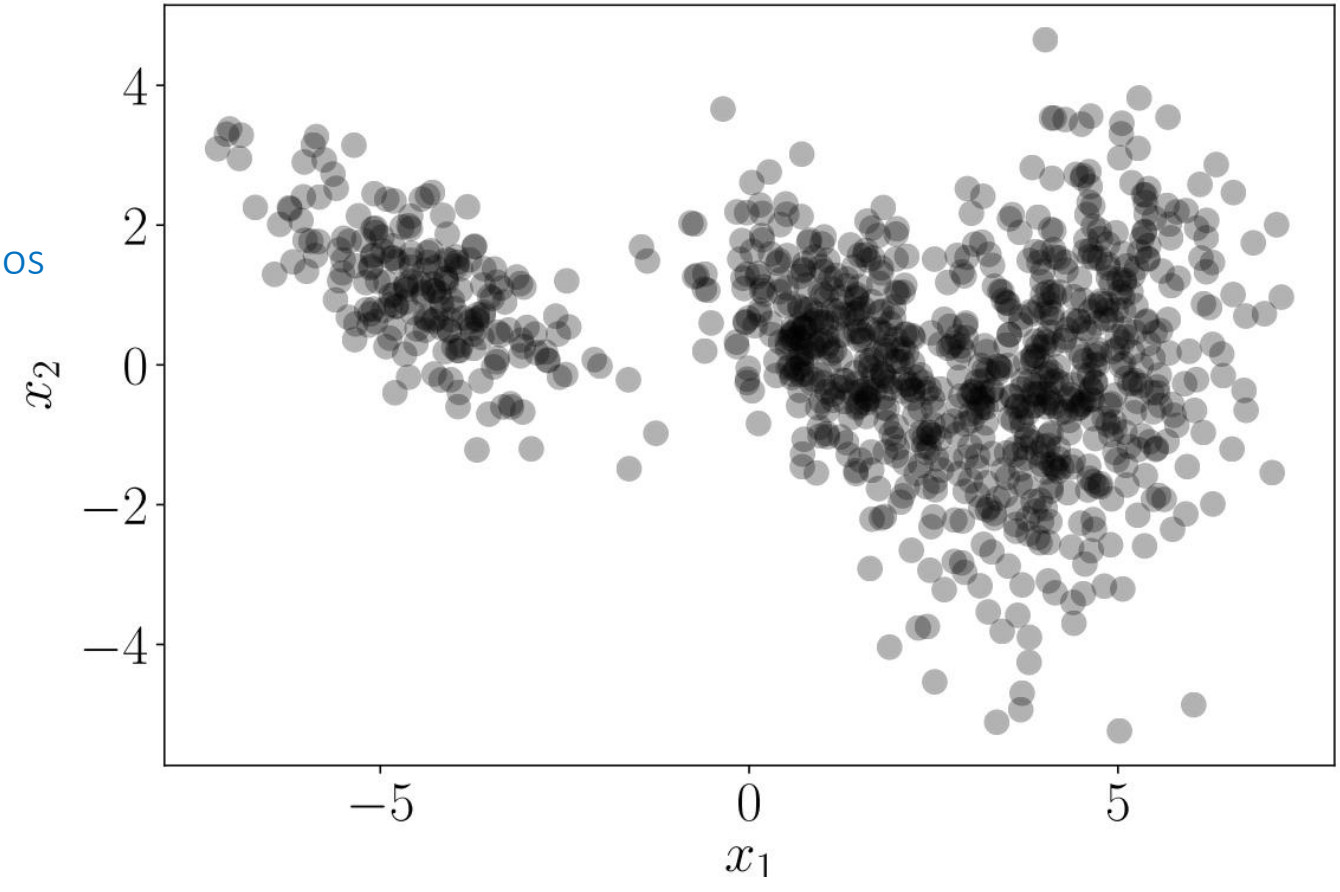
- ✓Modelos de mezclas
- ✓Aprendizaje de los parámetros
- ✓EM algorithm
- ✓Ejemplos ilustrativos

¿Cómo representamos estos datos?

¿Una distribución normal?

¿Cómo estimamos los parámetros?

MLE..



Modelos de mezcla

Los modelos de mezcla se pueden utilizar para describir una distribución $p(x)$ mediante una combinación convexa de K distribuciones simples

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$$

$$0 \leqslant \pi_k \leqslant 1\,, \quad \sum_{k=1}^{K} \pi_k = 1$$

donde los componentes p_k son miembros de una familia de distribuciones básicas, por ejemplo, gaussianas, bernoullis o gammas, y los π_k son pesos de mezcla.

Gaussian Mixture Model (GMM)

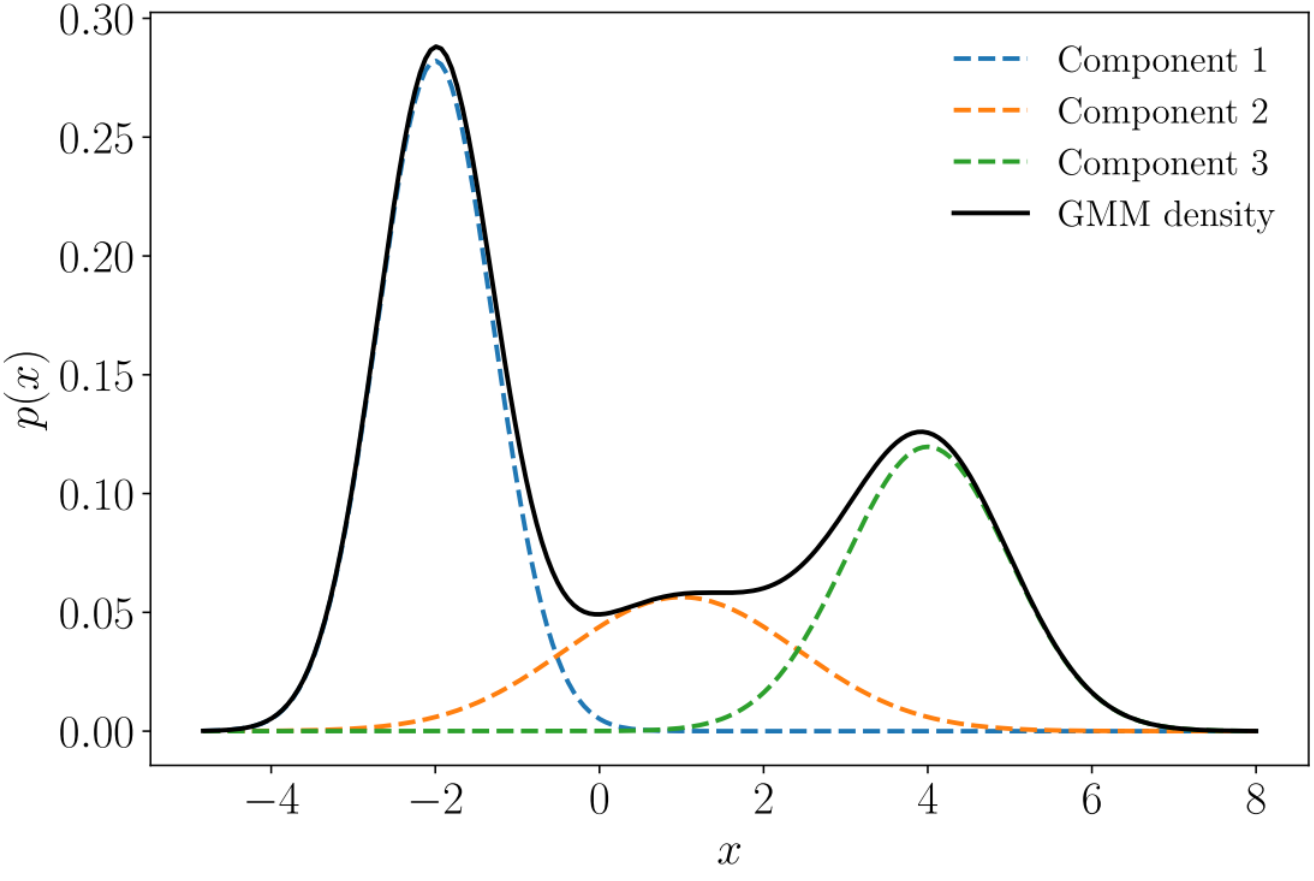
Un modelo de mezcla gaussiana es un modelo de densidad donde combinamos un número finito de K distribuciones gaussianas $N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ de modo que

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$0 \leqslant \pi_k \leqslant 1\,, \quad \sum_{k=1}^{K} \pi_k = 1\,,$$

donde definimos $\theta := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k : k = 1, \ldots, K\}$ como la colección de todos los parámetros del modelo.

