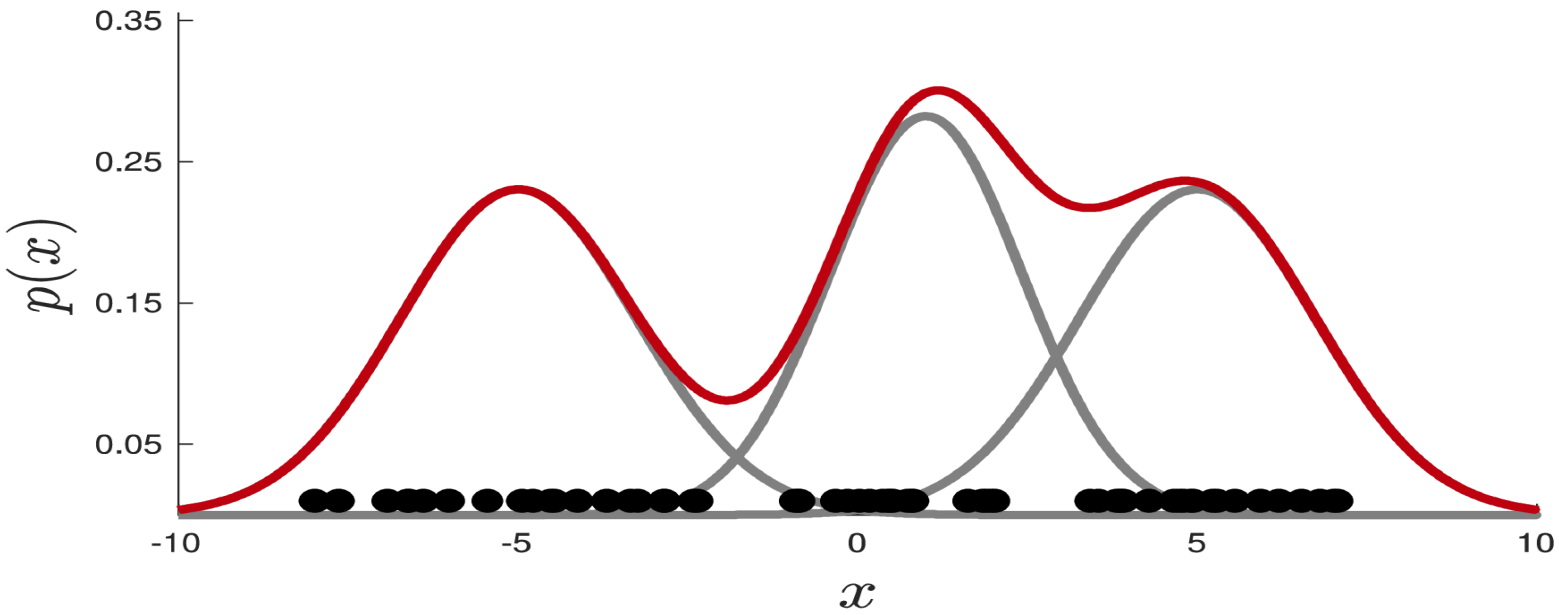
Clasificación Bayesiana
Estimación paramétrica de densidades desconocidas

FaMAF

2019

Modelos de mezclas de gaussianas



Es frecuente encontrar conjuntos de datos que requieren la combinación de varias funciones de densidad para modelarlos de manera más cercana a la distribución real. Se requiere una combinación de tres funciones de densidad para modelar el conjunto de datos representados con puntos negros.

Modelos de mezclas de gaussianas

- Modelar una distribución de probabilidad desconocida mediante la combinación lineal de m funciones de densidad (componentes o modelos):

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{m} p\left(\mathbf{x}|\boldsymbol{\theta}_j\right) p_j \tag{1}$$

  donde $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, p_1, \ldots, p_m]^T$

- Y se satisface

$$\sum_{j=1}^{m} p_j = 1 \quad \text{y} \int_{\mathbf{x}} p\left(\mathbf{x}|\boldsymbol{\theta}_j\right) d\mathbf{x} = 1 \tag{2}$$

Maximización de la expectativa

- Función de log-verosimilitud en función de $\boldsymbol{\Theta}$:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{\ln} \sum_{j=1}^{N} p\left(\mathbf{x}_i | \boldsymbol{\theta}_j\right) p_j \tag{3}$$

- Imposible generar una solución analítica (resolver $\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta})$); no hay información sobre cuál componente generó la muestra $\mathbf{x}_i$.
- El algoritmo maximización de la expectativa (EM) introduce una variable indicadora $y_i = j$ si la muestra $\mathbf{x}_i$ fue generada por la j-ésima componente.
- EM maximiza la expectativa de $\mathcal{L}(\boldsymbol{\Theta})$ condicionada a las muestras observadas $\mathbf{X}$ y la iteración actual del estimado de $\boldsymbol{\Theta}$.

