PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

Cierre de la clase 5 – Clustering basado en densidad

DBSCAN:

¿Qué es transformación de dominio? ¿Cómo se usa en DBSCAN? ¿En cuál otra técnica vista en la asignatura se usa?

¿Qué es un k-dist plot? ¿Para qué se usa en DBSCAN?

HDBSCAN:

¿Para qué se usa el minimum spanning tree en HDBSCAN?

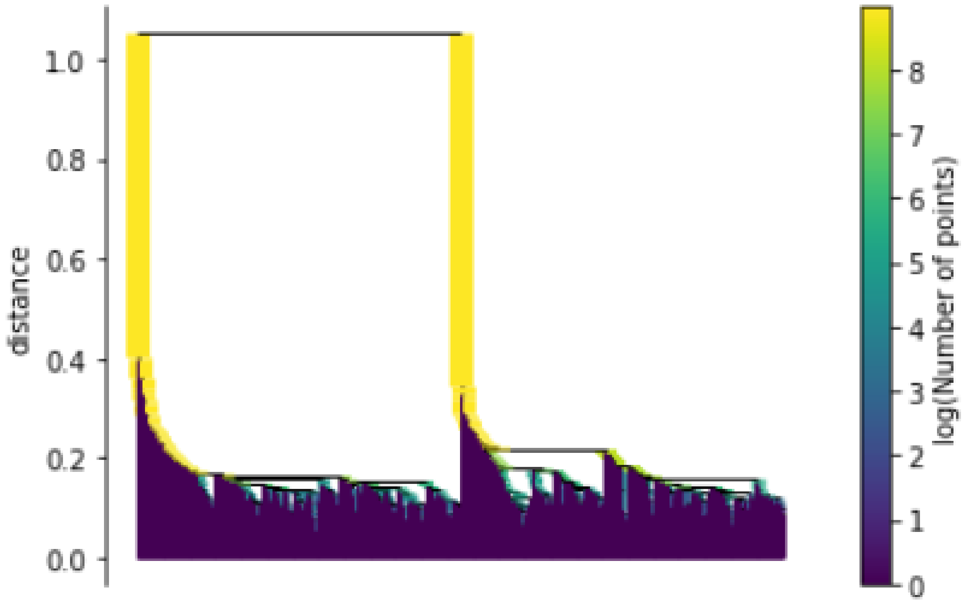
¿Qué es el lambda de nacimiento y muerte?

¿Cómo se define la estabilidad de un cluster en HDBSCAN?

¿Cómo se desagrega un cluster usando la definición de estabilidad?

Cierre de la clase 5 – Actividad formativa

¿Cuál cluster es más estable?



- GMM -

Clustering probabilístico

Modelos de mezcla

Vectores de características

Datos vectoriales: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

pesos de la mezcla

probabilidad del dato

parámetros del modelo

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\theta_k)$$

componentes de la mezcla

Pesos de la mezcla:

$$0 \le \pi_k \le 1 \ (k = 1, \ldots, K), \quad \text{y} \quad \sum_{k=1}^{K} \pi_k = 1.$$

Clustering probabilístico

$\Theta = \{\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K\}$

Modelos de mezcla

Inferencia de parámetros del modelo: $\{\pi_k\}$ y $\{\theta_k\}$.

El número de componentes K se considera fijo (hiper-parámetro).

Asumimos que los datos son muestreados i.i.d., la probabilidad de generación del dataset es:

$$p(\mathbf{X}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\theta_k)$$

y en forma logarítmica:

$$\log p(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\theta_k).$$

Se usa el enfoque de máxima verosimilitud para inferencia:

$$\Theta_{ML} = \arg\max_{\Theta}\{\log p(\mathbf{X}|\Theta)\}$$

MLE: la mejor estimación es aquella que maximiza la probabilidad de generar las observaciones.