Gaussian Mixture Model (GMM)



$$p(x \mid \boldsymbol{\theta}) = 0.5\mathcal{N}\left(x \mid -2, \tfrac{1}{2}\right) + 0.2\mathcal{N}(x \mid 1, 2) + 0.3\mathcal{N}(x \mid 4, 1)$$

Aprendizaje de parámetros a través de máxima verosimilitud

- Supongamos que se nos da un conjunto de datos $X = \{x_1, \ldots, x_N\}$, donde x_n, $n = 1, \ldots, N$, se extraen (i.i.d.) de una distribución desconocida $p(\boldsymbol{x})$.

- Nuestro objetivo es encontrar una buena aproximación o representación de esta distribución desconocida $p(\boldsymbol{x})$ por medio de un GMM con K componentes de mezcla.

- Los parámetros del GMM son las K medias $\boldsymbol{\mu}_k$, las covarianzas $\boldsymbol{\Sigma}_k$ y los pesos de mezcla π_k. Resumimos todos estos parámetros en $\theta := \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k : k = 1, \ldots, K\}$.

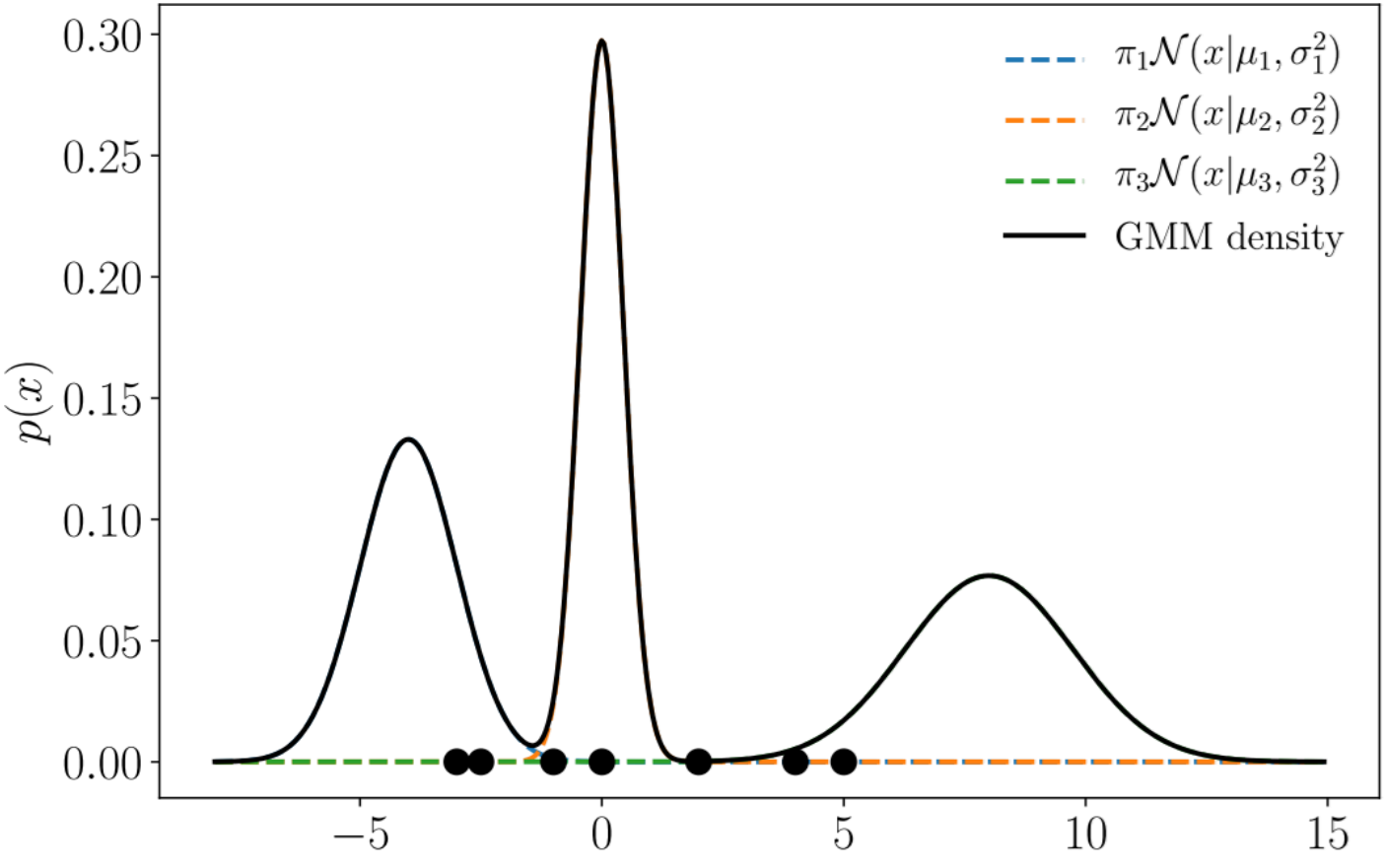
Ejemplo ilustrado: configuración inicial

Consideramos un conjunto de datos unidimensional $X = \{-3, -2.5, -1, 0, 2, 4, 5\}$ que consta de siete puntos de datos y deseamos encontrar un GMM con $K = 3$ componentes que modele la densidad de los datos. Inicializamos los componentes de la mezcla como

$$p_1(x) = \mathcal{N}(x \mid -4,1)$$
$$p_2(x) = \mathcal{N}(x \mid 0, 0.2)$$
$$p_3(x) = \mathcal{N}(x \mid 8, 3)$$

y les asignamos pesos iguales $\pi_1 = \pi_2 = \pi_3 = 1$

Ejemplo ilustrado: configuración inicial



Aprendizaje de parámetros a través de MLE

- Comenzamos calculando la verosimilitud, es decir, la distribución predictiva de los datos de entrenamiento dados los parámetros. Nuestra suposición i.i.d. conduce a la verosimilitud factorizada:

$$p(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\theta}), \quad p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

donde cada término $p(x_n \mid \theta)$ de probabilidad individual es una densidad de mezcla gaussiana.

- Entonces, el logaritmo de la verosimilitud es :

$$\log p(\mathcal{X} \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \underbrace{\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}_{=:\mathcal{L}}$$

Aprendizaje de parámetros a través de MLE

- Nuestro objetivo es encontrar parámetros θ^*_{ML} que maximicen el logaritmo de la verosimilitud L
- Nuestro procedimiento "normal" sería calcular el gradiente $dL/d\theta$ del log de la verosimilitud con respecto a los parámetros del modelo θ, establecerlo en 0 y resolver para θ.
- Sin embargo, a diferencia otros problemas de MLE, no podemos obtener una solución de forma cerrada.
- Sin embargo, podemos explotar un esquema iterativo para encontrar buenos parámetros de modelo θ_{ML}, lo que resultará en el algoritmo EM para GMM. La idea clave es actualizar un parámetro del modelo a la vez mientras se mantienen los demás fijos.

Aprendizaje de parámetros a través de MLE

$$\log p(\mathcal{X} \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \underbrace{\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{=:\mathcal{L}}$$

Cualquier óptimo local de una función exhibe la propiedad de que su gradiente con respecto a los parámetros debe desaparecer (condición necesaria).

Entonces, obtenemos las siguientes condiciones necesarias cuando optimizamos el log de la verosimilitud con respecto a los parámetros GMM $\mu_k, \Sigma_k \pi_k$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top,$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0},$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \pi_k} = 0.$$

Responsabilidades

$$r_{nk} := \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

La responsabilidad r_{nk} de la k-ésima componente de la mezcla para el punto de datos x_n es proporcional a la probabilidad

$$p(\boldsymbol{x}_n \mid \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

de la componente de la mezcla dado el punto de datos.

