Aprendizaje automatizado

Modelos de mezclas

Gibran Fuentes Pineda

Abril 2023

Modelando con variables latentes

- En muchos fenómenos las observaciones (variables observadas) dependen de variables no directamente visibles (variables latentes)

- Un modelo con variables no visibles se conoce como **modelo de variable latente** (MVL)

- Ventajas
  1. Son modelos más compactos en general
  2. Es posible aprender ciertas estructuras en los datos sin supervisión

Modelos de variables latentes (MVL)

- Sea $\mathbf{x}$ una observación muestreada aleatoriamente de una distribución no conocida, se presupone que

$$\mathbf{x} \sim P_{\boldsymbol{\theta}}(\mathbf{x}) \approx P_{real}(\mathbf{x}).$$

- Los modelos de variables latentes representan la distribución $P_{\boldsymbol{\theta}}(\mathbf{x})$ usando variables observadas $\mathbf{x}$ y variables latentes $\mathbf{z}$

$$P_{\boldsymbol{\theta}}(\mathbf{x}) = \int P_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z}.$$

- La distribución conjunta comúnmente se factoriza como

$$P_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = P_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \cdot P_{\boldsymbol{\theta}}(\mathbf{z}).$$

- La probabilidad a posteriori está dada por

$$P_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{P_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{P_{\boldsymbol{\theta}}(\mathbf{x})}.$$

Dependencia local en MVLs

- Suposición: relación entre variables observadas se da únicamente a través de variables latentes

- Ejemplo (de Lazarsfeld and Henry): 1000 personas fueron encuestadas sobre si leen la revista A y B.

	Leyó A	No leyó A	Total
Leyó B	260	140	400
No leyó B	240	360	600
Total	500	500	1000

Dependencia local en MVLs

- Suposición: relación entre variables observadas se da únicamente a través de variables latentes

- Ejemplo (de Lazarsfeld and Henry): 1000 personas fueron encuestadas sobre si leen la revista A y B.

High education	Read A	Did not read A	Total
Read B	240	60	300
Did not read B	160	40	200
Total	400	100	500
Low education	**Read A**	**Did not read A**	**Total**
Read B	20	80	100
Did not read B	80	320	400
Total	100	400	500

Tipos de MVLs

- Los modelos de variables latentes se pueden clasificar por la naturaleza de sus variables latentes y observadas

	V. observadas	
V. latentes	**Continua**	**Categórica**
Continua	Análisis de factores	Teoría de la respuesta al reactivo
Discreta	Modelo de mezclas	Análisis de clases latentes

