Introducción al Aprendizaje No Supervisado

Hernán Felipe García Arias PhD

Especialización en Analítica de Datos
Facultad de Ingeniería
Universidad de Antioquia

Febrero, 2025

Contenido

Definiciones

- **Aprendizaje no supervisado.** En aprendizaje no supervisado no se cuenta con información sobre la variable de salida.
- Existen diferentes tipos de aprendizaje no supervisado: agrupamiento, estimación de densidad, y reducción de dimensionalidad.
- A continuación se estudia el problema de agrupamiento.
- Dos tipos de agrupamiento: agrupamiento determinístico y agrupamiento probabilístico.

Contenido

Algoritmo de las K-medias (I)

- Se busca identificar grupos de datos en un espacio multidimensional.
- Un grupo se puede entender como un conjunto de datos cuya distancia entre sí es pequeña comparada con la distancia a los puntos por fuera del grupo.
- Se supone un conjunto de vectores $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ en $\mathbb{R}^D$.
- Se introduce un conjunto de K vectores $\{\boldsymbol{\mu}_k\}_{k=1}^K \in \mathbb{R}^D$.
- Cada vector $\boldsymbol{\mu}_k$ es un prototipo asociado al k-ésimo grupo.

Algoritmo de las K-medias (II)

- Encontrar una asignación de los datos observados $\mathbf{X}$ a los K grupos.
- También se busca encontrar el conjunto de vectores $\boldsymbol{\mu}_k$ tal que se minimice la suma de los cuadrados de las distancias entre cada punto y su $\boldsymbol{\mu}_k$ más cercano.
- Se define una *medida de distorsión*

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

donde $r_{n,k}$ es una variable binaria que indica a cuál de los K grupos se asigna el vector de observación $\mathbf{x}_n$.

- **Objetivo:** encontrar valores de $\{r_{n,k}\}$ y $\{\boldsymbol{\mu}_k\}$ que minimicen J.

Algoritmo de las K-medias (III)

- Lo anterior se puede lograr mediante un proceso iterativo de dos pasos.
  - Se escogen los $\boldsymbol{\mu}_k$ y se minimiza J con respecto a los $\{r_{n.k}\}$ manteniendo los $\boldsymbol{\mu}_k$ fijos.
  - Se minimiza J con respecto a los μ_k manteniendo los $r_{n,k}$ fijos.

- Los dos pasos se repiten hasta lograr la convergencia.

- El primer paso se consigue seleccionando los $r_{n,k}$ como

$$r_{n,k} = \begin{cases} 1, & \text{si } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0, & \text{de otra forma.} \end{cases}$$

- Esto debido a que J es una función lineal de $r_{n,k}$, y los $\mathbf{x}_n$ son independientes.

Algoritmo de las K-medias (IV)

- En el segundo paso se obtiene la derivada de J con respecto a $\boldsymbol{\mu}_k$, y se iguala a cero,