Por lo tanto, las componentes de mezcla tienen una alta responsabilidad para un punto de datos cuando éste punto puede ser una muestra plausible de esa componente de mezcla.

- Tenga en cuenta que $\boldsymbol{r}_n := [r_{n1}, \ldots, r_{nK}]^\top \in \mathbb{R}^K$ es un vector de probabilidad (normalizado), es decir, $\sum_k r_{nk} = 1$ with $r_{nk} \geqslant 0$.
- Este vector de probabilidad distribuye la masa de probabilidad entre los K componentes de mezcla, y podemos pensar en $\boldsymbol{r}_n$ como una "asignación suave" de $\boldsymbol{x}_n$ a las K componentes de la mezcla.
- La responsabilidad r_{nk} representa la probabilidad de que x_n haya sido generado por la k-ésima componente de mezcla.

EM Algorithm

En el ejemplo ilustrado del modelo de mezcla gaussiana, elegimos valores iniciales para μ_k, Σ_k, π_k y alternamos hasta la convergencia entre

- ✓ **E-step**: Evalua las responsabilidades r_{nk} (probabilidad posterior de que el punto de datos n pertenezca a la mezcla k).
- ✓ **M-step**: Utiliza las responsabilidades actualizadas para volver a estimar los parámetros μ_k, Σ_k, π_k.

- Cada paso en el algoritmo EM aumenta el log de la verosimilitud.
- Para la convergencia, podemos comprobar el log de la verosimilitud de los parámetros directamente.

EM Algorithm

1. Inicializar μ_k, Σ_k, π_k.
2. E-step: Evaluar responsabilidades r_{nk} para cada punto de datos $\boldsymbol{x}_n$ usando los parámetros actuales π_k, μ_k, Σ_k:

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$$

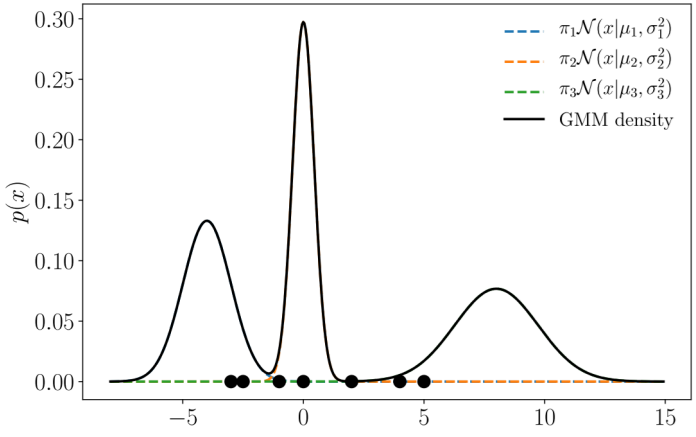
3. M-step: Se re-estiman los parámetros π_k, μ_k, Σ_k utilizando las responsabilidades actuales r_{nk} (del E-step):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n \,,$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \,,$$

$$\pi_k = \frac{N_k}{N} \,.$$

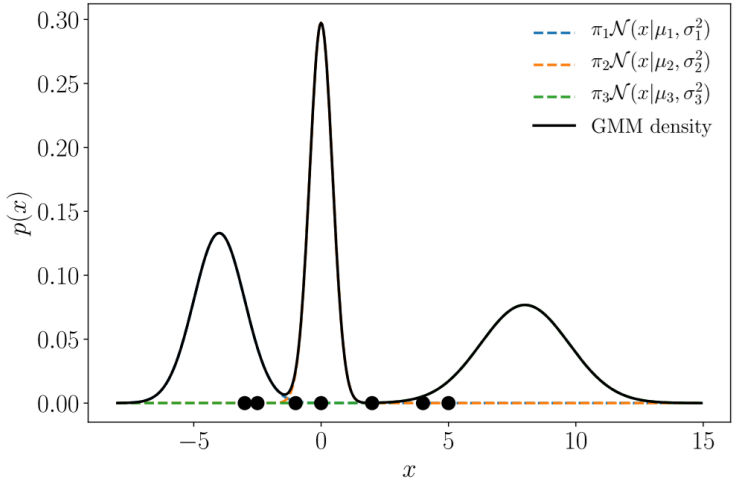
Ejemplo ilustrado: configuración inicial



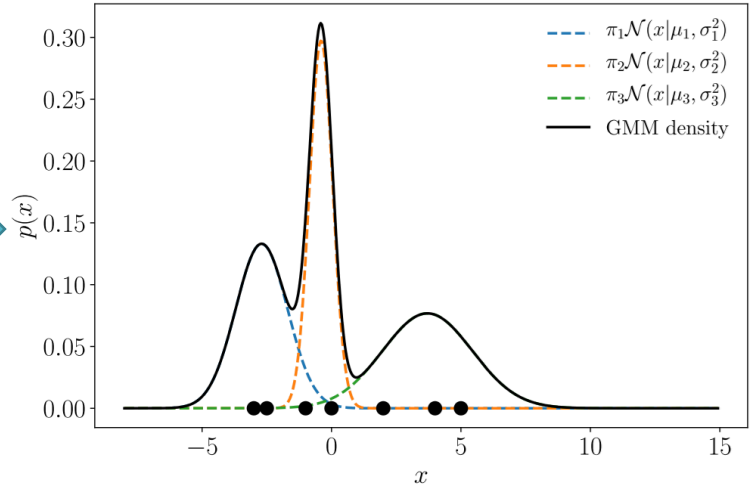
$$r_{nk} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 0.057 & 0.943 & 0.0 \\ 0.001 & 0.999 & 0.0 \\ 0.0 & 0.066 & 0.934 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \in \mathbb{R}^{N\times K}$$

- ✓ Aquí la n-ésima fila nos dice las responsabilidades de todos los componentes de la mezcla para x_n.
- ✓ La suma de todas las K responsabilidades para un punto de datos (suma de cada fila) es 1.
- ✓ La k-ésima columna nos da una visión general de la responsabilidad de la k-ésima componente de mezcla.
- ✓ Podemos ver que la tercera componente de la mezcla (tercera columna) no es responsable de ninguno de los primeros cuatro puntos de datos, sino que asume mucha responsabilidad de los puntos de datos restantes.
- ✓ La suma de todas las entradas de una columna nos da los valores N_k, es decir, la responsabilidad total de la k-ésima componente de mezcla.
- ✓ En este ejemplo, obtenemos $N_1 = 2.058, N_2 = 2.008, N_3 = 2.934$.

Ejemplo ilustrado: actualizaciones de las medias

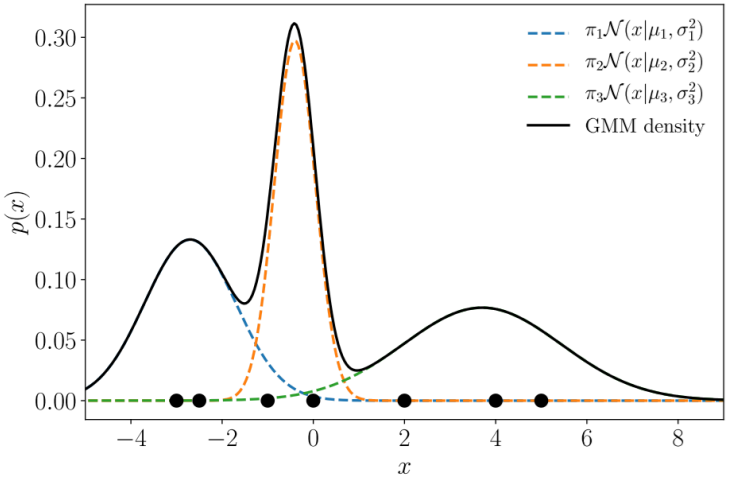


$$\mu_1 : -4 \rightarrow -2.7$$
$$\mu_2 : 0 \rightarrow -0.4$$
$$\mu_3 : 8 \rightarrow 3.7$$

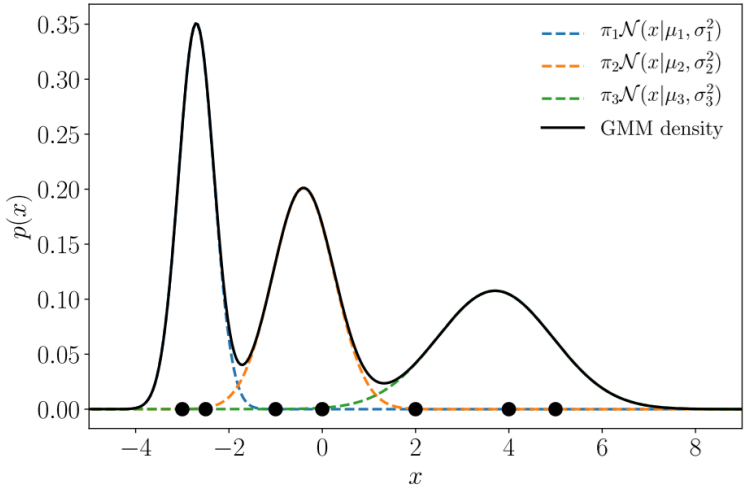


Aquí vemos que las medias de la primera y tercera componentes de mezclas se mueven hacia el régimen de los datos, mientras que la media de la segunda componente no cambia tan dramáticamente.