Maximización de la expectativa

Input: $\boldsymbol{X} = \{\mathbf{x}_i | i = 1, \ldots, n\}, m, \epsilon$
Output: $\hat{\boldsymbol{\Theta}}$

1. $\hat{\boldsymbol{\Theta}}^0 \leftarrow \left[\left(\hat{\boldsymbol{\theta}}_1, \hat{p}_1\right)^0, \ldots, \left(\hat{\boldsymbol{\theta}}_m, \hat{p}_m\right)^0\right]$
2. $t \leftarrow 0$
3. **do**
4. $\quad t \leftarrow t + 1$
5. $\quad$ **Paso-E:** $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right) \leftarrow \mathbb{E}\left[\sum_{i=1}^{\infty} \ln\left(p\left(\mathbf{x}_i | j; \hat{\boldsymbol{\Theta}}_j^t\right) p_j^t\right)\right]$
6. $\quad$ **Paso-M:** $\hat{\boldsymbol{\Theta}}^{t+1} \leftarrow \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right)$
7. **until** $\left|Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right) - Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{t+1}\right)\right| < \epsilon$

Maximización de la expectativa

El algoritmo EM para mezclas Gaussianas:

1. Inicialización: definir los parámetros iniciales de m componentes Gaussianas: $\hat{\mathbf{\Theta}}^0 = \left[\left(\hat{\boldsymbol{\mu}}_1, \hat{\mathbf{\Sigma}}_1, \hat{p}_1\right)^0, \ldots, \left(\hat{\boldsymbol{\mu}}_m, \hat{\mathbf{\Sigma}}_m, \hat{p}_m\right)^0\right]$
2. Paso-E: Computar la esperanza de $\mathcal{L}(\mathbf{\Theta})$:

$$Q\left(\mathbf{\Theta}; \hat{\mathbf{\Theta}}^t\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \ln\left(p\left(\mathbf{x}_i|j; \hat{\boldsymbol{\theta}}_j^t\right)\hat{p}_j^t\right) p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right) \qquad (4)$$

donde

$$p\left(j|\mathbf{x}; \hat{\mathbf{\Theta}}^t\right) = \frac{p\left(\mathbf{x}|j; \hat{\boldsymbol{\theta}}_j^t\right) p_j^t}{\sum_{k=1}^{m} p_k\left(\mathbf{x}|\hat{\boldsymbol{\theta}}_k^t\right)\hat{p}_k^t}$$

y

$$\ln\left(p\left(\mathbf{x}|j; \hat{\boldsymbol{\theta}}_j\right)\hat{p}_j\right) = -\frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}})^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\hat{\mathbf{\Sigma}}| + \ln\hat{p}_j$$

Maximización de la expectativa

3 Paso-M: Maximizar $Q\left(\mathbf{\Theta}; \hat{\mathbf{\Theta}}^t\right)$ con respecto de $\hat{\boldsymbol{\mu}}_j, \hat{\mathbf{\Sigma}}_j$ y $\hat{p}_j$:

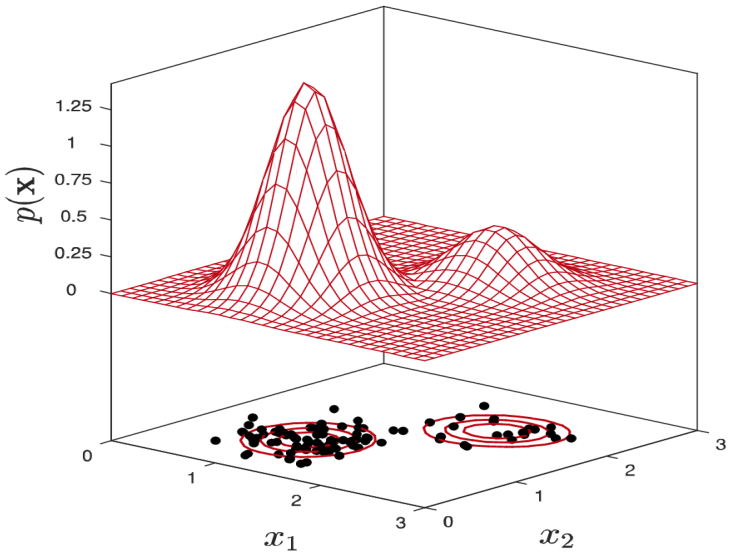
$$\hat{\boldsymbol{\mu}}_j^{t+1} = \frac{\sum_{i=1}^n p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right) \mathbf{x}_i}{\sum_{i=1}^n p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right)}$$