Modelos de mezclas

- Variable latente discreta $z \in \{1, \ldots, K\}$

$$z \sim Cat(\boldsymbol{\pi})$$

- K distribuciones base $P(\mathbf{x}|z = k) = f_k(\mathbf{x})$

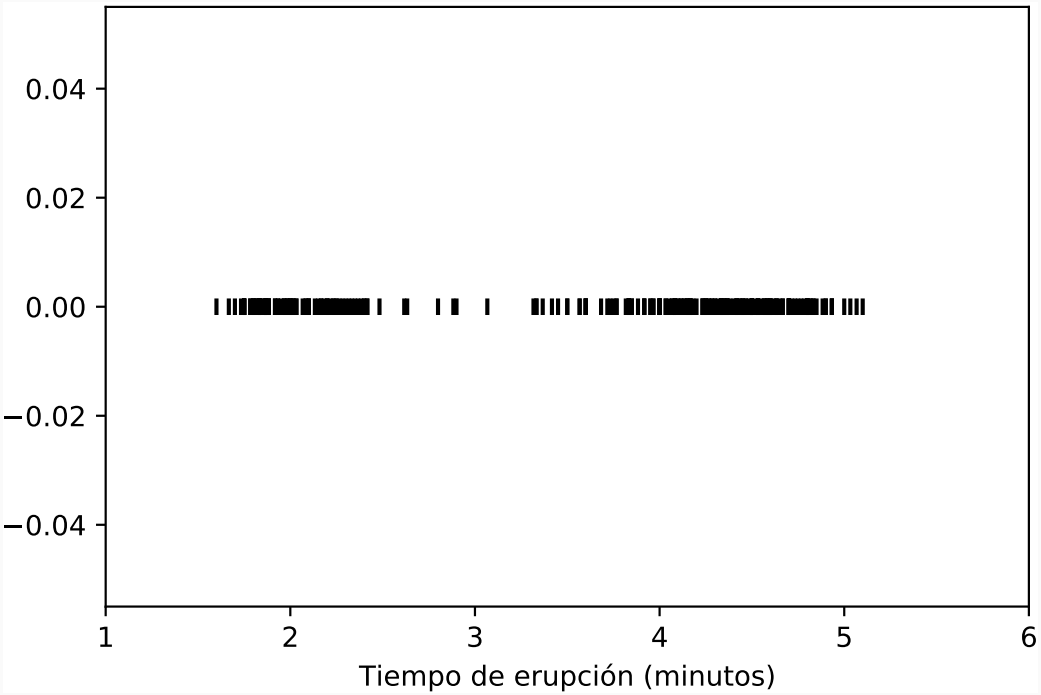
$$\mathbf{x}|z \sim f_k(\mathbf{x})$$

- Distribución de $\mathbf{x}$ se puede expresar como

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x})$$

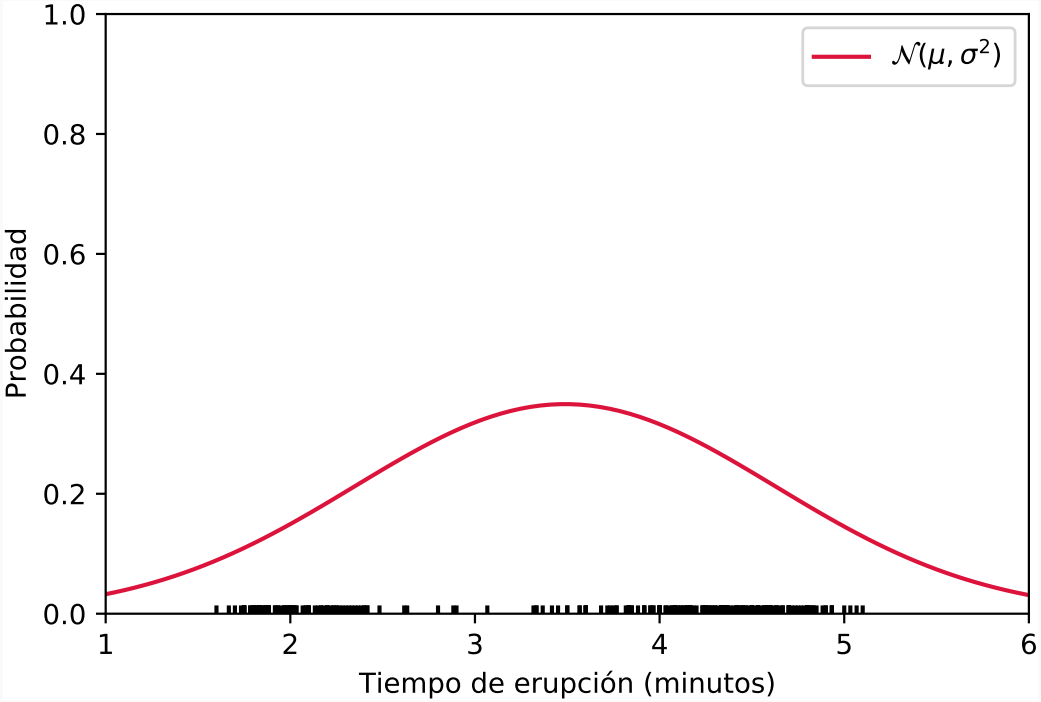
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



