Aprendizaje no supervisado y análisis de agrupamientos

Luis Vázquez

GTI - IIE

Facultad de Ingeniería

Universidad de la República

Aprendizaje no supervisado

- Partimos de conjunto de entrenamiento $\{x_j\,; \quad j = 1, 2, \cdots, N\}$ tq. no conocemos sus etiquetas de clase γ_i.
- Cuando:
  - no disponemos del conocimiento de un experto
  - el etiquetado de cada muestra individual es impracticable
- **Ejemplo:** Aplicaciones con sensores remotos, como imágenes satelitales de terrenos.

  Costoso o imposible recoger información real del tipo de suelo sensado en cada punto.

Heurística del aprendizaje

- Dado conjunto de entrenamiento suficientemente grande podríamos inferir

$$p(\mathbf{x}) = \sum_{i=1}^{m} P(\omega_i) p(\mathbf{x}|\omega_i)$$

- Si la densidad conjunta es multimodal cada modo debería corresponder a cada clase presente.
- Si cada clase fuera normal se podría recuperar parámetros de cada distribución y luego seguir con el diseño del clasificador ...
- **Problema práctico:** estimar y analizar densidad conjunta en espacio de dimensión n:

Impracticable por su complejidad computacional.

Distribución conjunta multimodal

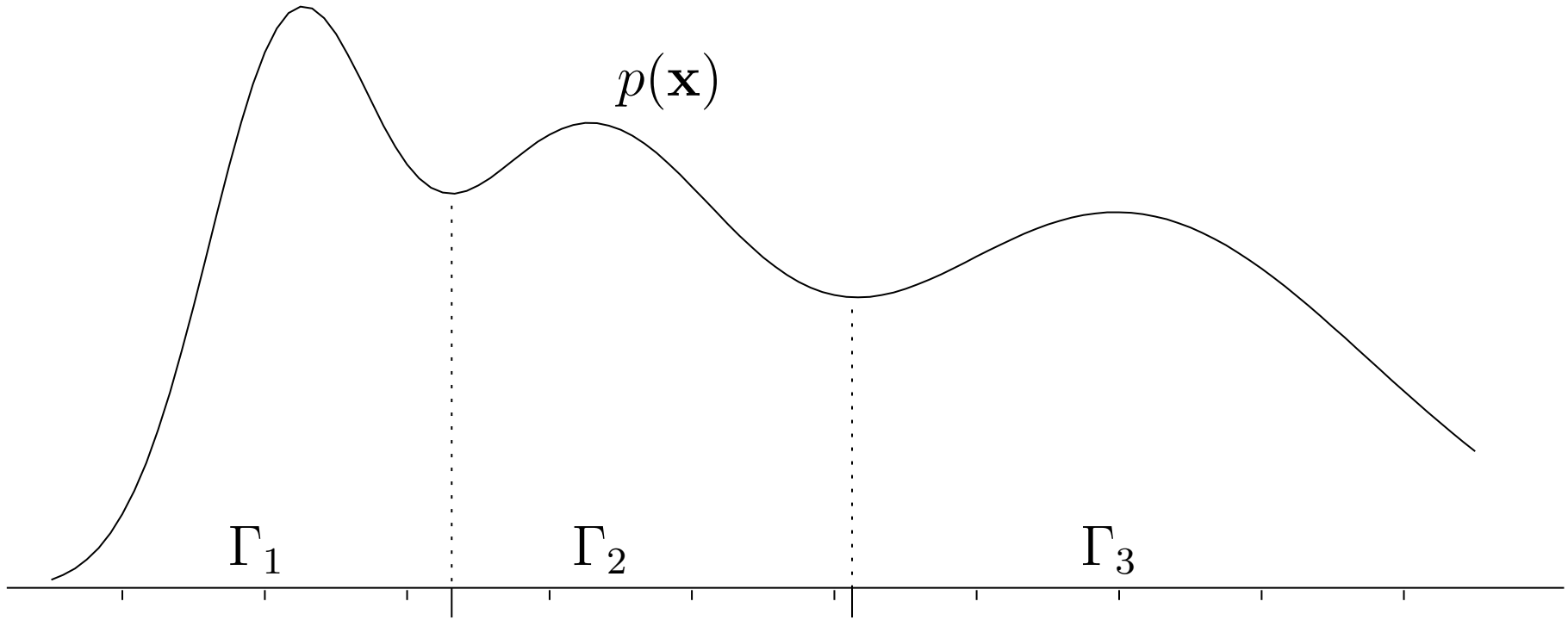


Figure 1: **Regiones asociadas a cada clase**

Heurística del aprendizaje II

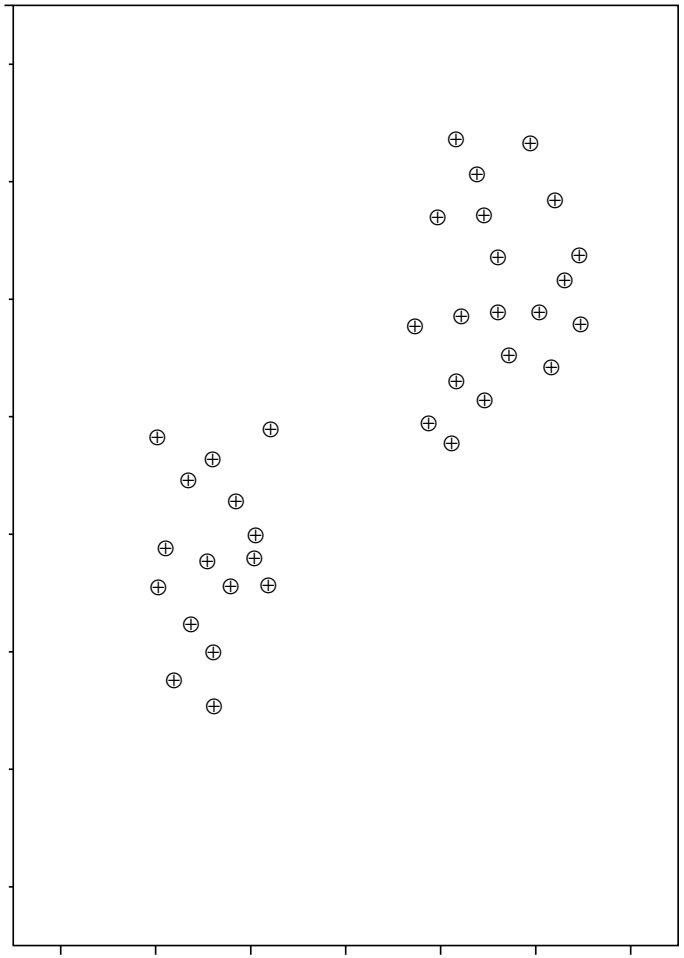
- Opción alternativa: clasificar los patrones en el conjunto de entrenamiento y usar estas etiquetas para un aprendizaje supervisado
- Solo necesitamos un método indirecto que permita etiquetar automáticamente los patrones de entrenamiento.
- Queremos forma de particionar el conjunto en clases con misma etiqueta

métodos de agrupamiento o *clustering*

Análisis de Agrupamientos

- Nos interesan las modas en la funcion de densidad conjunta $p(\mathbf{x})$
- **Intuitivamente:** estarán asociadas a regiones con alta densidad en el espacio de observación.
- Proposito de las técnicas de agrupamiento:
  - detectar y agrupar ***enjambres de puntos***
  - analizar y extraer la estructura presente en el conjunto de muestras de entrenamiento.
- **Conjunto de datos bien estructurado:** contiene varias regiones de alta densidad separadas por otras relativamente vacias o con poca densidad.