$$\hat{\mathbf{\Sigma}}_j^{t+1} = \frac{\sum_{i=1}^n p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right) \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{t+1}\right) \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{t+1}\right)^T}{\sum_{i=1}^n p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right)} \quad (5)$$

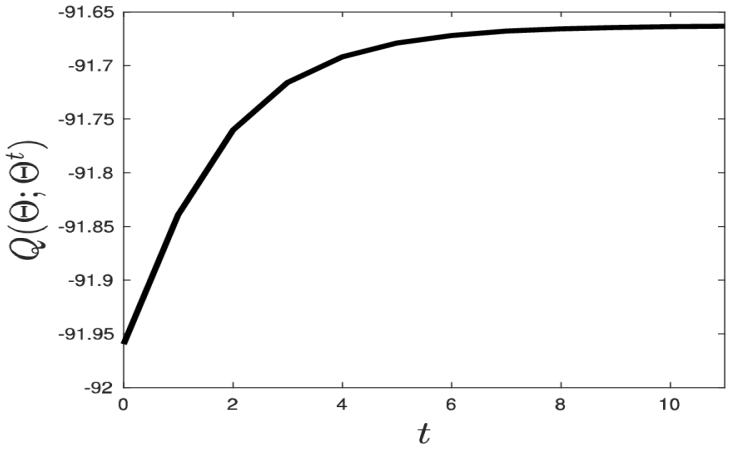
$$\hat{p}_j^{t+1} = \frac{1}{n} \sum_{i=1}^n p\left(j|\mathbf{x}_i; \hat{\mathbf{\Theta}}^t\right)$$

4 Finalizar si $\left|Q\left(\mathbf{\Theta}; \hat{\mathbf{\Theta}}^t\right) - Q\left(\mathbf{\Theta}; \hat{\mathbf{\Theta}}^{t+1}\right)\right| \leq \epsilon$; en caso contrario regresar al Paso-E.

Maximización de la Expectativa



(a)



(b)

(a) Muestras bivariadas originadas a partir de dos funciones normales y estimación de la función de densidad mediante el algoritmo EM. (b) Valores de log-verosimilitud en función del número de iteraciones.

Inicialización de parámetros

1. Estimar las medias $\hat{\boldsymbol{\mu}}_j$ para $j = 1, \ldots, m$ componentes mediante el algoritmo k -means $(k = m)$
2. Generar variables indicadoras: $y_i = \underset{j=1,\ldots,m}{\arg\min} \left\| \mathbf{x}_i - \hat{\boldsymbol{\mu}}_j \right\|_2$
3. Estimar la probabilidad $\hat{p}_j = n_j / n$ para $j = 1, \ldots, m$
4. Estimar la matriz de covarianza $\hat{\boldsymbol{\Sigma}}_j$:

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j - 1} \sum_{\mathbf{x}_i \in j} \left(\mathbf{x}_i - \boldsymbol{\mu}_j \right) \left(\mathbf{x}_i - \boldsymbol{\mu}_j \right)^T, \quad j = 1, \ldots, m$$

5. Generar el arreglo $\hat{\boldsymbol{\Theta}}^0 = \left[\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1, \hat{p}_1 \right), \ldots, \left(\hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m, \hat{p}_m \right) \right]$

Número óptimo de componentes

Principio de descripción de longitud mínima (MDL): minimiza la suma de la complejidad del modelo $L(\boldsymbol{M})$ y la eficiencia de la descripción de los datos con respecto al modelo $L(\boldsymbol{X}, \boldsymbol{M})$:

$$L(\boldsymbol{X}, \boldsymbol{M}) = \mathrm{mn}_{\boldsymbol{M} \in \mathcal{M}}[L(\boldsymbol{M}) + L(\boldsymbol{X}|\mathcal{M})] \tag{6}$$