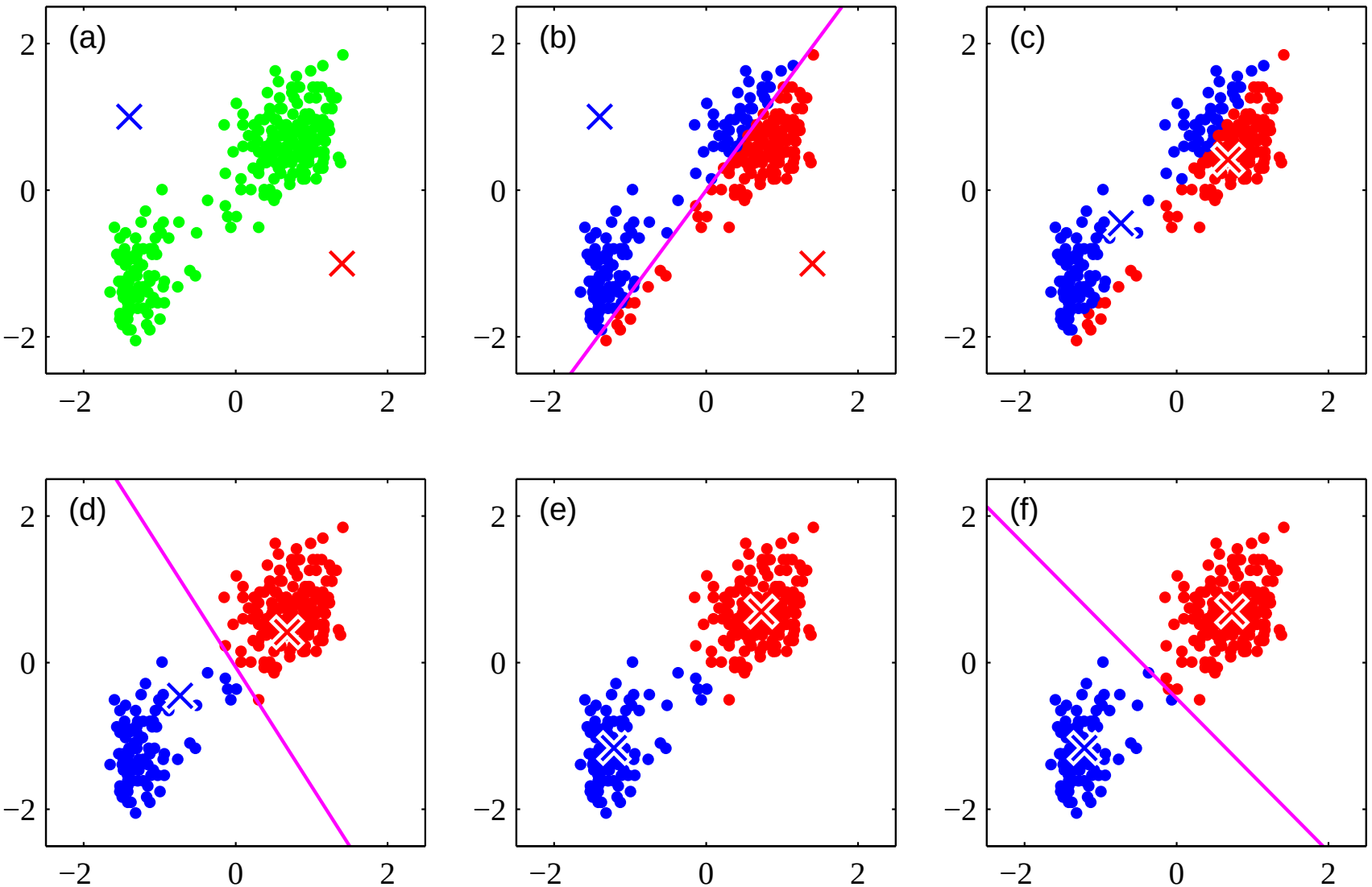
$$2\sum_{n=1}^{N} r_{n,k}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0.$$