Clustering probabilístico

Bayes

Modelos de mezcla

$$\Theta_{ML} = \arg\max_{\Theta} \{\log p(\mathbf{X}|\Theta)\}$$

Verosimilitud (cuan probable es la evidencia condicionada al modelo)

Prior (nuestra creencia antes de observar la evidencia)

$$P(\Theta|X) = \frac{P(X|\Theta) \cdot P(\Theta)}{P(X)}$$

Marginal (distribución de la evidencia, antes de pasar por el modelo)

Posterior (cuan probable es el modelo condicionado a la evidencia)

Clustering probabilístico

Carl Gauss

Modelos de mezcla Gaussianas

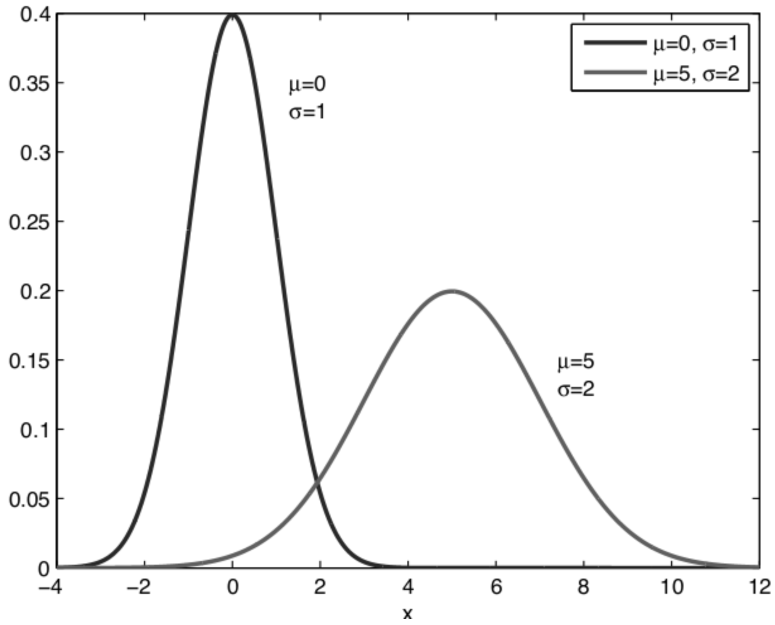
Cada componente de la mezcla es una distribución Gaussiana.

media

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

varianza

Ej.:



Clustering probabilístico

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nm}^2 \end{pmatrix}$$

Carl Gauss

Modelos de mezcla Gaussianas

Si $\mathbf{x}$ es *D* dimensional:

Matriz de covarianza $D \times D$

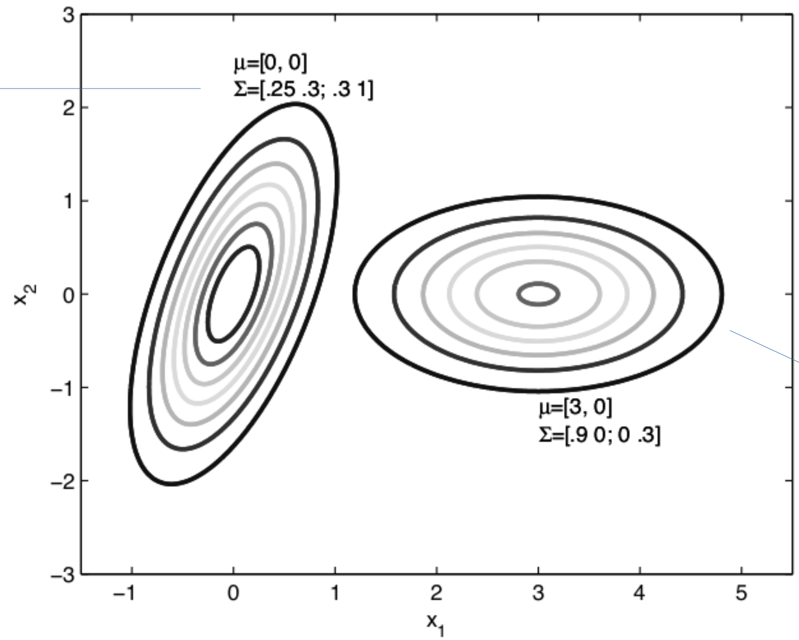
$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\}$$

Vector de medias D-dimensional

Determinante de Σ

Ej.:

$$\begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$



$$\begin{bmatrix} 0.9 & 0.0 \\ 0.0 & 0.3 \end{bmatrix}$$

Clustering probabilístico

Carl Gauss

Modelos de mezcla Gaussianas

Cada componente está representada por los parámetros de una Gaussiana multivariada $p(\mathbf{x}_k|\theta_k) = \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$:

$$p(\mathbf{x}_n|\Theta) = p(\mathbf{x}_n|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k).$$

Dado $\mathbf{X}$, la función de verosimilitud queda dada por:

$$l(\Theta) = \log p(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k).$$

Los parámetros del modelo son π_k, μ_k, y Σ_k.