Estructura de datos



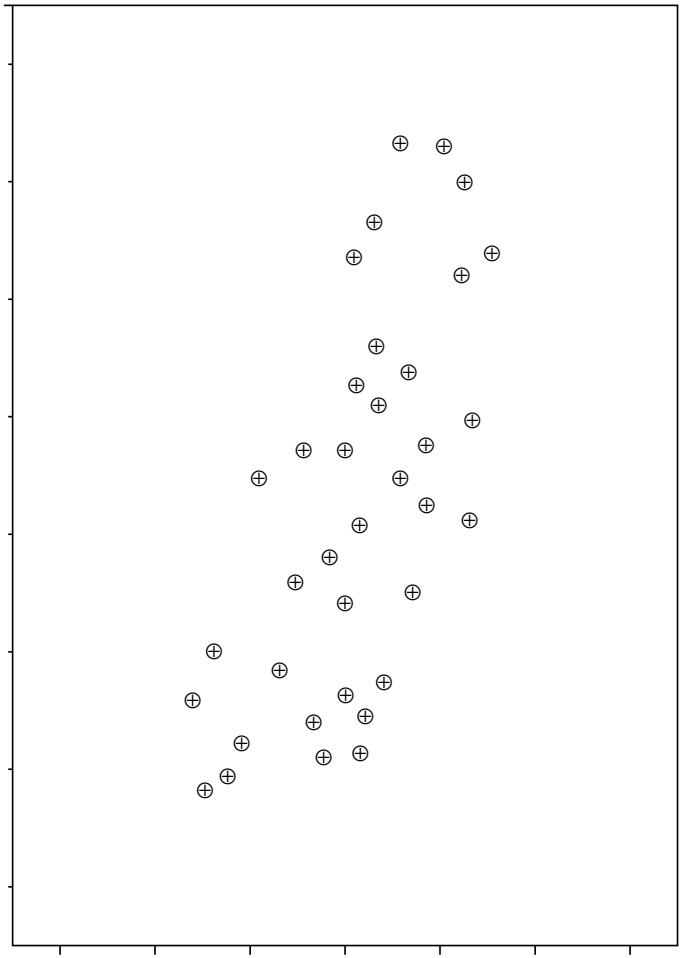


Figure 2: **Datos structurados y no estructurados**

Medidas de Similitud

- Para decidir si $\mathbf{x}$ pertenece a un agrupamiento necesitamos ***medida de proximidad o similitud***.
- Se han sugerido gran número de tales medidas pero las más usadas son medidas de distancia, en particular la distancia Euclideana.
- Medir afinidad de un punto a un agrupamiento:
  - similitud con otros puntos en el agrupamiento
  - similitud con modelo definido para el agrupamiento
- **Ejemplo:** representar grupo i por su media μ_i. La afinidiad entre un punto y el grupo podría ser:

$$d(\mathbf{x}, \mu_i) = \left[(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i)\right]$$

Criterios de Agrupamiento

- Problema: particionar un conjunto de puntos en agrupamientos *óptimos*
- Necesitamos algun ***criterio de agrupamiento***
- Permite definir *cuantitativamente* cuando una partición es mejor que otra.
- El criterio de agrupamiento que definamos y el algoritmo de agrupamiento asociado estarán relacionados con la medida de similitud usada.

Algoritmos de Agrupamiento

- Conocidos medida de similitud y criterios encontrar particion *óptima* $\Rightarrow$ **problema de optimización!**
- Problema práctico: considerar todas las particiones

$$\frac{k^M}{k!} \quad \text{siendo} \quad \begin{cases} k & \text{nro. de grupos} \\ M & \text{nro. de patrones} \end{cases}$$

Inviable en la práctica!!

$$M = 100,\ k = 5 \quad \Rightarrow \quad 10^{67}\text{!!}$$

- Se deben buscar métodos no exaustivos, aunque no siempre garanticen óptimos globales.