Ejemplo ilustrado: actualizaciones de varianzas

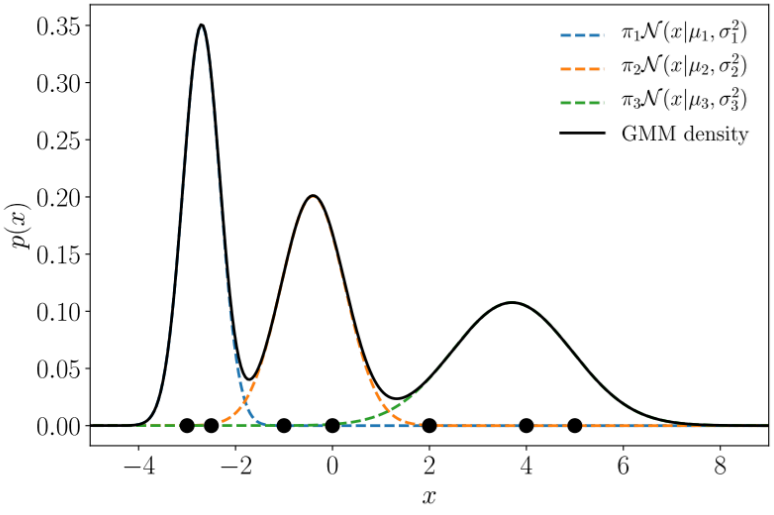


$$\sigma_1^2 : 1 \rightarrow 0.14$$
$$\sigma_2^2 : 0.2 \rightarrow 0.44$$
$$\sigma_3^2 : 3 \rightarrow 1.53$$

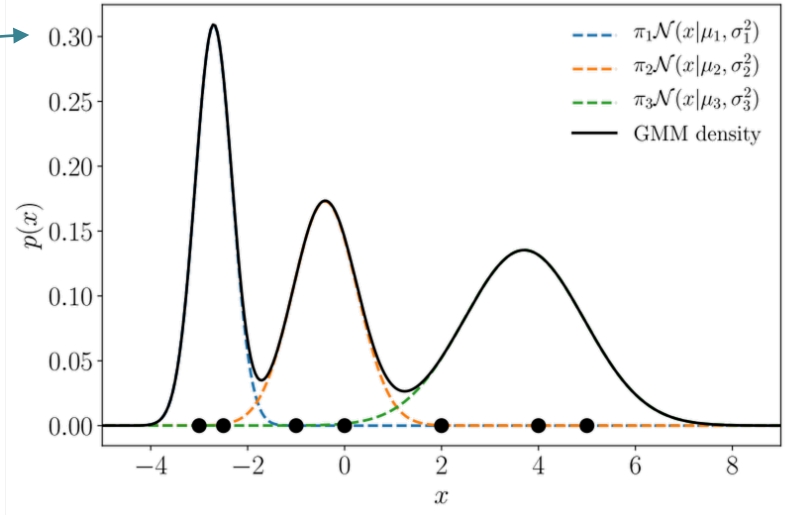


Aquí vemos que las varianzas de la primera y tercera componentes se reducen significativamente, mientras que la varianza de la segunda componente aumenta ligeramente.

Ejemplo ilustrado: actualizaciones de los pesos

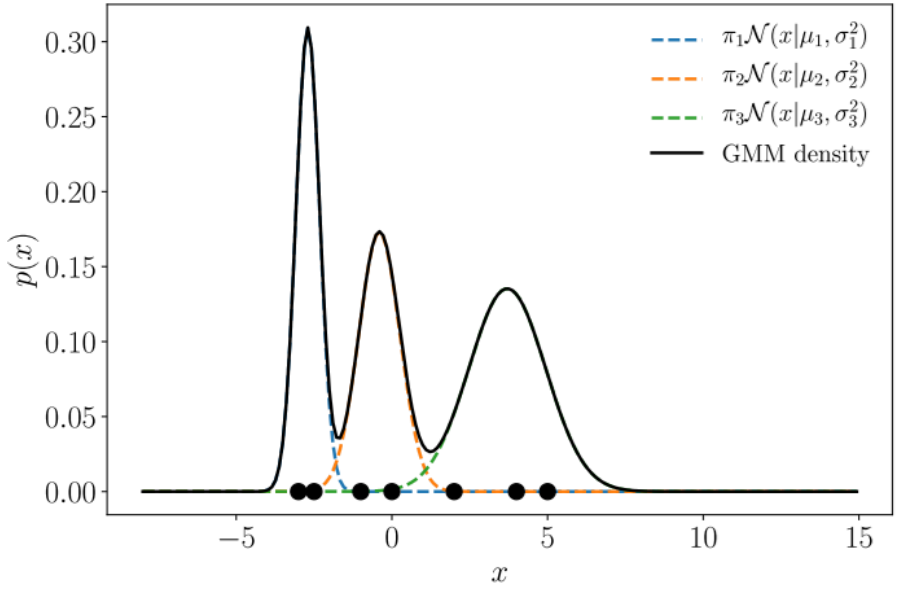


$$\pi_1 : \frac{1}{3} \rightarrow 0.29$$
$$\pi_2 : \frac{1}{3} \rightarrow 0.29$$
$$\pi_3 : \frac{1}{3} \rightarrow 0.42$$

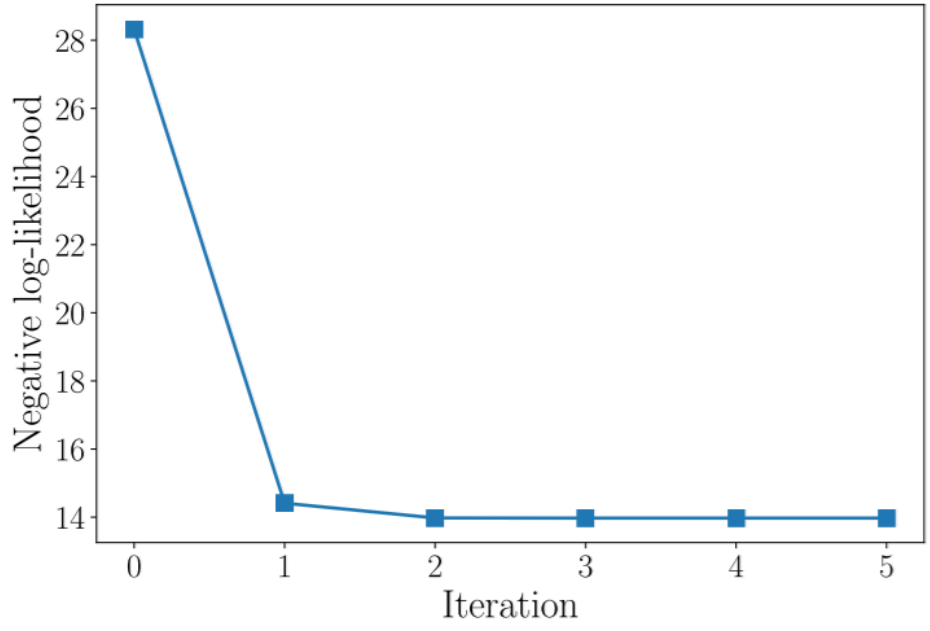


- Aquí vemos que la tercera componente obtiene más peso/importancia, mientras que las otras componentes se vuelven un poco menos importantes.
- En general, habiendo actualizado las medias, las varianzas y los pesos una vez, obtenemos el GMM que se muestra en la figura de la derecha.