Número óptimo de componentes

- Principio MDL para mezclas Gaussianas con m componentes:

$$L\left(\boldsymbol{X}, \hat{\boldsymbol{\Theta}}_m\right) = \frac{\alpha_m}{2} \ln n - \sum_{i=1}^{n} \ln \sum_{j=1}^{m} p\left(\mathbf{x}_i | j; \hat{\boldsymbol{\theta}}_j\right) \hat{p}_j \tag{7}$$

donde

$$\alpha_m = (m-1) + md + \frac{md(d+1)}{2}$$

- Número óptimo de componentes:

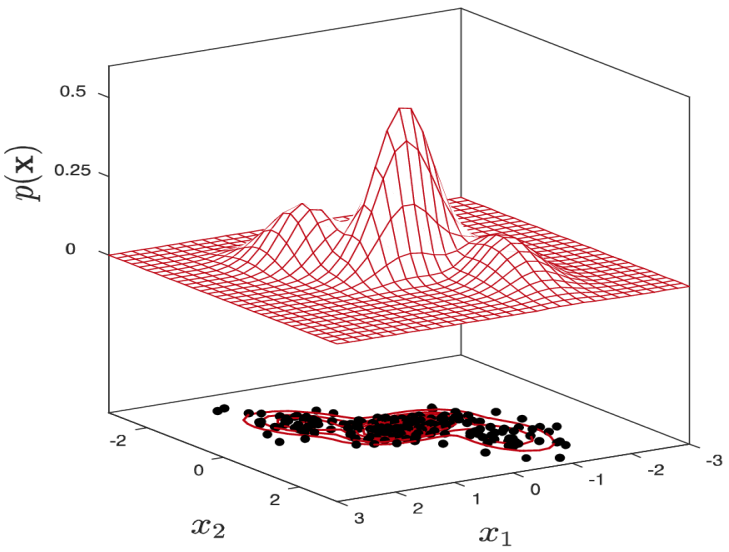
$$m^* = \underset{k=1,\ldots,m}{\arg\min} \left[L\left(\boldsymbol{X}, \boldsymbol{\Theta}_k\right)\right] \tag{8}$$

Número óptimo de componentes

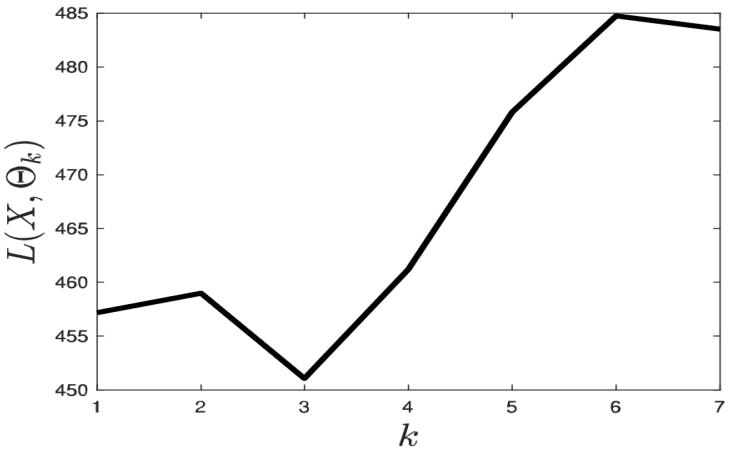
Input: $\boldsymbol{X} = \{\mathbf{x}_i | i = 1, \ldots, n\}, m_{\text{mx}}, \epsilon$
Output: m^*

1. $m^* \leftarrow 1$
2. for $(m = 1; m \leq m_{\text{máx}}; m++)$ do
3. $\quad \hat{\Theta}_k \leftarrow \text{EM}(\boldsymbol{X}, m, \epsilon)$ // Algoritmo EM
4. $\quad L\left(\boldsymbol{X}, \hat{\boldsymbol{\Theta}}_m\right) \leftarrow \frac{\alpha_m}{2} \ln n - \sum_{i=1}^{n} \ln \sum_{j=1}^{m} p\left(\mathbf{x}_i | j; \hat{\boldsymbol{\theta}}_j\right) \hat{p}_j$
5. $\quad$ if $(m > 1) \wedge \left(L\left(\boldsymbol{X}, \hat{\Theta}_m\right) < L\left(X, \hat{\Theta}_{m-1}\right)\right)$ then
6. $\quad\quad m^* \leftarrow m$
7. $\quad$ end
8. end