Inferencia: basada en algoritmo Expectation - Maximization (EM).

Algorithm EM for Gaussian Mixtures

Given a set of data points and a Gaussian mixture model, the goal is to maximize the log-likelihood with respect to the parameters.

1: Initialize the means μ_k^0, covariances Σ_k^0, and mixing probabilities π_k^0.

2: **E-step**: Compute

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}.$$

3: **M-step**: Compute

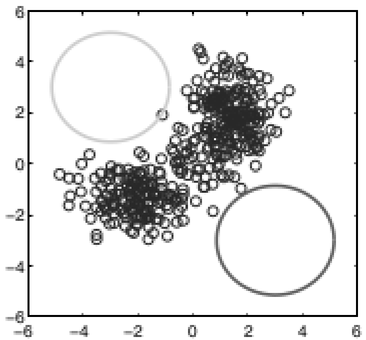
$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}. \quad \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

4: Compute the log-likelihood using $\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$.

Inferencia: basada en algoritmo Expectation - Maximization (EM).

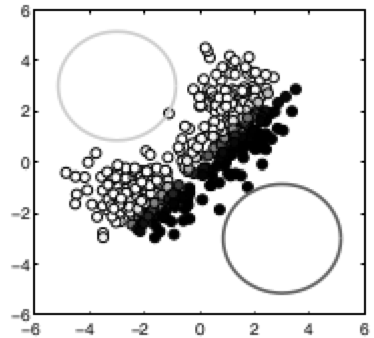
datos|distribución

Algoritmo EM:



Initialize μ_k^0, Σ_k^0, and π_k^0.

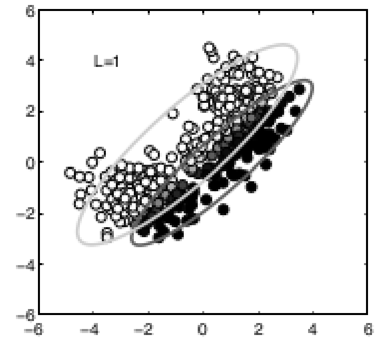
E-step



$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}$$

distribución|datos

M-step

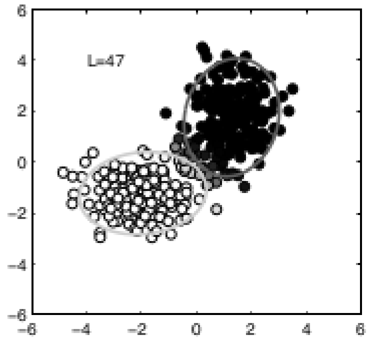


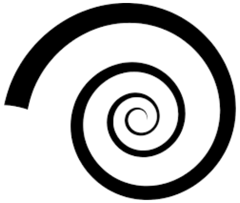
$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

E-step





Iterar hasta converger

¿Cuántas componentes necesita la mezcla?

Criterio de información de Akaike (AIC): maneja un tradeoff (diferencia) entre la bondad de ajuste del modelo (L) y la complejidad del modelo (k).