Ejemplo ilustrado: resultados finales

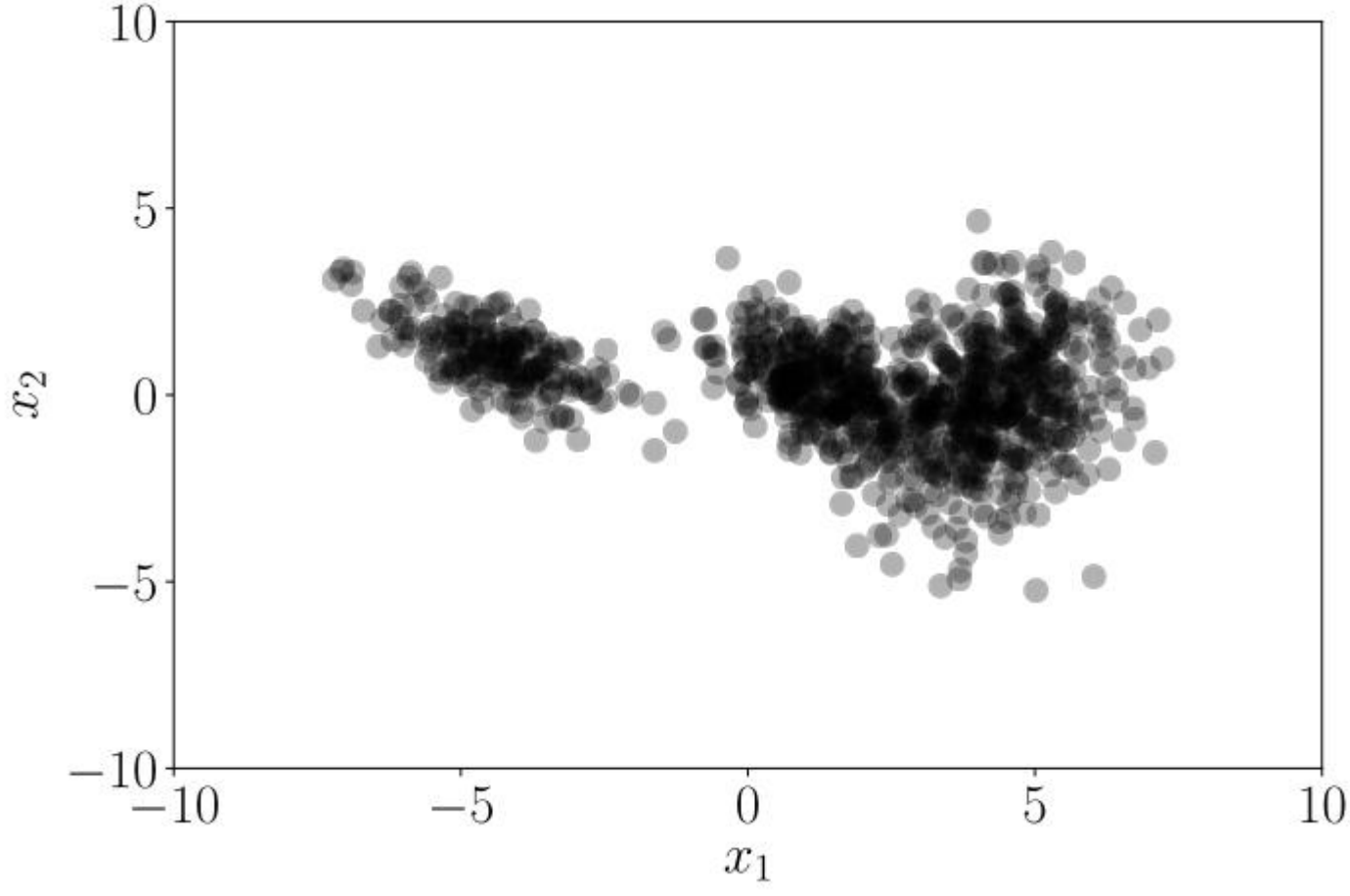


Ajuste final de GMM. Después de cinco iteraciones, el algoritmo EM converge y devuelve este GMM.

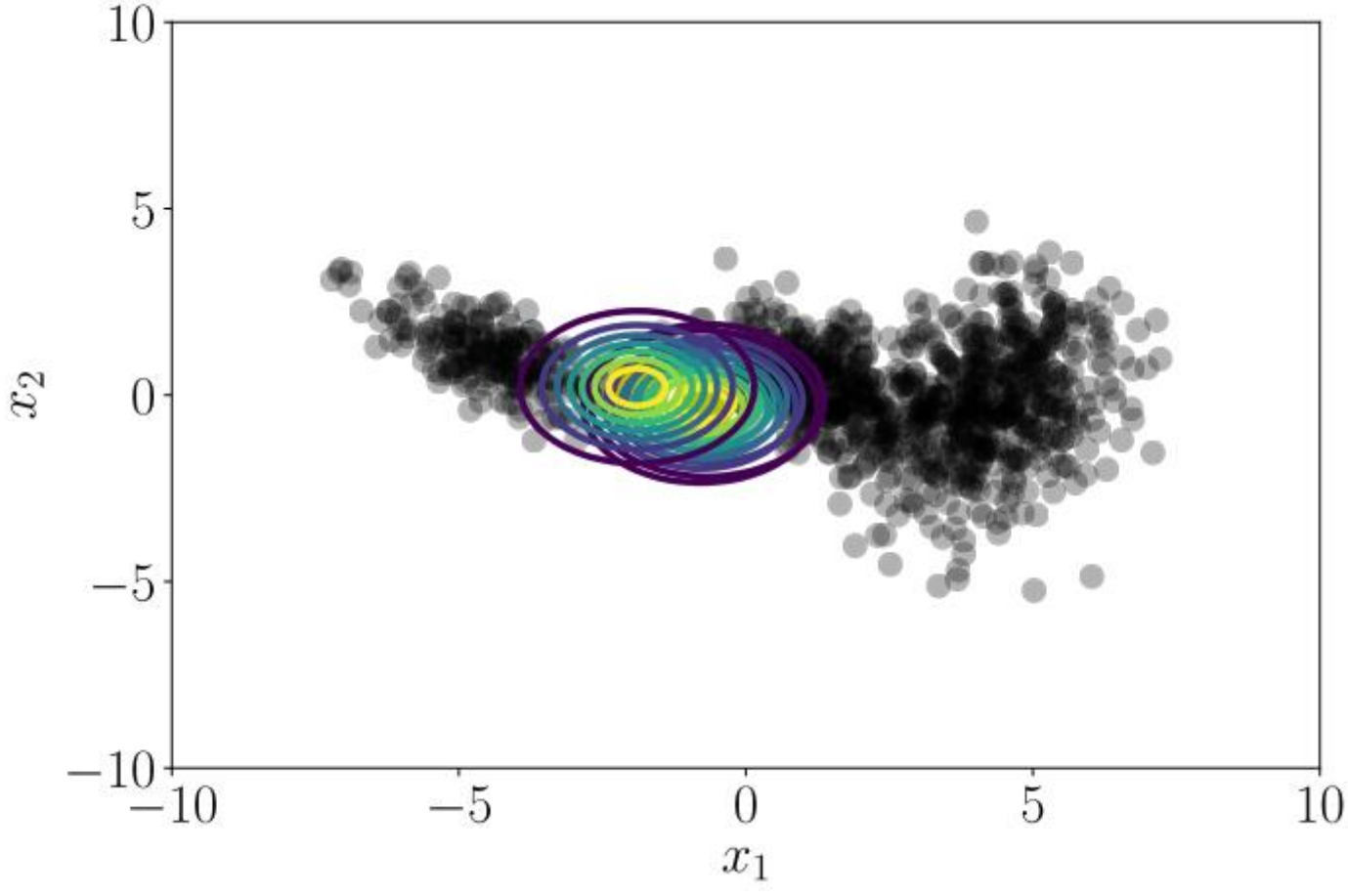


El negativo del log de la verosimilitud como una función de las iteraciones EM .