- Despejando se obtiene,

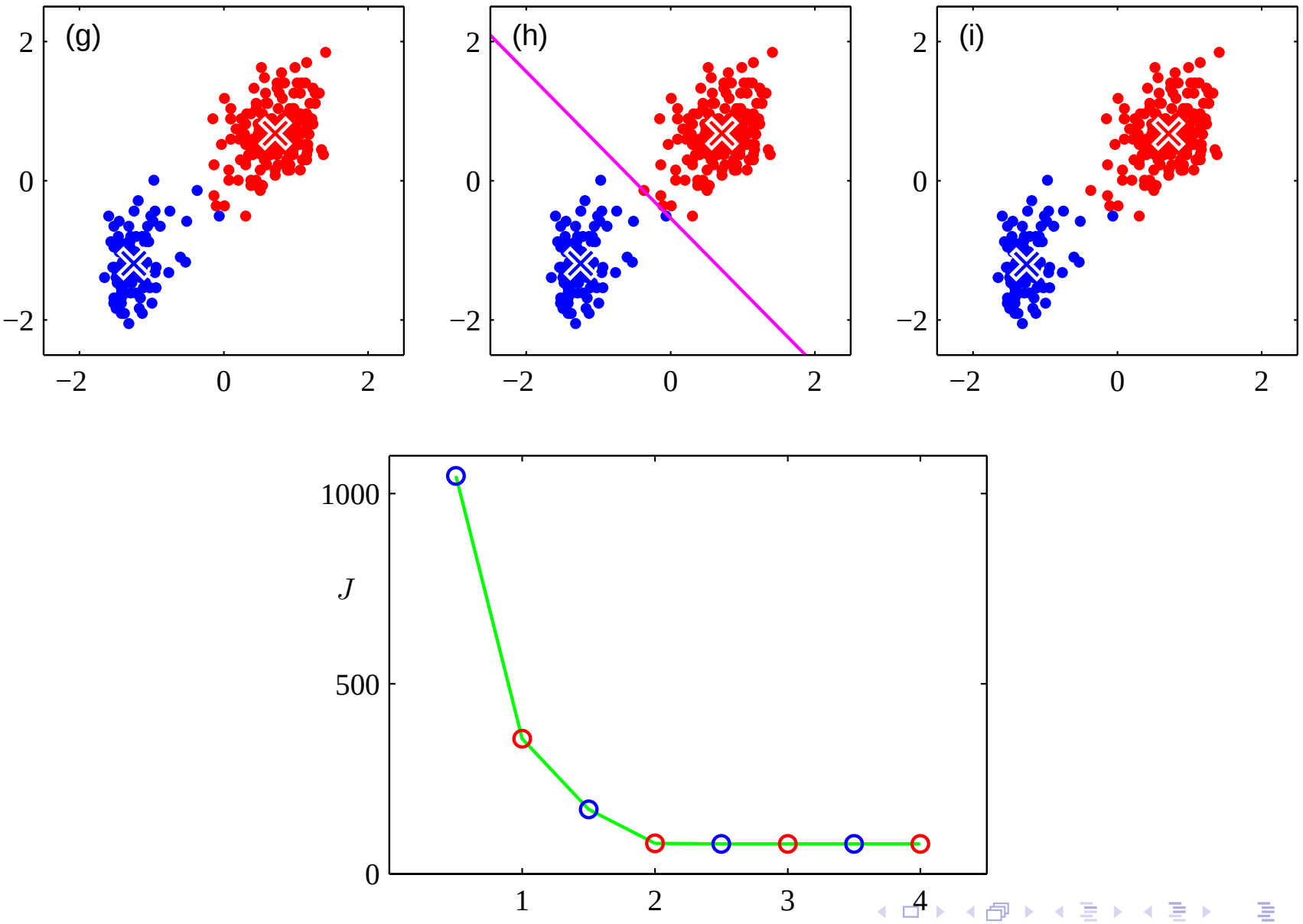
$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{n,k}\mathbf{x}_n}{\sum_{n=1}^{N} r_{n,k}}.$$

- Nótese que el denominador es igual al número de puntos asignados al grupo k.
- Igualmente $\boldsymbol{\mu}_k$ es la media de los datos $\mathbf{x}_n$ asignados al grupo k.
- Las dos fases de asignación de datos y cálculo de las medias se repiten hasta que no existan cambios en la asignación de grupos.

Algoritmo de las K-medias: ejemplo (I)



Algoritmo de las K-medias: ejemplo (II)



Algoritmo de las K-medias: otros espacios

- La distancia Euclidiana puede reemplazarse por una medida de disimilaridad $\mathcal{V}(\mathbf{x}, \mathbf{x}')$ que dependa de la aplicación y datos específicos.

- En este caso, la función de costo está dada como

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k).$$

Aplicación: segmentación de imágenes (I)

- El objetivo de la segmentación es dividir una imagen en regiones que tengan una apariencia visual razonablemente homogénea.
- Esas regiones suelen corresponder a objetos o partes de objetos.
- En esta aplicación, cada pixel se representa por un vector de intensidad [R,G,B], donde cada variable toma valores entre 0 y 1.
- Se usa agrupamiento por K-medias para diferentes valores de K.
- La imagen se redibuja cambiando el valor [R,G,B] de cada punto por el valor [R,G,B] dado por el centro μ_k al que ese punto ha sido asignado.