$$AIC = 2k - 2\ln(L)$$

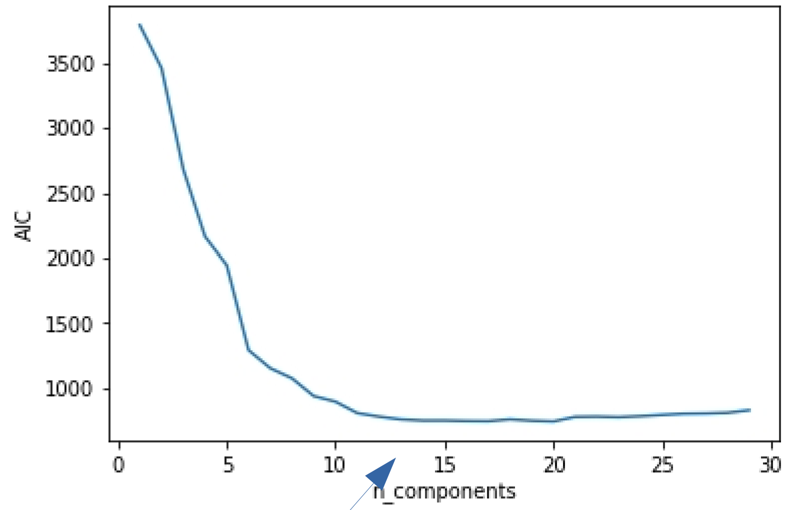
Si existe un balance entre la complejidad y la verosimilitud, AIC $\rightarrow$ 0.

¿Cuántas componentes necesita la mezcla?

Criterio de información de Akaike (AIC): maneja un tradeoff (diferencia) entre la bondad de ajuste del modelo (verosimilitud(y la complekidad del modelo (k).

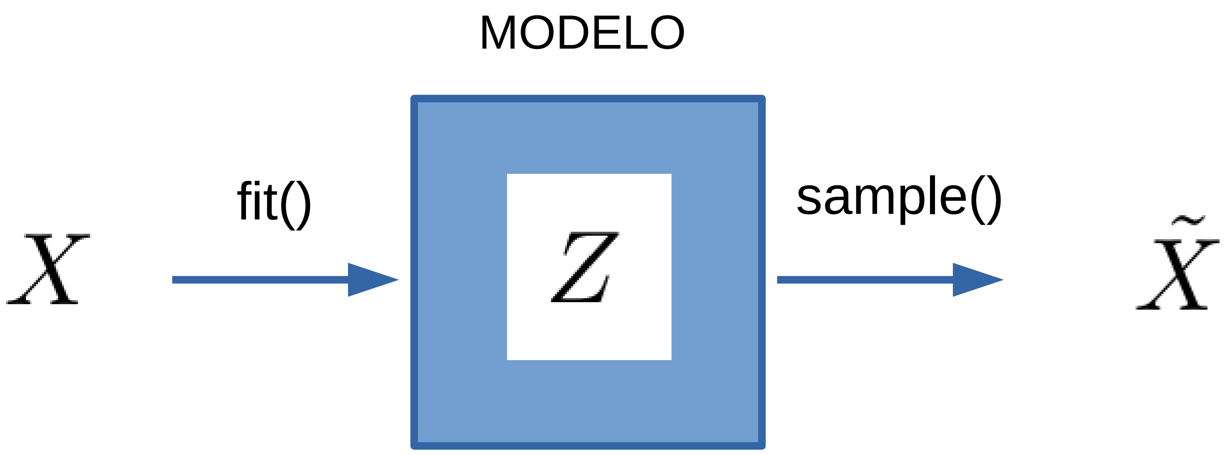
$$AIC = 2k - 2 \ln(L)$$

Si existe un balance entre la complejidad y la verosimilitud, AIC $\rightarrow$ 0.