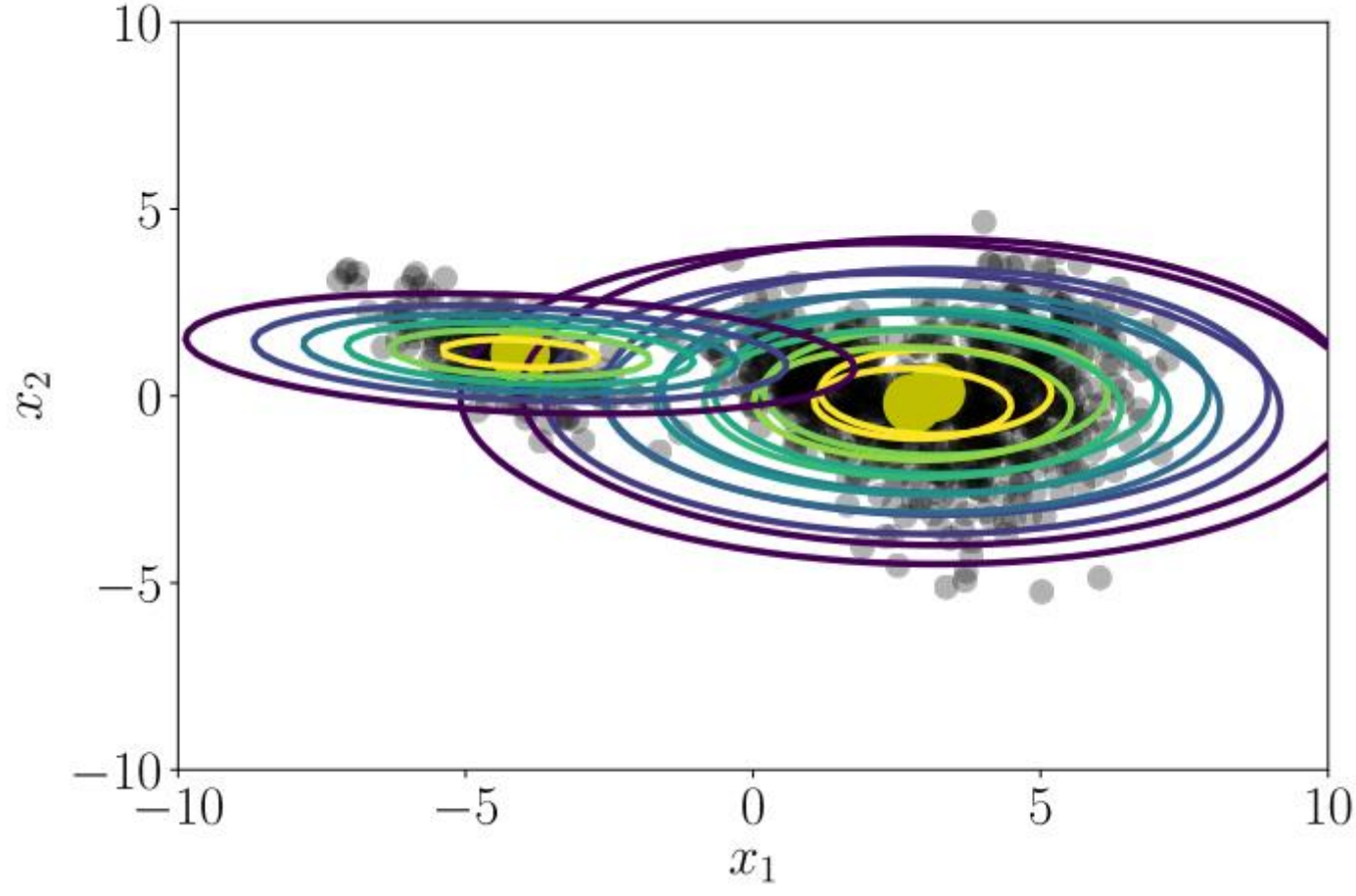
Ejemplo 2: conjunto de datos



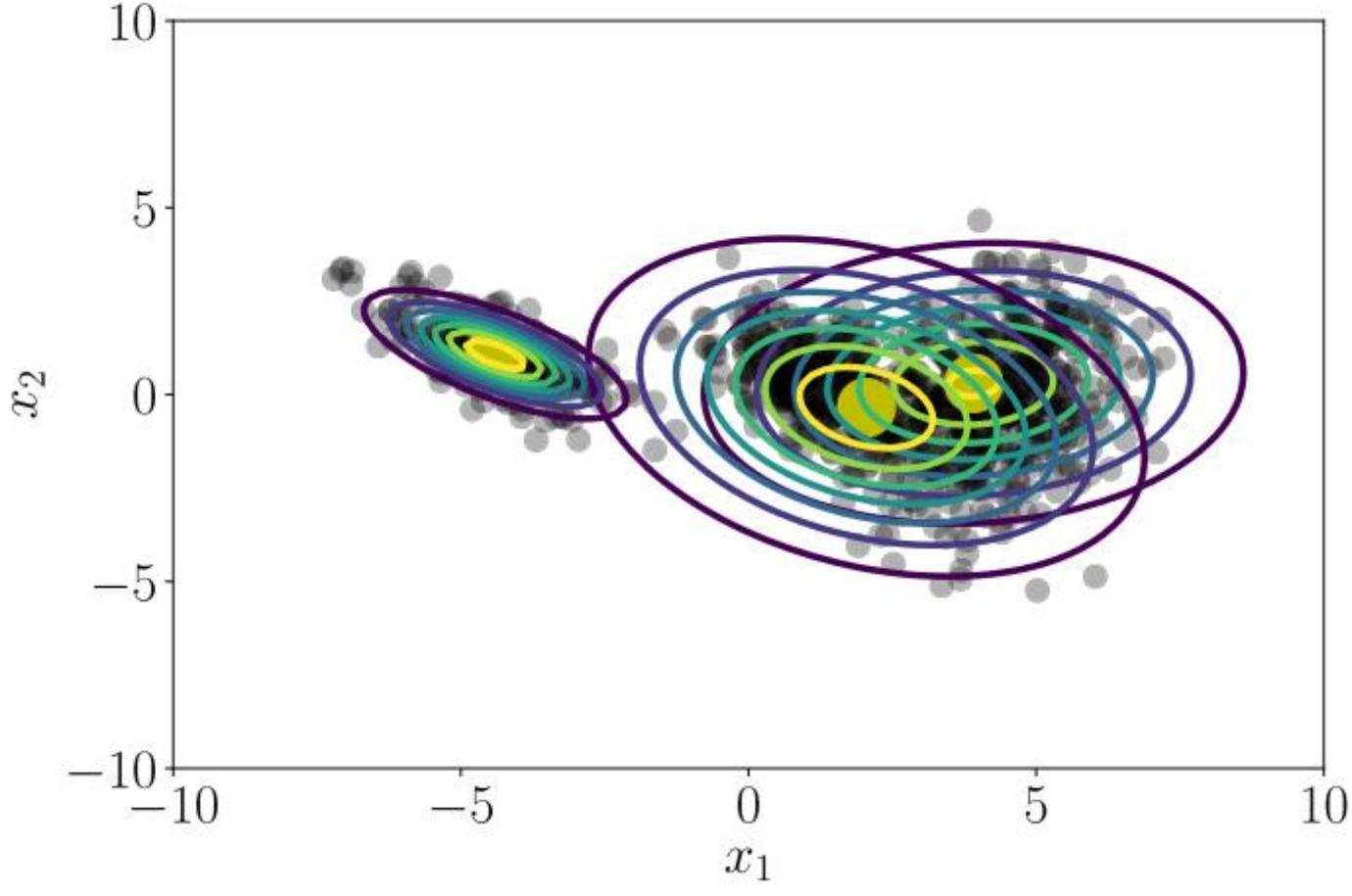
Ejemplo 2: inicialización



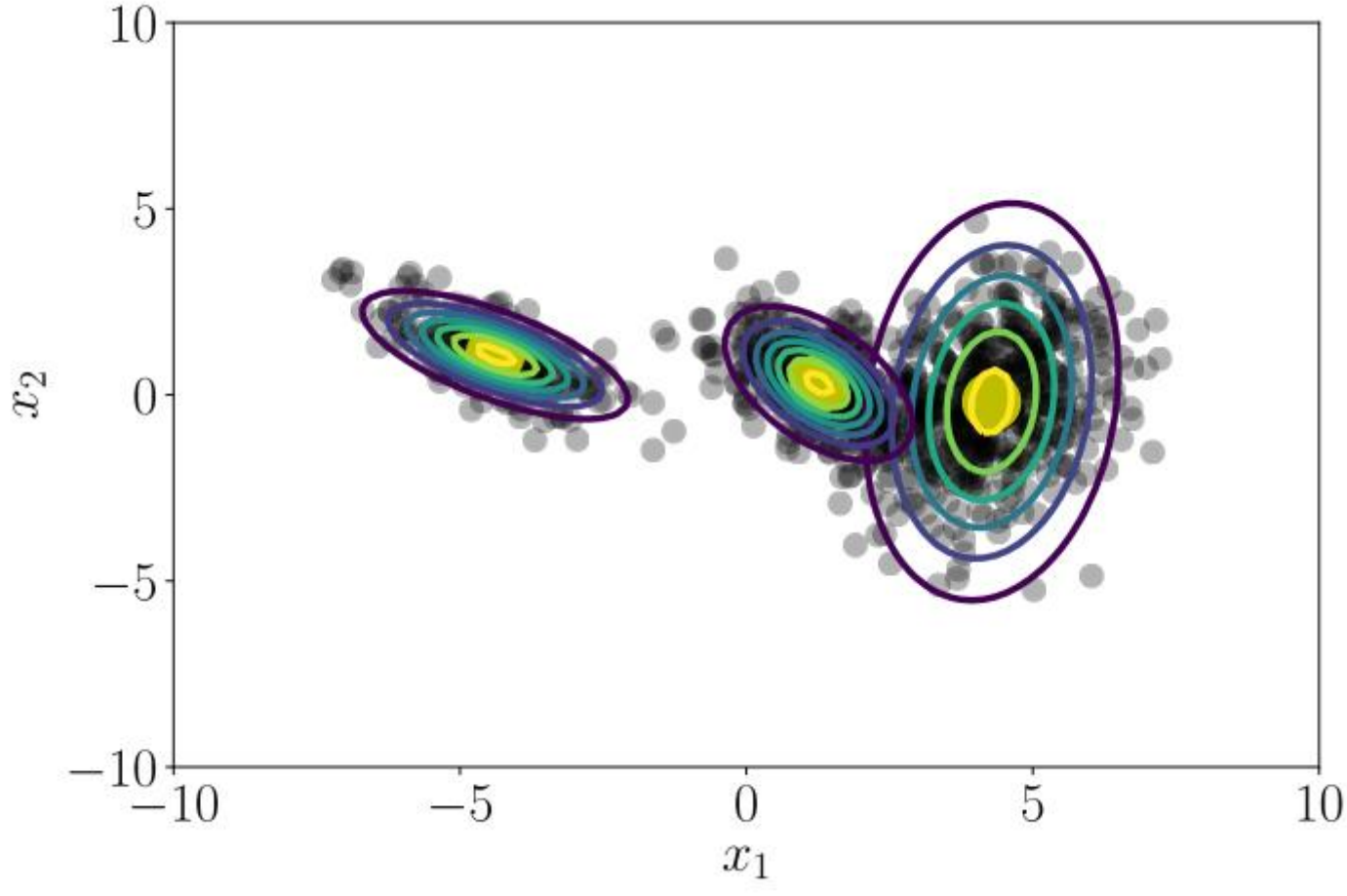
Ejemplo 2: una iteración



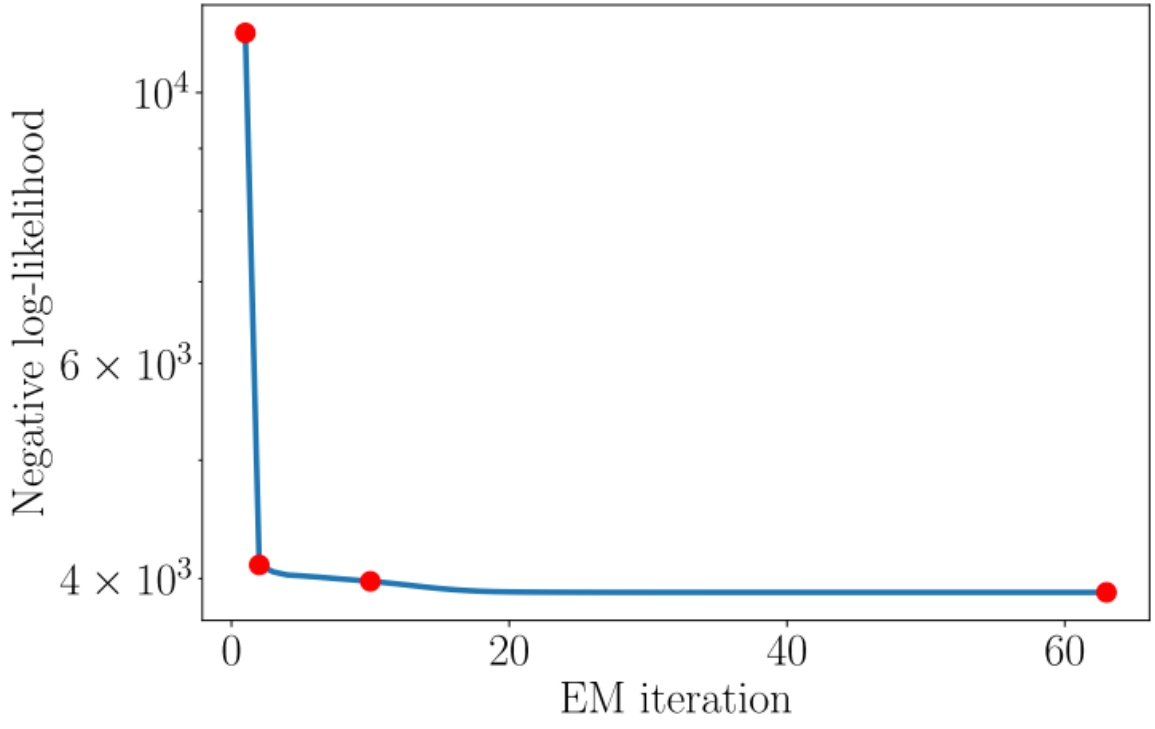