Clasificación de Algoritmos

- Por estrategia de agrupamiento
  - Aglomerativos o incrementales: ***botton up***
  - Divisivos o decrementales: ***top down***
  - Mixtos: ***creación y mezcla***
- Por criterio de busqueda del óptimo
  - Heurísticos: no se optimiza la función criterio
  - Por Optimización: se usa una función criterio
- Por información a priori
  - Número de clases conocido
  - Número de clases desconocido

Medidas Objetivas Importantes

- ***Para mezclar***
  - Distancias entre los centros de los agrupamientos
- ***Para dividir***
  - Varianza interna en un agrupamiento
  - Distancia entre puntos extremos de un agrupamiento

Algoritmo de k-medias

- El conjunto de datos $\mathbf{X}$ contiene k agrupamientos $\mathbf{X}_i$ que pueden representarse adecuadamente con su valor medio μ_i.
  - *Medida de similitud:* distancia Euclideana a μ_i
  - *Criterio de agrupamiento:* suma total de la distancia cuadrática de cada punto al vector medio de su agrupamiento.
- **Objetivo del algoritmo:**

  Encontrar entre todas las particiones de $\mathbf{X}$ en k conjuntos $\{\mathbf{X}_i \; ; \quad i = 1, 2, \cdots, k\}$ aquella que minimiza el criterio de agrupamiento.

Algoritmo de k-medias: Planteo Formal

- Encontrar los agrupamientos $\{\mathbf{X}_i\}$ que minimizan

$$J = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k} \sum_{j=1}^{N_i} d(\mathbf{x}_{ij}, \mu_i) \quad \mathbf{x}_{ij} \in \mathbf{X}_i,\ N_i = \#\mathbf{X}_i$$

  entre todas las posibles k-particiones de $\mathbf{X}$.

- Considero el efecto de un cambio atómico en la configuración de agrupamientos:

  sacar $\mathbf{x}$ *de* $\mathbf{X}_l$ *y pasarlo a otro grupo* $\mathbf{X}_r$.

- Esta reasignación solo afecta los grupos l y r con:

$$\bar{\mu}_l = \mu_l + \frac{1}{N_l - 1}(\mu_l - \mathbf{x}) \qquad \bar{\mu}_r = \mu_r - \frac{1}{N_r + 1}(\mu_r - \mathbf{x})$$

Deducción

- Calculamos el valor medio de $\mathbf{X}_i$ antes y despues de la reasignación

$$\boldsymbol{\mu}_l = \frac{1}{N_l}\sum_{j=1}^{N_l}\mathbf{x}_j \qquad \bar{\boldsymbol{\mu}}_l = \frac{1}{N_l - 1}\left(\sum_{j=1}^{N_l}\mathbf{x}_j - \mathbf{x}\right)$$

- De aqui resulta

$$(N_l - 1)\bar{\boldsymbol{\mu}}_l = N_l\,\boldsymbol{\mu}_l - \mathbf{x} \;\Rightarrow\; \bar{\boldsymbol{\mu}}_l = \frac{N_l}{N_l - 1}\,\boldsymbol{\mu}_l - \frac{1}{N_l - 1}\,\mathbf{x}$$

$$\Rightarrow \quad \bar{\boldsymbol{\mu}}_l = \boldsymbol{\mu}_l + \frac{1}{N_l - 1}(\boldsymbol{\mu}_l - \mathbf{x})$$

- Análogamente se verifica la segunda identidad.

Deducción de cambio atómico global

- Para hallar el cambio global en J basta calcular los cambios en J_l y J_r
- Para el nuevo agrupamiento l-ésimo tengo

$$\bar{J}_l = \sum_{j=1}^{N_l-1} d(\mathbf{x_j}, \bar{\boldsymbol{\mu}}_l) = \sum_{j=1}^{N_l-1} (\mathbf{x}_j - \bar{\boldsymbol{\mu}}_l)^T (\mathbf{x}_j - \bar{\boldsymbol{\mu}}_l) =$$

$$= \sum_{j=1}^{N_l} \Big(\mathbf{x}_j - \boldsymbol{\mu}_l + \frac{\boldsymbol{\mu}_l - \mathbf{x}}{N_l - 1}\Big)^T \Big(\mathbf{x}_j - \boldsymbol{\mu}_l + \frac{\boldsymbol{\mu}_l - \mathbf{x}}{N_l - 1}\Big) - \Big\|\mathbf{x} - \boldsymbol{\mu}_l + \frac{\boldsymbol{\mu}_l - \mathbf{x}}{N_l - 1}\Big\|^2$$

$$= J_l - \frac{2}{N_l - 1}(\boldsymbol{\mu}_l - \mathbf{x}) \underbrace{\sum_{j=1}^{N_l} (\mathbf{x}_j - \boldsymbol{\mu}_l)}_{0} + \frac{N_l - N_l^2}{(N_l - 1)^2}(\boldsymbol{\mu}_l - \mathbf{x})^T (\boldsymbol{\mu}_l - \mathbf{x})$$