Aplicación: segmentación de imágenes (II)

$K = 2$

$K = 3$

$K = 10$

Original image

Aplicación: compresión de imágenes (I)

- Para los N datos, se almacena únicamente la identidad del grupo al que pertenece cada dato.
- También se almacenan los valores de los centros $\boldsymbol{\mu}_k$.
- Si se transmitiera la imagen en codificación [R,G,B] con 8 bits de precisión, para transmitir la imagen completa se necesitarían

$$24 \times N \text{ bits} .$$

- Si se corre primero K-medias sobre la imagen, la información a transmitir consistiría en la identidad del grupo al que pertenece cada pixel ($\log_2 K$ bits), más la codificación [R,G,B] de los K centros

$$24 \times K + N \log_2 K \text{ bits} .$$

Aplicación: compresión de imágenes (II)

- En el ejemplo anterior, las imágenes tienen dimensiones de 240x180 píxeles.
- Esto da un valor de $N = 43200$ muestras.
- Transmitir la imagen completa implicaría transmitir

$$24 \times N \text{ bits } = 1.036.800 \text{ bits.}$$

- Transmitir haciendo K-medias primero implicaría transmitir ($K = 2$),

$$24 \times K + N \log_2 K \text{ bits } = 43248 \text{ bits.}$$

Contenido

Contenido

Mezcla de funciones de probabilidad

- Una forma de aproximar funciones de probabilidad multimodales es a través de una mezcla de funciones de probabilidad.

- De las mezclas de funciones de probabilidad, la mezcla de Gaussianas es una de las más conocidas,

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

donde K es el número de componentes de la mezcla, y los parámetros π_k son probabilidades que satisfacen

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

Variable latente **z**

- Se introduce una variable aleatoria binaria de K dimensiones $\mathbf{z}$ con representación 1 de K.
- El vector $\mathbf{z}$ puede tomar uno de K estados, de acuerdo a cuál de los elementos es diferente de cero.
- La distribución marginal sobre $\mathbf{z}$ se especifica como

$$p(z_k = 1) = \pi_k.$$

- De forma compacta, esta distribución se escribe como

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

Distribución condicional de **x** dado **z**

- La distribución condicional de **x** dado un valor particular de **z**, es una Gaussiana

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- En forma compacta,

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Distribución marginal de $\mathbf{x}$

- La probabilidad conjunta está dada por $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.
- La distribución marginal de $\mathbf{x}$ se obtiene como,

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

- Si existen varios datos observados, a cada dato observado $\mathbf{x}_n$ le corresponde una variable latente $\mathbf{z}_n$.
- Este es una nueva formulación de la distribución de mezclas usando variables latentes, lo que permite trabajar con la distribución conjunta $p(\mathbf{x}, \mathbf{z})$.

Distribución condicional de **z** dado **x**

- Otra cantidad que juega un papel importante es la probabilidad condicional de **z** dado **x**.
- Esta probabilidad está dada como