Número óptimo de componentes



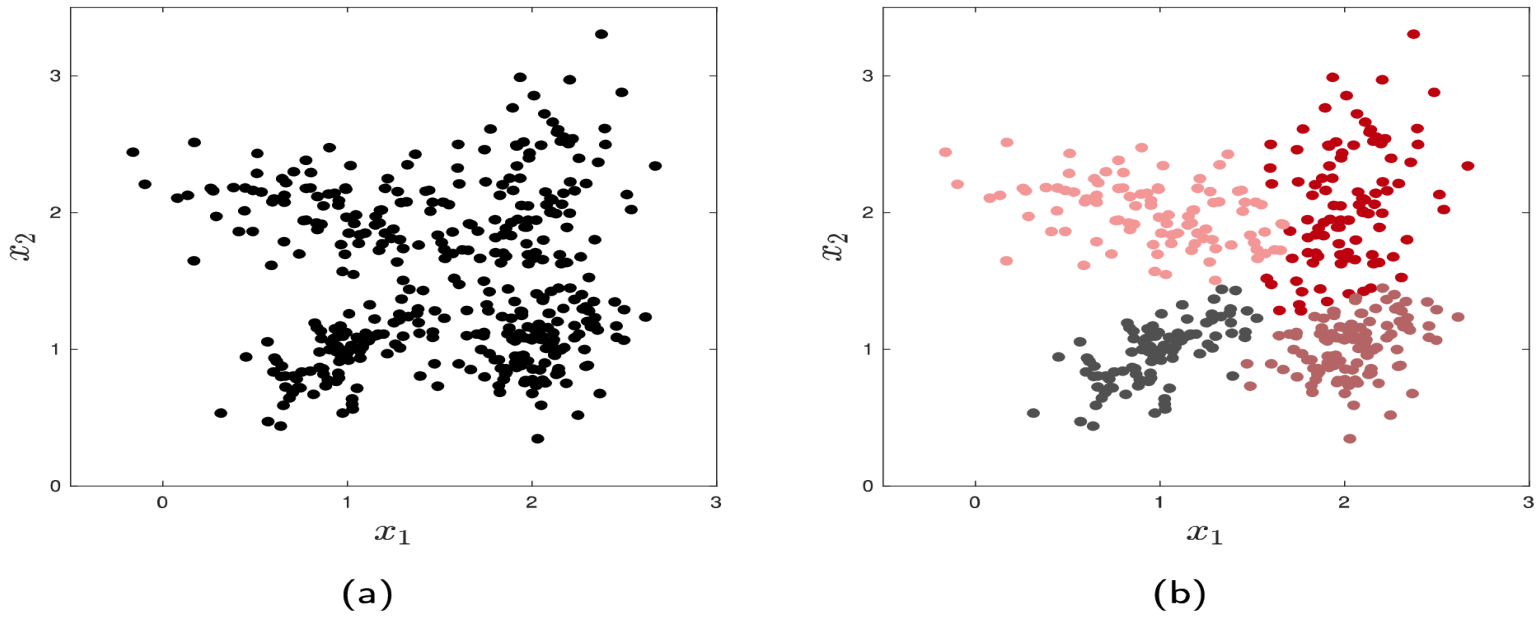
(a)



(b)