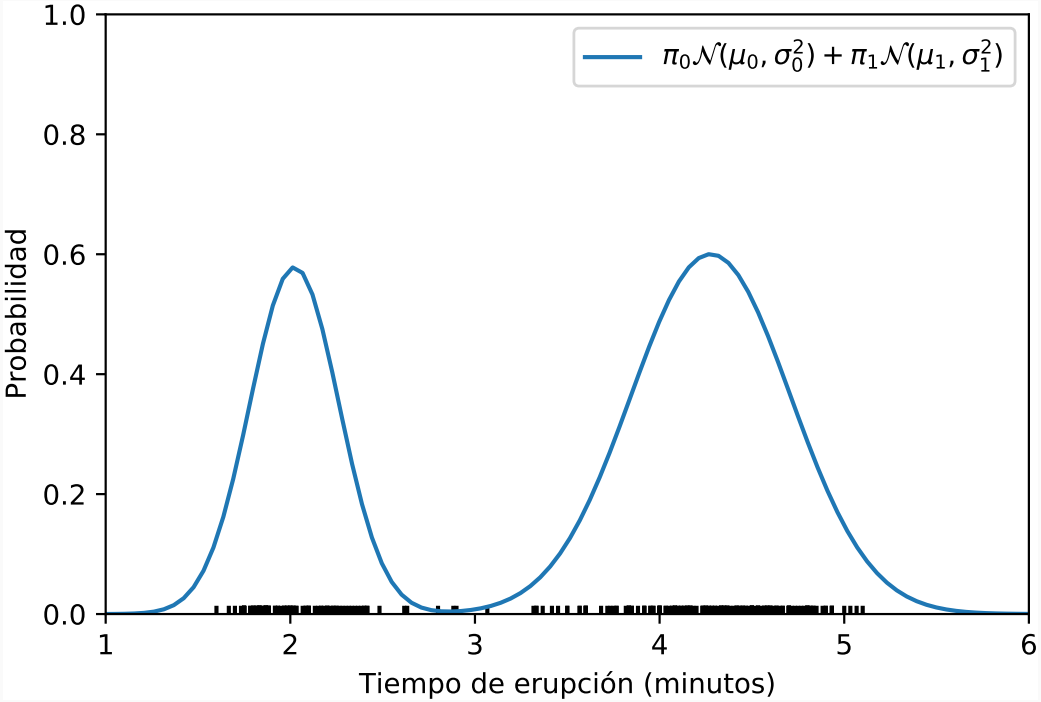
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



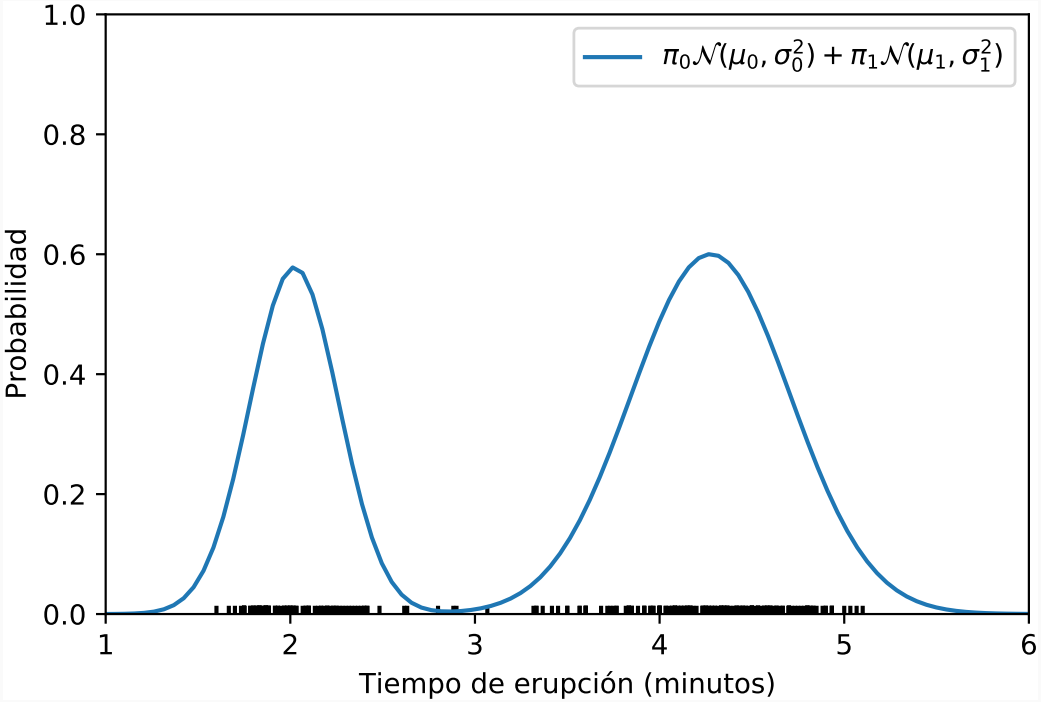
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