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

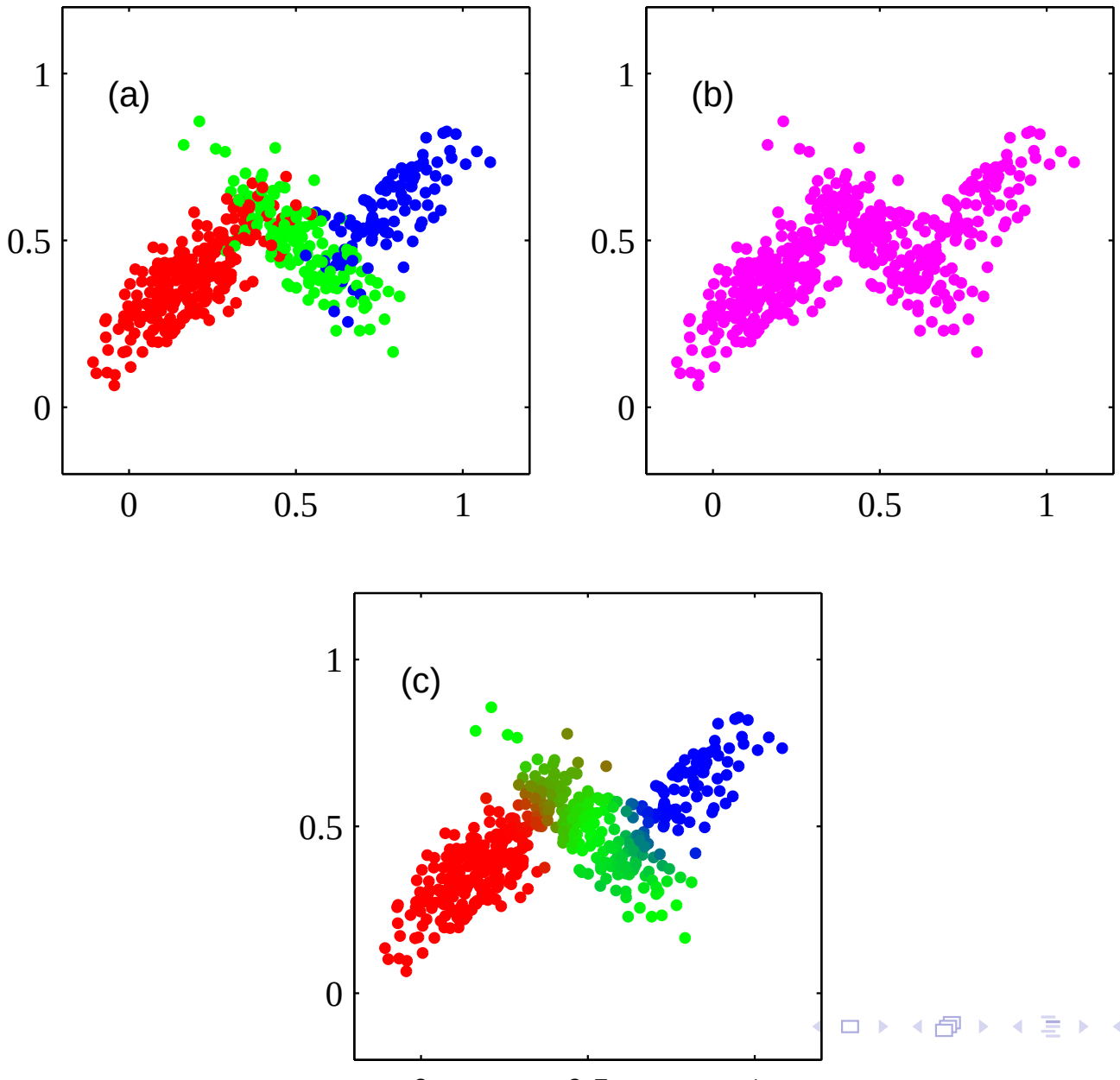
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- Se puede ver a π_k como la probabilidad a priori de que $p(z_k = 1)$ y a $\gamma(z_k)$ como la probabilidad a posteriori correspondiente una vez se ha observado **x**.
- Esta cantidad puede verse como la responsabilidad que la componente k asume para explicar la observación **x**.

Definición del tipo de datos

- Al conjunto $\{\mathbf{X}, \mathbf{Z}\}$ se le conoce como el conjunto completo de datos.
- Al conjunto de datos observados $\mathbf{X}$ se le conoce como los datos incompletos.
- Del conjunto $\{\mathbf{X}, \mathbf{Z}\}$ sólo se conoce $\mathbf{X}$. La única información sobre $\mathbf{Z}$ está en la función de probabilidad $p(\mathbf{Z}|\mathbf{X}, \theta)$.