Ejemplo 2: 10 iteraciones



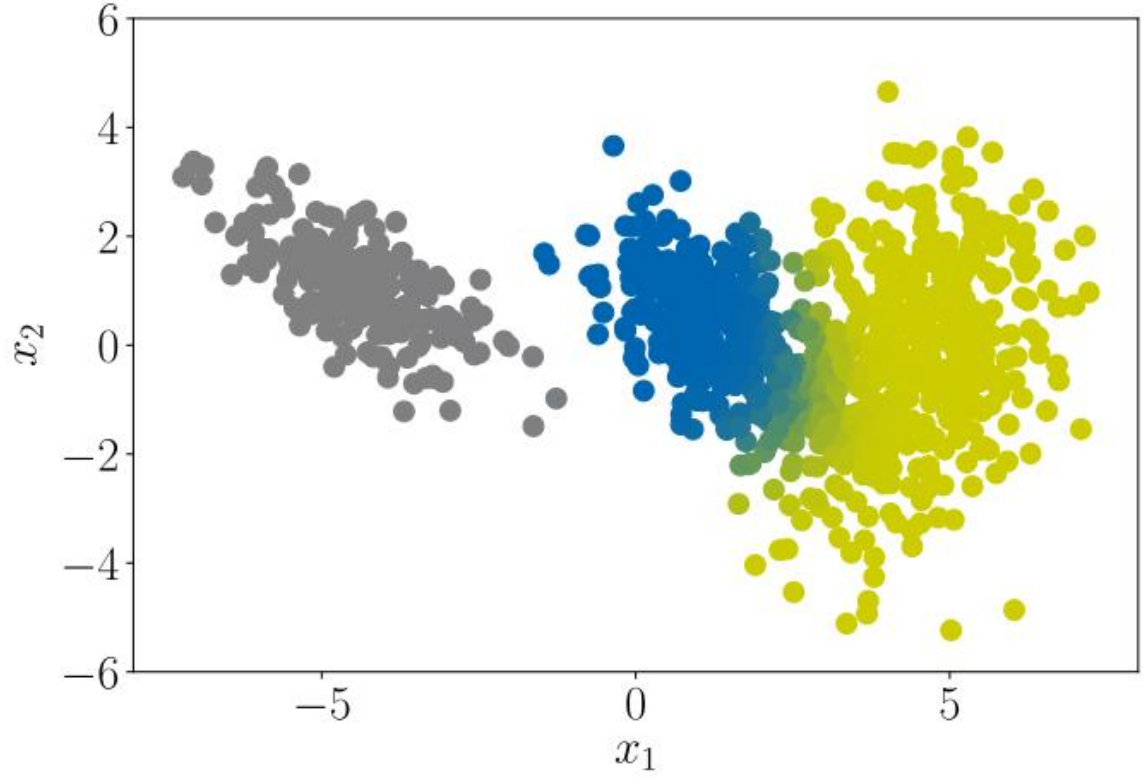
Ejemplo 2: 62 iteraciones



Ejemplo 2: negativo del log de la verosimilitud



Ejemplo 2: Conjunto de datos coloreado de acuerdo con las responsabilidades de los componentes de la mezcla.