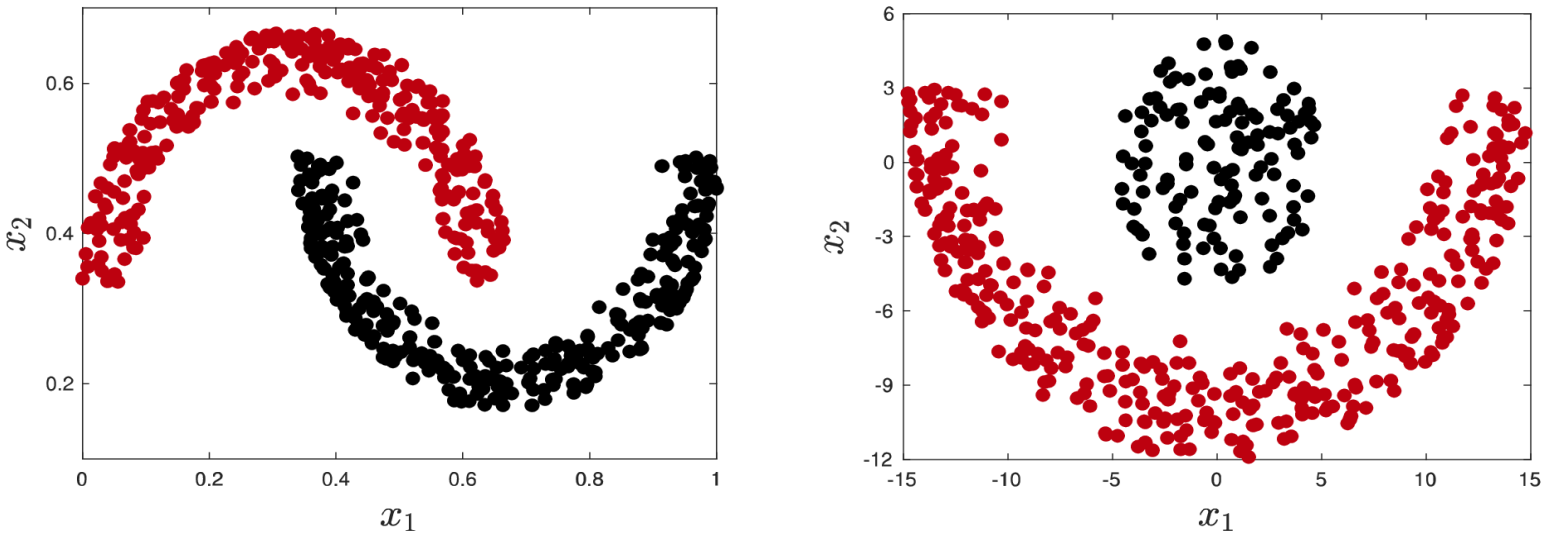
(a) Muestras bivariadas originadas a partir de tres funciones normales y estimación de la función de densidad mediante el algoritmo EM. El valor m^* se estimó mediante el principio MDL. (b) MDL en función del número de componentes donde $m^* = 3$.

Aprendizaje no supervizado



El método de mezclas Gaussianas es útil para agrupamiento de datos: (a) datos sin etiquetar y (b) grupos formados mediante mínima distancia Mahalanobis.

Aprendizaje supervizado



Para modelar distribuciones de clases con formas complejas, se puede utilizar un modelo de mezclas Gaussianas para aproximar las distribución real de cada clase.

Aprendizaje supervizado

Definir una colección de c modelos de mezclas Gaussianas:

$$\hat{\Theta}_\Omega = \left\{ \hat{\Theta}_{\omega_1}, \ldots, \hat{\Theta}_{\omega_c} \right\}$$

donde el modelo para la$i$ -ésima clase es

$$\hat{\Theta}_{\omega_i} = \{M_i, S_i, p_i\}$$

donde cada elemento en $\hat{\Theta}_{\omega_i}$ es

$$M_i = \left\{ \hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_{m_i^*} \right\}, \quad \boldsymbol{S}_i = \left\{ \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_{m_i^*} \right\}, \quad \boldsymbol{p}_i = \left[\hat{p}_1, \ldots, \hat{p}_{m_i^*} \right]$$

donde m_i^* es el número óptimo de modelos para la i -ésima clase.

Aprendizaje supervisado

Input: $\boldsymbol{X} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$ con $y_i \in \{\omega_1, \ldots, \omega_c\}$ $m_{\mathrm{mx}}, \epsilon$
Output: $\boldsymbol{\Theta}_\Omega$

1. $\hat{\boldsymbol{\Theta}}_\Omega \leftarrow \emptyset$
2. for $(i = 1; i \leq c; i++)$ do
3. Identificar patrones de la clase ω_i : $\boldsymbol{X}_{y=\omega_i} \in \boldsymbol{X}$
4. $\hat{\boldsymbol{\Theta}}_{\omega_i} \leftarrow \mathrm{MDL}\left(\boldsymbol{X}_{y=\omega_i}, m_{\max}, \epsilon\right)$ // Algoritmo MDL
5. $\hat{\boldsymbol{\Theta}}_\Omega \leftarrow \left\{\hat{\boldsymbol{\Theta}}_\Omega + \hat{\boldsymbol{\Theta}}_{\omega_i}\right\}$
6. end
7. end