Datos incompletos y completos



Contenido

Función de verosimilitud logarítmica

- Se parte de un conjunto de datos $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ que se desean modelar con una mezcla de Gaussianas.
- Este conjunto de datos se representan con una matriz $\mathbf{X}$ de dimensiones $N \times D$ y filas $\mathbf{x}_n^\top$.
- Similarmente, las variables latentes correspondientes se denotan por una matriz $\mathbf{Z}$ con filas $\mathbf{z}_n^\top$ y de dimensiones $N \times K$.
- La función de verosimilitud logarítmica está dada por

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

- Encontrar los parámetros $\theta^{\text{old}} = \{\{\pi_k\}_{k=1}^K, \{\mu_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K\}$, que maximicen la función de verosimilitud de los datos incompletos.

Algoritmo EM

Dada una distribución conjunta $p(\mathbf{X}, \mathbf{Z}|\theta)$, el objetivo es maximizar la función de verosimilitud $p(\mathbf{X}|\theta)$ con respecto a los parámetros θ.

1. Escoger un valor inicial para los parámetros θ^{old}.
2. **Paso E**. Evaluar $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$.
3. **Paso M**. Evaluar θ^{new} dada por

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}}),$$

donde

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

4. Verificar la convergencia de la función de verosimilitud o de los parámetros. Si no se satisface el criterio de convergencia, luego $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$ y volver al paso 2.

Mezcla de Gaussianas: Aplicación del paso E

- Comenzando con un valor de θ^{old} se calcula la probabilidad a posteriori de las variables latentes $\mathbf{Z}$ dados los datos $\mathbf{X}$ y los parámetros θ^{old}.

- La función de probabilidad $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ tiene como elementos $\gamma(z_{n,k})$.

- Las probabilidades $\gamma(z_{n,k})$ están dadas como

$$\gamma(z_{n,k}) \equiv p(z_{n,k} = 1|\mathbf{x}_n) = \frac{p(z_{n,k} = 1)p(\mathbf{x}_n|z_{n,k} = 1)}{\sum_{j=1}^{K} p(z_{n,j} = 1)p(\mathbf{x}_n|z_{n,j} = 1)}$$

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

- $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ es una tabla de dimensiones $N \times K$.

Mezcla de Gaussianas: Aplicación del paso M (I)

- Encontremos primero la función $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$.
- La función $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ está dada como

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$$

- La función de verosimilitud de los datos completos para la mezcla de Gaussianas está dada como

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

- La verosimilitud logarítmica está como

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}.$$

Mezcla de Gaussianas: Aplicación del paso M (II)

- La función $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ está dada como

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}.\end{aligned}$$

- Nótese en la ecuación anterior, que $\mathbb{E}_{\mathbf{Z}}[z_{nk}]$ coincide con $\gamma(z_{nk})$,

$$\mathbb{E}_{\mathbf{Z}}[z_{nk}] = \sum_{z_{nk}} z_{nk} p(z_{nk}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = p(z_{nk} = 1|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \gamma(z_{nk}).$$

- Dado $\gamma(z_{nk})$ la idea es ahora maximizar $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ con respecto a los parámetros $\boldsymbol{\theta} = \{\{\pi_k\}_{k=1}^K, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\Sigma}_k\}_{k=1}^K\}$.

Mezcla de Gaussianas: Aplicación del paso M (III)

- La maximización de $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ con respecto a π_k conduce a

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

donde $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

- La maximización de $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ con respecto a $\boldsymbol{\mu}_k$ conduce a

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

- La maximización de $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ con respecto a $\boldsymbol{\Sigma}_k$ conduce a

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

Resumen

1. Se escoge un valor inicial para θ^{new}.
2. Paso **E**. Se calcula $\gamma(z_{nk})$, para $n = 1, \ldots, N$ y $k = 1, \ldots, K$.
3. Paso **M**. Se usan las fórmulas de actualización para π_k^{new}, $\boldsymbol{\mu}_k^{\text{new}}$ y $\boldsymbol{\Sigma}_k^{\text{new}}$, para $k = 1, \ldots, K$.
4. Se verifica la convergencia de la función de verosimilitud o de los parámetros. Si no se satisface el criterio de convergencia, luego $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$ y se repite desde el paso 2.

Ejemplo