- ¿Cómo estimamos los parámetros?

Mezclas gaussianas

- K distribuciones base $f_k(\mathbf{x})$ gaussianas

$$\mathbf{x}|z \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Rightarrow P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- La verosimilitud logarítmica está dada por

$$\log\{P(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\} = \sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

- No hay solución cerrada analítica, necesitamos usar algoritmos de optimización iterativa.

Esperanza-Maximización

- Algoritmo para estimar parámetros por máxima verosimilitud o máximo a posteriori en problemas con datos faltantes y modelos de variables latentes

- Procedimiento general
  1. **Paso E**: inferir valores faltantes o de variales latentes
  2. **Paso M**: optimizar parámetros usando datos inferidos

EM para estimación por máxima verosimilitud

- Considera que el conjunto de ejemplos está dado por los valores tanto de las variables observadas como las variables latentes $\{\mathcal{D}, \mathsf{Z}\}$

- Busca encontrar los valores de los parámetros $\boldsymbol{\theta}$ que maximicen la verosimilitud logarítmica de $\{\mathcal{D}, \mathsf{Z}\}$

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \log \left\{ \sum_{\mathsf{Z}} P(\mathcal{D}, \mathsf{Z}|\boldsymbol{\theta}) \right\}$$

Esperanza: paso E

- Como los valores de las variables latentes Z no se conocen, se calcula la distribución a posteriori $P(\mathsf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo})$ con los parámetros actuales $\boldsymbol{\theta}^{viejo}$

$$P(\mathsf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo}) = \frac{P(\mathcal{D}|\mathsf{Z}, \boldsymbol{\theta}^{viejo})P(\mathsf{Z}|\boldsymbol{\theta}^{viejo})}{P(\mathcal{D}|\boldsymbol{\theta}^{viejo})}$$

Maximización: paso M

- Se asignan parámetros que maximizan la esperanza de la verosimilitud logarítmica usando $P(\mathsf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo})$

$$\begin{aligned}
\boldsymbol{\theta}^{nuevo} &= \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{viejo}) \\
&= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathsf{Z}|\mathcal{D},\boldsymbol{\theta}^{viejo}} \left[\log \left\{ \sum_{\mathsf{Z}} P(\mathcal{D}, \mathsf{Z}|\boldsymbol{\theta}) \right\} \right] \\
&= \sum_{\mathsf{Z}} P(\mathsf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo}) \log \{P(\mathcal{D}, \mathsf{Z}|\boldsymbol{\theta})\}
\end{aligned}$$

EM general

1. Initializa parámetros $\boldsymbol{\theta}$
2. **Paso E:** Evaluar $P(\mathsf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo})$
3. **Paso M:** Re-estimar parámetros

$$\boldsymbol{\theta}^{nuevo} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{viejo})$$

4. Repetir 2 y 3 hasta que se cumpla el criterio de convergencia

Distribución a posteriori para modelo de mezclas gaussianas

- Probabilidad a posteriori $P(z = k|\mathbf{x}^{(i)})(responsabilidad)$ está dada por

$$P(z^{(i)} = k|\mathbf{x}^{(i)}) = \gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