Aprendizaje supervizado

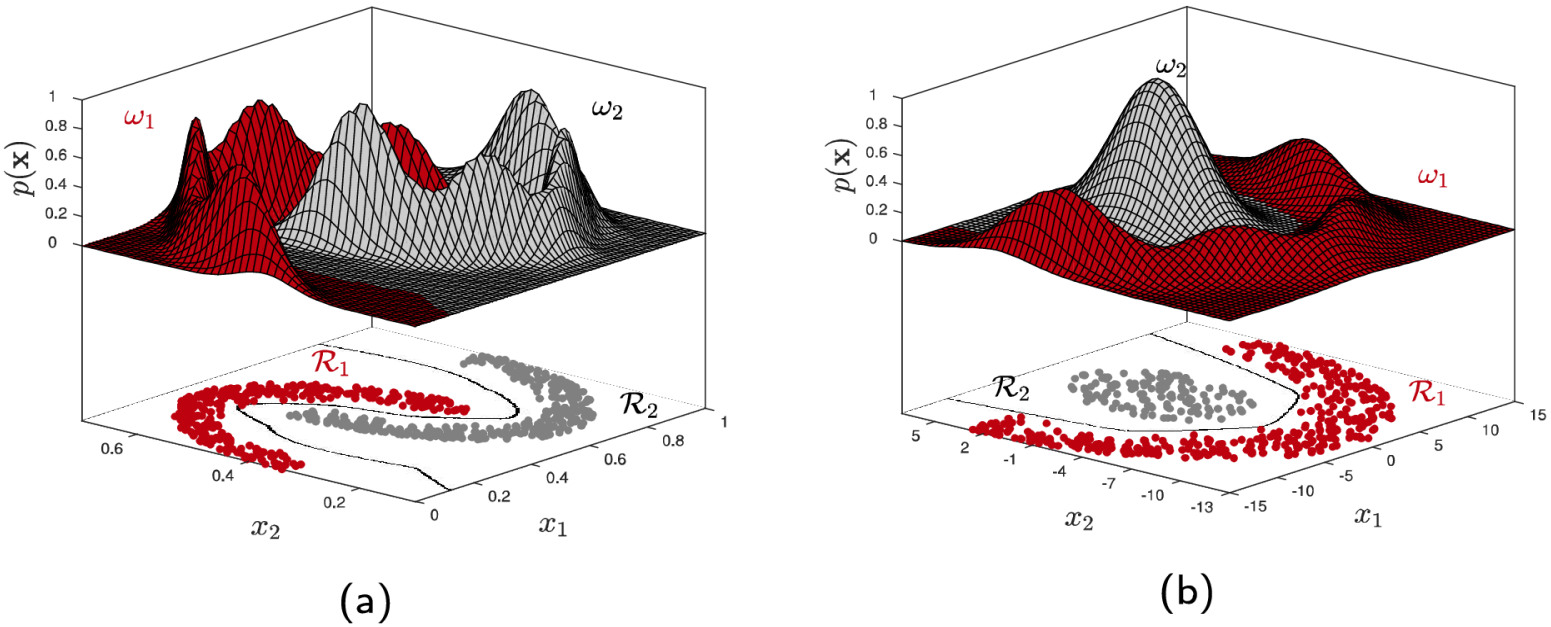
- Etapa de clasificación, un patrón de prueba $\mathbf{x}_t$ se evalúa en el modelo entrenado de mezclas Gaussianas de la i-ésima clase, $\hat{\mathbf{\Theta}}_{\omega_i}$:

$$g_i\left(\mathbf{x}_t\right) = \sum_{j=1}^{m_i^*} p\left(\mathbf{x}|\hat{\boldsymbol{\mu}}_j, \hat{\mathbf{\Sigma}}_j\right)\hat{p}_j, \qquad \text{para } i = 1, \ldots, c \tag{9}$$

- La regla de decisión es

$$\text{Decidir } \omega_i \text{ si } g_i\left(\mathrm{x}_t\right) > g_j\left(\mathrm{x}_t\right) \text{ para todo } i \neq j \tag{10}$$

Aprendizaje supervizado



Estimación de densidad con mezclas Gaussianas para clases que no siguen una distribución normal. La frontera de decisión se define mediante (10)

Aprendizaje supervizado con datos faltantes

- Supongamos tener una muestra de entrenamiento D con datos faltantes D_b, esto es, $D = \{x_1, \ldots, x_n\} = D_g \cup D_b$.
- El estimador de máxima verosimilitud es el vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$que maximiza la $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right)$, que no es calculable con datos faltantes.
- Una solución aproximada sería maximizar la siguiente función objetivo

$$\mathcal{L}(\boldsymbol{\theta}) = E_{D_b}\left(\sum_{i=1}^{n} \ln p\left(\mathbf{x}_i|D_g, \boldsymbol{\theta}\right)\right) \tag{11}$$

y utilizar el algoritmo de Maximizacion de la expectativa para calcular dicho maximo en forma iterativa.