$$p(x) = 0.29\mathcal{N}(x \mid -2.75, 0.06) + 0.28\mathcal{N}(x \mid -0.50, 0.25) + 0.43\mathcal{N}(x \mid 3.64, 1.63) .$$

¿Preguntas?

Extras

Aprendizaje de parámetros a través de MLE

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0}, \quad \frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \pi_k} = 0.$$

regla de la cadena

$$\frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

donde $\theta = \{\mu_k, \Sigma_k, \pi_k, k = 1, \ldots, K\}$ son los parámetros del modelo y

$$\frac{1}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} = \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Actualización de las medias GMM (prueba)

De $\frac{\partial \log p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{p(x_n|\theta)} \frac{\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, vemos que el gradiente del log de la verosimilitud respecto a los a las medias $\mu_k, k = 1, \ldots, K$, requiere que calculemos la derivada parcial

$$\frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{j=1}^{K} \pi_j \frac{\partial \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\partial \boldsymbol{\mu}_k} = \pi_k \frac{\partial \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k}$$
$$= \pi_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) ,$$

juntando todo para que la derivada parcial deseada de $\mathcal{L}$ con respecto a μ_k

$$\frac{1}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} = \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k}$$
$$= \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \underbrace{\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{= r_{nk}}$$
$$= \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} .$$

Actualización de las medias GMM (prueba)

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}$$

Ahora resolvemos para $\boldsymbol{\mu}_k^{new}$ tal que $\frac{\partial \mathcal{L}(\boldsymbol{\mu}_k^{\text{new}})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top$

$$\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n = \sum_{n=1}^{N} r_{nk}\boldsymbol{\mu}_k^{\text{new}} \iff \boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\boxed{\sum_{n=1}^{N} r_{nk}}} = \frac{1}{\boxed{N_k}} \sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n$$

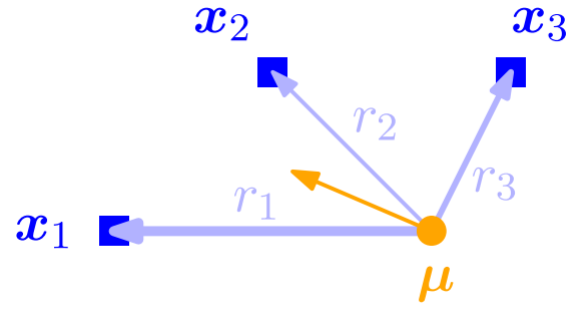
donde definimos

$$N_k := \sum_{n=1}^{N} r_{nk}$$

como la responsabilidad total de la k-ésima componente de mezcla para todo el conjunto de datos.

Actualización de las medias GMM (prueba)

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$



- Por lo tanto, la media μ_k es jalada hacia un punto de datos x_n con una fuerza dada por r_{nk}. Las medias son jaladas más fuerte hacia puntos de datos para los cuales la componente de mezcla correspondiente tiene una alta responsabilidad, es decir, una alta verosimilitud.

- También podemos interpretar la actualización de la media como el valor esperado de todos los puntos de datos bajo la distribución dada por $\boldsymbol{r}_k := [r_{1k}, \ldots, r_{Nk}]^\top / N_k$, el cual es un vector de probabilidad normalizado, es decir, $\boldsymbol{\mu}_k \leftarrow \mathbb{E}_{\boldsymbol{r}_k}[\mathcal{X}]$

Estimación de los parámetros

- A continuación, determinamos las actualizaciones de los parámetros del modelo μ_k, Σ_k, π_k para unas responsabilidades dadas.
- Veremos que todas las ecuaciones de actualización dependen de las responsabilidades, lo que hace imposible una solución de forma cerrada al problema de la estimación de máxima verosimilitud.
- Sin embargo, para determinadas responsabilidades actualizaremos un parámetro del modelo a la vez, mientras mantenemos los demás fijos. Después de esto, recalcularemos las responsabilidades.
- La iteración de estos dos pasos eventualmente convergerá a un óptimo local y es una instancia específica del algoritmo EM.

Actualizando las medias del GMM

La actualización de los parámetros de medias $\mu_k, k = 1, \ldots, K$, del GMM viene dada por

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

donde las responsabilidades r_{nk} se definen como

$$r_{nk} := \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

La actualización de las medias μ_k de las componentes individuales de la mezcla depende de todas las medias, las matrices de covarianza Σ_k y los pesos de mezcla π_k a través de r_{nk} dado. Por lo tanto, no podemos obtener una solución de forma cerrada para todos los μ_k a la vez.

Actualizando las medias de GMM

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

$$r_{nk} := \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

La actualización de los parámetros medios parece bastante sencilla. Sin embargo, tenga en cuenta que las responsabilidades r_{nk} son una función de π_j, μ_j, Σ_j para todo $j = 1, \ldots, K$, de modo que las actualizaciones dependen de todos los parámetros del GMM, y no se puede obtener una solución de forma cerrada.

Actualización de las covarianzas GMM

La actualización de los parámetros de covarianza $\Sigma_k, k = 1, \ldots, K$ de GMM viene dada por

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$$

donde

$$r_{nk} := \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$N_k := \sum_{n=1}^{N} r_{nk}$$

Al igual que con las actualizaciones de los parámetros de las medias, esta actualización depende de todos los $\pi_j, \mu_j, \Sigma_j, j = 1, \ldots, K$, a través de las responsabilidades r_{nk}, lo que prohíbe una solución de forma cerrada.

Actualización de los pesos de mezcla GMM

Los pesos de mezcla del GMM se actualizan como

$$\pi_k^{new} = \frac{N_k}{N}, \quad k = 1, \ldots, K$$

donde N es el número de puntos de datos y N_k se define como

$$N_k := \sum_{n=1}^{N} r_{nk}$$

- Podemos identificar el peso de la mezcla como la relación entre la responsabilidad total del k-ésimo clúster y el número de puntos de datos.
- Dado que $N = \sum_k N_k$, el número de puntos de datos también puede interpretarse como la responsabilidad total de todos los componentes de la mezcla juntos, de modo que π_k es la importancia relativa de la k-ésima componente de la mezcla para el conjunto de datos.
- Puesto que $N_k = \sum_{i=1}^{N} r_{nk}$, la ecuación de actualización para los pesos de la mezcla π_k también depende de todos los $\pi_j, \mu_j, \Sigma_j, j = 1, \ldots, K$ a través de las responsabilidades r_{nk}.

EM Algorithm

EM Algorithm

- Desafortunadamente, las actualizaciones no constituyen una solución de forma cerrada para las actualizaciones de los parámetros μ_k, Σ_k, π_k del modelo de mezcla porque las responsabilidades r_{nk} dependen de esos parámetros de una manera compleja.
- Sin embargo, los resultados sugieren un esquema iterativo simple para encontrar una solución al problema de estimación de parámetros a través de MLE.
- El algoritmo de maximización de la Esperanza (EM algorithm) se propuso como un esquema iterativo general para el aprendizaje de parámetros en modelos de mezcla y, más generalmente, en modelos de variables latentes.