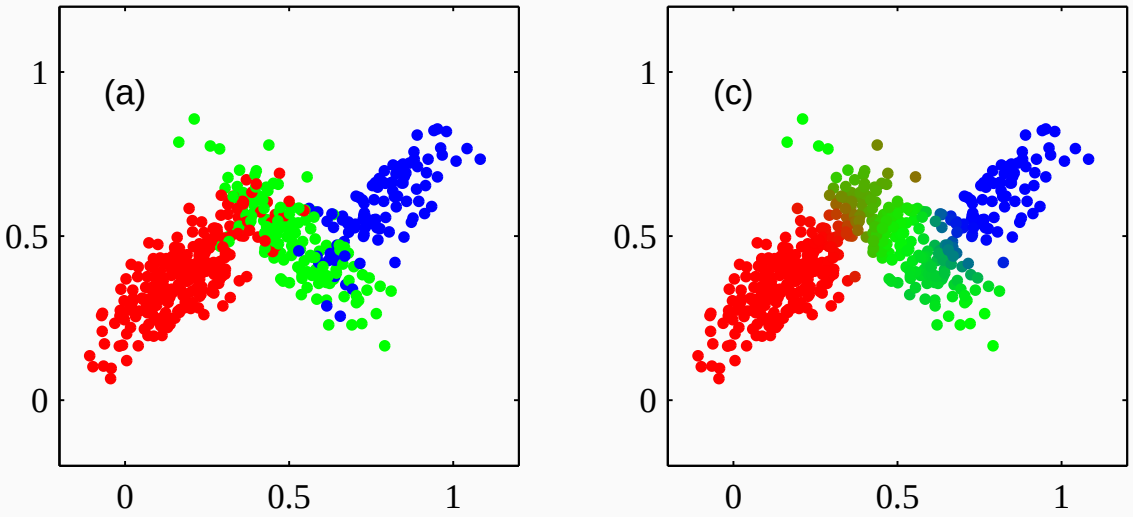


Imagen tomada de Bishop, PRML 2007

EM para modelos de mezclas gaussianas

1. Inicializa $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ y $\boldsymbol{\pi}_k$
2. **Paso E**: Evalúa responsabilidades con parámetros actuales

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(i)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(i)} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. **Paso M**: Recalcula parámetros $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ y $\boldsymbol{\pi}_k$ a partir de $\gamma(z_{nk})$

$$n_k = \sum_{i=1}^{n} \gamma(z_{ik})$$

$$\boldsymbol{\mu}_k^{nuevo} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \cdot \mathbf{x}^{(i)}$$

$$\boldsymbol{\Sigma}_k^{nuevo} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \cdot (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{nuevo})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{nuevo})^T$$

$$\boldsymbol{\pi}_k^{nuevo} = \frac{n_k}{n}$$

4. Evalúa verosimilitud logarítmica

$$\log \left\{ P(\mathcal{D} | \boldsymbol{\mu}^{nuevo}, \boldsymbol{\Sigma}^{nuevo}, \boldsymbol{\pi}^{nuevo}) \right\}$$