Maximización de la expectativa

1. Inicialize $\mathbf{\Theta}^0$, T, $i = 0$
2. **do** $i \leftarrow i + 1$
3. **Paso-E:** $Q\left(\mathbf{\Theta}; \hat{\mathbf{\Theta}}^i\right)$
4. **Paso-M:** $\mathbf{\Theta}^{i+1} \leftarrow \arg\max_{\mathbf{\Theta}} Q\left(\mathbf{\Theta}; \mathbf{\Theta}^i\right)$

$$Q\left(\theta; \theta^i\right) = E_{D_b}\left(\sum_{i=1}^{n} \ln p\left(\mathbf{x}_i | D_g, \boldsymbol{\theta}\right)\right) \tag{12}$$

5. **until** $\left|Q\left(\mathbf{\Theta}; \mathbf{\Theta}^i\right) - Q\left(\mathbf{\Theta}; \mathbf{\Theta}^{i+1}\right)\right| < T$
6. return $\hat{\mathbf{\Theta}} \leftarrow \mathbf{\Theta}^{i+1}$

Ejemplo

Supongamos que tenemos dos distribuciones normales y los siguientes datos

$$D = \{x_1, x_2, x_3, x_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\} = D_g \cup D_b \tag{13}$$

Entonces $D_b = x_{41} \quad \theta = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$. Sea $\theta^0 = (0, 0, 1, 1)$ el punto de inicio. La función objetivo es

$$Q\left(\theta; \theta^0\right) = E_{D_b}\left(\ln p\left(D_g, D_b\right) | D_g; \theta^0\right) \tag{14}$$

$$= \int_{-\infty}^{\infty} \left[\sum_{k=1}^{3} \ln p\left(x_k | \theta\right) + \ln p\left(x_4 | \theta\right) \right] p\left(x_{41} | \theta^0; x_{42} = 4\right) dx_{41} \tag{15}$$

Ejemplo

$$= \sum_{k=1}^{3} \ln p\left(x_k|\theta\right) + \int_{-\infty}^{\infty} \ln p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix}|\theta\right) \cdot \frac{p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix}|\theta^0\right)}{\int_{-\infty}^{\infty} p\left(\begin{pmatrix} x'_{41} \\ 4 \end{pmatrix}|\theta^0\right) dx'_{41}} dx_{41}$$

La integral del denominador es constante y puede ser sacada fuera de la integral general. Sustituyendo $p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix}|\theta^0\right)$ la densidad Gaussiana en el valor de inicio θ^0 el paso E resulta

Ejemplo

$$= \sum_{k=1}^{3} \ln p\left(x_k|\theta\right) + \frac{1}{K}\int_{-\infty}^{\infty} \ln p\left(\begin{pmatrix} x_{41} \\ 4 \end{pmatrix}|\theta\right).\frac{\exp\left[-\frac{1}{2}(x_{41}^2+4^2)\right]}{2\pi\left|\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right|}dx_{41} \tag{16}$$

$$= \sum_{k=1}^{3} \ln p\left(x_k|\theta\right) - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{\left(1-\mu_2\right)^2}{2\sigma_2^2} - \ln\left(2\pi\sigma_1\sigma_2\right) \tag{17}$$

Esta función $Q\left(\theta;\theta^0\right)$ tiene un máximo en $\theta^1 = \begin{pmatrix} 0.75 \\ 2 \\ 0.938 \\ 2 \end{pmatrix}$

Iterando de la misma forma en tres pasos se obtiene $\theta = \begin{pmatrix} 1 \\ 2 \\ 0.667 \\ 2 \end{pmatrix}$

Ejemplo

Es importante destacar que el óptimo para la verosimilitud con todos los datos quizás sea otro valor. Por ejemplo, si el dato faltante es $x_{14} = 1$, la solución es $\theta = \begin{pmatrix} 1 \\ 2 \\ 0.5 \\ 2 \end{pmatrix}$ y el cálculo de $Q\left(\theta; \theta^0\right)$ es mucho más simple pues no involucra integrales.