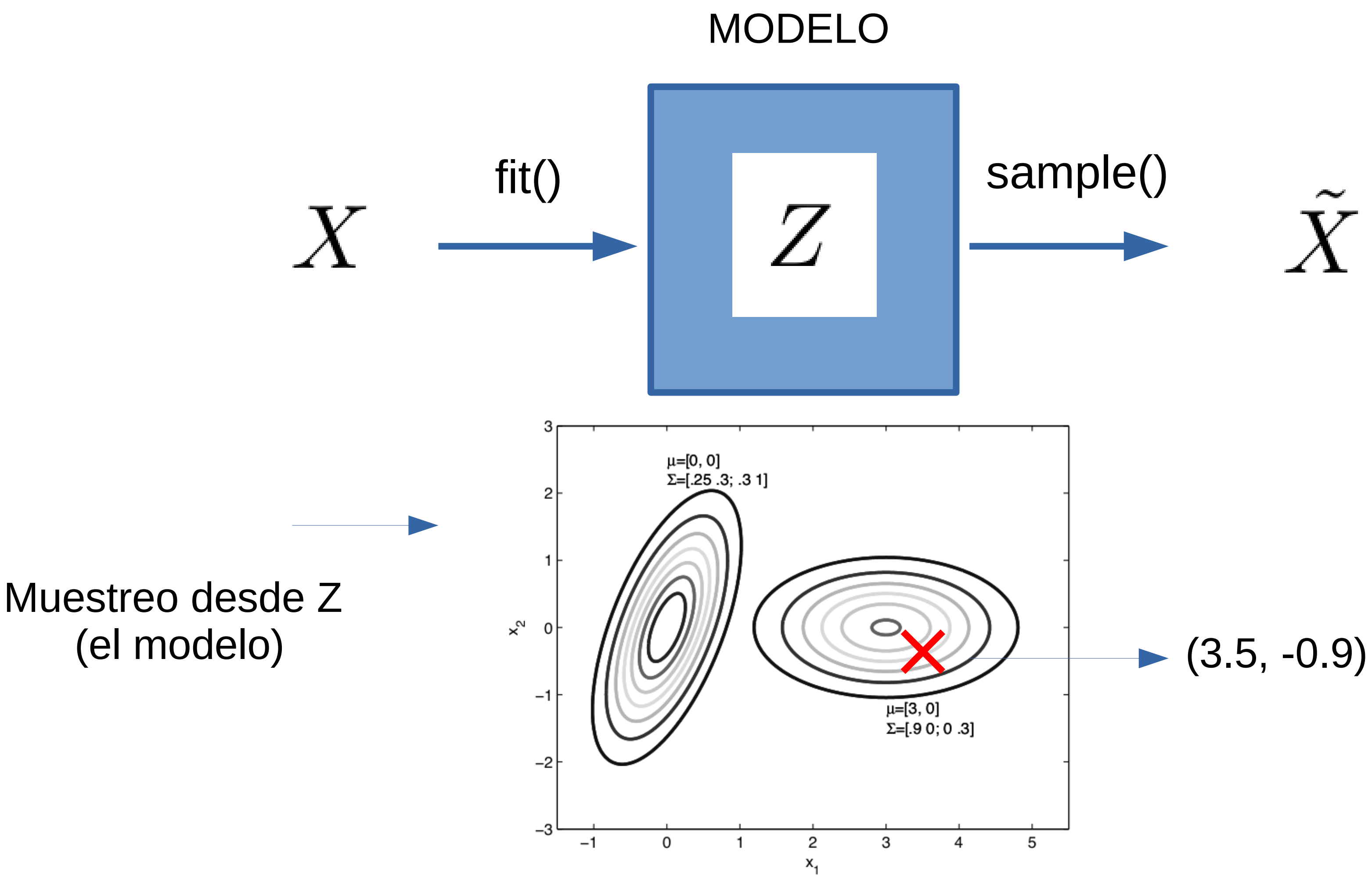


Buscar el mínimo

La idea del modelo generativo



La idea del modelo generativo



- ANEXO -

MLE: debemos calcular las derivadas de $\log p(\mathbf{X}|\pi,\mu,\Sigma)$ w. r. t. π_k, μ_k, y Σ_k.

$$\frac{d}{dx}(\log(x)) = \frac{1}{x}$$

$$l(\Theta) = \log p(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k).$$

$$\frac{d}{dx}(\exp(f(x))) = e^{f(x)}\, f'(x)$$

posterior

$$\frac{\partial l}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j,\Sigma_j)} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k) = \sum_{n=1}^{N} \gamma(z_{nk})\Sigma_k^{-1}(\mathbf{x}_n - \mu_k) = 0$$

$$\begin{aligned}
\frac{\partial \mathbf{L}}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\left(\frac{\partial (\boldsymbol{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}\right) \\
&= -\frac{1}{2}\left(-2\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\
&= \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})
\end{aligned}$$

Análogamente, obtenemos:

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}.$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j,\Sigma_j)}.$$

MLE: debemos calcular las derivadas de $\log p(\mathbf{X}|\pi, \mu, \Sigma)$ w. r. t. π_k, μ_k, y Σ_k.

… falta un poco:

se agrega ya que los π_k deben ser positivos y sumar 1.

$$\log p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right).$$

$$\boxed{\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}}$$