Modelo de mezclas gaussianas y EM en acción

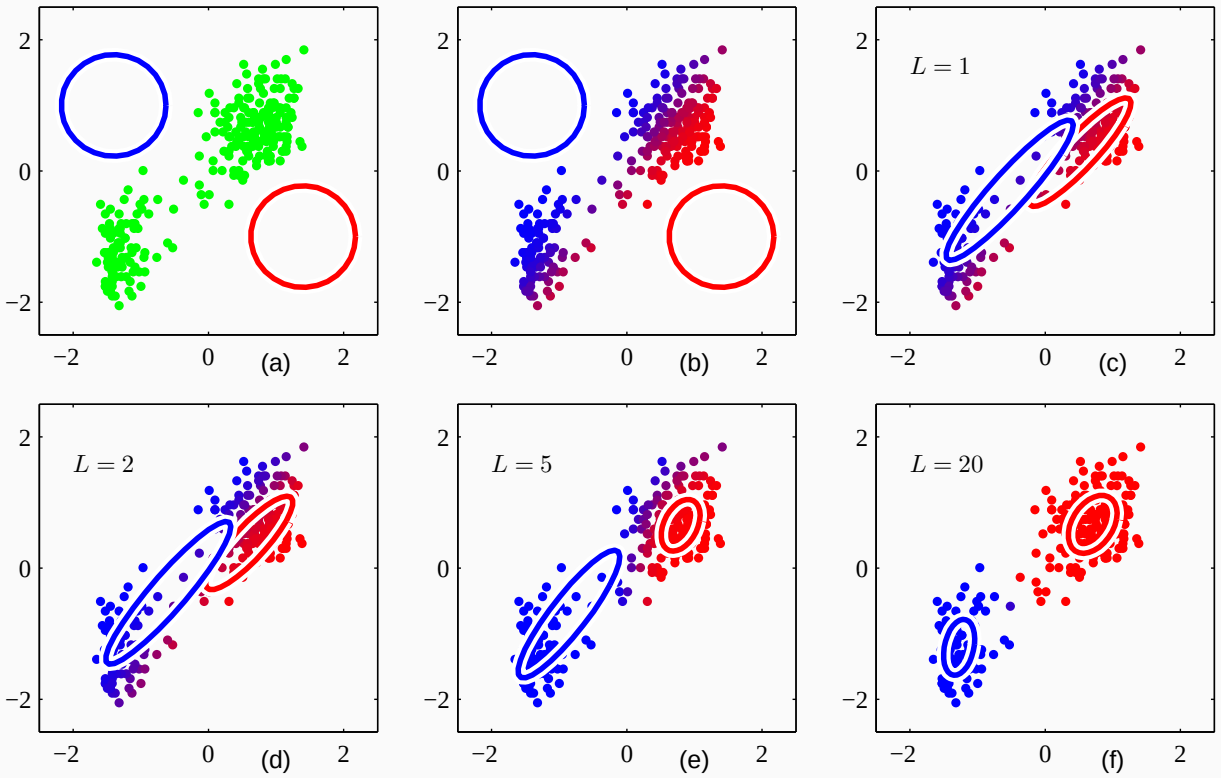


Imagen tomada de Bishop, PRML 2007

Modelo de mezclas de Bernoulli (análisis de clases latentes)

- Ejemplos con d variables binarias $\mathbf{x}^{(i)} = \{x_1, \ldots, x_d\}$

$$P(\mathbf{x}^{(i)}|\mathbf{q}) = \prod_{j=1}^{d} q_j^{x_j^{(i)}} \left(1 - q_j^{x_j^{(i)}}\right)^{\left(1-x_j^{(i)}\right)}$$

- Mezcla de K de estas distribuciones

$$P(\mathbf{x}^{(i)}|\mathbf{Q}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \cdot \left[\prod_{j=1}^{d} q_{kj}^{x_j^{(i)}} \left(1 - q_{kj}^{x_j^{(i)}}\right)^{\left(1-x_j^{(i)}\right)}\right]$$

donde $\mathbf{Q} = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K\}$ y $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$

EM para modelo de mezclas de Bernoulli

1. Inicializa $\boldsymbol{q}_k$ y $\boldsymbol{\pi}_k$
2. **Paso E**: Evalúa responsabilidades con parámetros actuales

$$\gamma(z_{ik}) = \frac{\pi_k \cdot \left[\prod_{j=1}^{d} q_{kj}^{x_j^{(i)}} \left(1 - q_{kj}^{x_j^{(i)}}\right)^{\left(1-x_j^{(i)}\right)} \right]}{\sum_{l=1}^{K} \pi_l \cdot \left[\prod_{j=1}^{d} q_{lj}^{x_j^{(i)}} \left(1 - \mu_{lj}^{x_j^{(i)}}\right)^{\left(1-x_j^{(i)}\right)} \right]}$$