Cambio global en reasignación

- Luego de agrupar resulta:

$$\begin{cases} \bar{J}_l = J_l - \dfrac{N_l}{N_l - 1}(\boldsymbol{\mu}_l - \mathbf{x})^T(\boldsymbol{\mu}_l - \mathbf{x}) = J_l - \dfrac{N_l}{N_l - 1}\, d(\mathbf{x}, \boldsymbol{\mu}_l) \\ \bar{J}_r = J_r + \dfrac{N_r}{N_r - 1}(\boldsymbol{\mu}_r - \mathbf{x})^T(\boldsymbol{\mu}_r - \mathbf{x}) = J_r + \dfrac{N_r}{N_r - 1}\, d(\mathbf{x}, \boldsymbol{\mu}_r) \end{cases}$$

- Obtenemos descenso en $J \;\Leftrightarrow\; (J_l - \bar{J}_l) > (\bar{J}_r - J_r)$

$$\Leftrightarrow \quad \frac{N_l}{N_l - 1}\, d(\mathbf{x}, \boldsymbol{\mu}_l) > \frac{N_r}{N_r - 1}\, d(\mathbf{x}, \boldsymbol{\mu}_r)$$

- Si N_l y N_r son grandes $\Rightarrow$ esta desigualdad se cumple $\Leftrightarrow \mathbf{x}$ esta más cerca de $\boldsymbol{\mu}_r$ que de $\boldsymbol{\mu}_l$.

El algoritmo k-mean

De lo anterior se deducen los pasos del algoritmo

- **Paso 1:** Elegir una partición inicial en k grupos al azar
- **Paso 2:** Calcular las medias $\boldsymbol{\mu}_i$ de los *clusters*
- **Paso 3:** Seleccionar secuencialmente un punto $\mathbf{x}$ del conjunto y, si corresponde, reasignarlo al *cluster* que minimiza $d(\mathbf{x}, \boldsymbol{\mu}_i)$
- **Paso 4:** Si no hay reasignaciones en todo el conjunto terminar; sino volver al **Paso 2**.

Algoritmos de agrupamiento dinámico

- El k-mean es caso particular de una familia de algoritmos

  Dynamical Clustering Algorithms

- El grupo $\mathbf{X_i}$ se representa por un nucleo $K_i(\mathbf{X_i})$:

  Puede ser un una función paramétrica, un conjunto de puntos u otro modelo

- Se define una medida de similitud $\Delta(\mathbf{x}, K_i)$ y el criterio de agrupamiento

$$J = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \Delta(\mathbf{x}_{ji}, K_i) \qquad \mathbf{x}_{ji} \in \mathbf{X_i}$$

Algoritmo de agrupamiento genérico

Por analogía el algoritmo de agrupamiento dinámico genérico sería

- **Paso 1:** Elegir una partición inicial en k grupos al azar
- **Paso 2:** Calcular el modelo $K_i(\mathbf{X_i})$ para cada *cluster*
- **Paso 3:** Seleccionar uno a uno cada punto $\mathbf{x}$ del conjunto y, si corresponde, reasignarlo al *cluster* con mayor afinidad
- **Paso 4:** Si no hay reasignaciones terminar el algoritmo; sino volver al **Paso 2**.

Ejemplo de Modelo Gassiano

- Un ejemplo importante de representacion de un cluster es el *modelo gaussiano*

$$K_i(\mathbf{x}, \mathbf{X_i}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\Big)\cdot\frac{N_i}{N}$$

- La medida de similitud asociada es

$$\Delta(\mathbf{x}_{ji}, K_i) = (\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log|\boldsymbol{\Sigma}_i| - 2\log N_i$$

- Este modelo se usa para recuperar las componentes de una mezcla de distribuciones normales.