3. **Paso M**: Re-estima parámetros $\boldsymbol{\mu}_k$ y $\boldsymbol{\pi}_k$ a partir de $\gamma(z_{nk})$

$$\boldsymbol{\mu}_k = \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}^{(i)}$$

$$\pi_k = \frac{n_k}{n}$$

$$n_k = \sum_{i=1}^{n} \gamma(z_{ik})$$

4. Evalúa verosimilitud logarítmica $\log P(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\pi})$

Desventajas de EMV en MMG

- Cuando una media es exactamente igual a un ejemplo $\mathbf{x}^{(i)} = \boldsymbol{\mu}_k$, la verosimilitud logarítmica se vuelve infinita ya que $\sigma_k \to 0$ y $\mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^0$. [1]

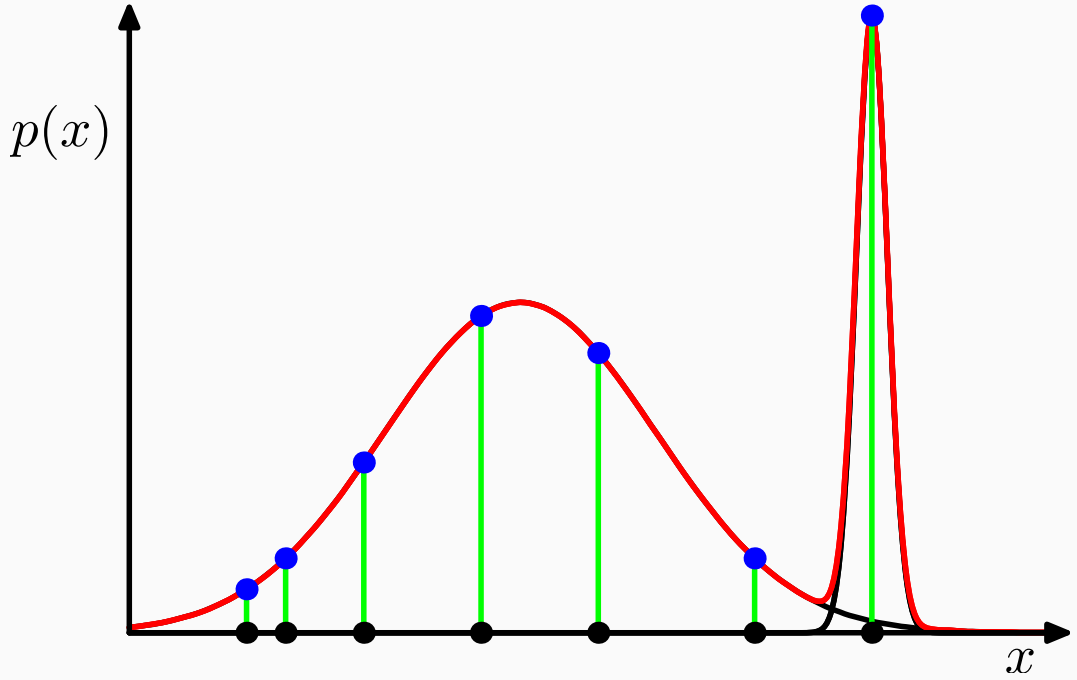


Figura tomada de Bishop, PRML 2007

[1] A esto se le conoce como el problema del colapso de la varianza.

Desventajas de EMV en MMG

- Singularidades
  - Cuando una media es exactamente igual a un ejemplo $\mathbf{x}^{(i)} = \boldsymbol{\mu}_k$, la verosimilitud logarítmica se vuelve infinita ya que $\sigma_k \to 0$ y $\mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^0$. [2]

- No identificabilidad
  - Existen $K!$ soluciones equivalentes.

[2] A esto se le conoce como el problema del colapso de la varianza.