Aprendizaje no Supervisado

Aprendizaje Automático

Ingeniería Informática

Fernando Fernández Rebollo y Daniel Borrajo Millán

Grupo de Planificación y Aprendizaje (PLG)
Departamento de Informática
Escuela Politécnica Superior
Universidad Carlos III de Madrid

27 de febrero de 2009

En Esta Sección:

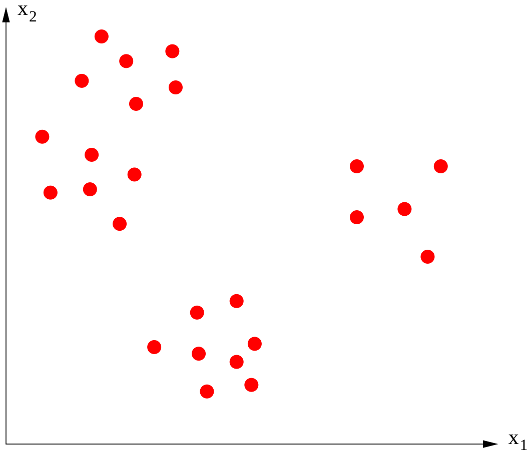
Aprendizaje no Supervisado

- Distintos objetivos enmarcados dentro del a aprendizaje no supervisado:
  - Agrupación: dados dos ejemplos sin etiquetar (sin campo de clase), agruparlos siguiendo algún criterio predefinido.
  - Generación de jerarquías: dados unos datos en un mismo nivel, generar jerarquías que organicen dichos datos
  - Reducción de dimensionalidad: dados unos datos, reducir la dimensión o número de atributos que caracterizan dichos datos
  - Visualización: dados unos datos con representación compleja, permitir su visualización

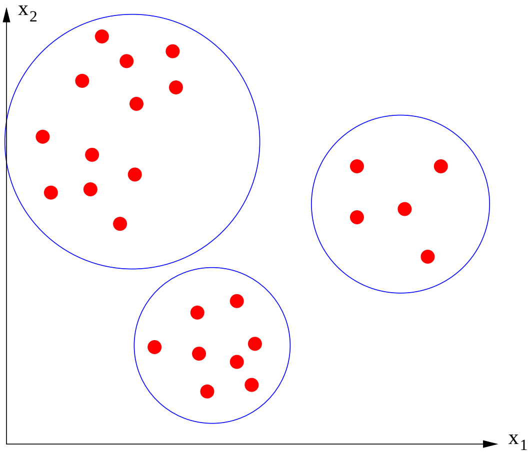
Aprendizaje no Supervisado: Agrupación

- Objetivo: dados dos ejemplos sin etiquetar (sin campo de clase), agruparlos siguiendo algún criterio predefinido
- Ese criterio suele venir por:
  - Aprendizaje paramétrico: parámetros asumidos, por ejemplo, que los datos siguen una determinada densidad de probabilidad
  - Aprendizaje no paramétrico: alguna medida de distancia
- Cuestiones principales:
  - ¿Cuántos grupos hay?
  - ¿Cómo se decide a qué grupo pertenece una nueva instancia?

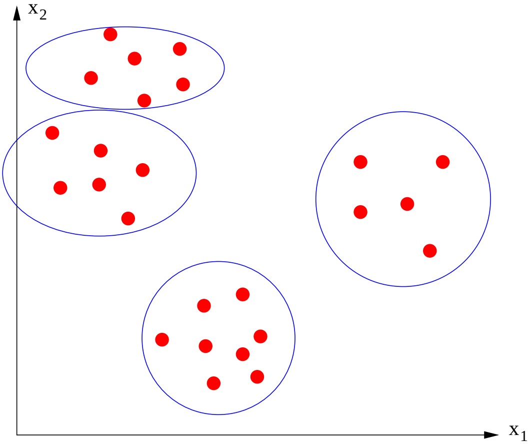
Ejemplo



Ejemplo



Ejemplo



Clasificación Bayesiana y Aprendizaje no Supervisado

- Recordamos de la Teoría Bayesiana que:

$$P(\omega_i|\vec{x},\chi) = \frac{p(\vec{x}|\omega_i,\chi_i)P(\omega_i)}{\sum_{j=1}^{c} p(\vec{x}|\omega_j,\chi_j)P(\omega_j)} \quad (37)$$

- Donde estamos asumiendo que:
  - Diponemos de conocimiento del dominio que nos permite parametrizar esas densidades de probabilidad (por ejemplo, que siguen una distribución normal)
  - Diponemos de un conjunto de entrenamiento, χ, del que podemos aprender los parámetros de las funciones de densidad

Clasificación Bayesiana y Aprendizaje no Supervisado

- Por tanto, podemos calcular el estimador de máxima verosimilitud del conjunto de parámetros
- El aprendizaje no supervisado de una mezcla de distribuciones es equivalente al aprendizaje supervisado de los parámetros de varias distribuciones
- Problema: ¿qué se hace antes, la asignación de clases o la estimación de los parámetros?

Algoritmo EM

- Permite estimar, dado un conjunto de datos generado con c distribuciones gausianas, las medias de las distribuciones: $< \mu_1, \ldots, \mu_c >$
- Proceso iterativo:
  - Para cada $\vec{x}_j \in \chi$,
    - Calcular:
$$\omega_j = \arg_{\omega_i} \max p(\vec{x}|\omega_i, \chi_i) P(\omega_i) \qquad (38)$$
    - Incluir $\vec{x}$ en χ_j
  - Paso 2: Recalcular $< \mu_1, \ldots, \mu_c >$, donde μ_i es el estimador de máxima verosimilitud de la distribución normal i:
$$\mu_j = \frac{1}{|\chi_j|} \sum_{\vec{x} \in \chi_j}^{|\chi_j|} \vec{x} \qquad (39)$$

Métodos no Paramétricos

- Se basan en agrupar las instancias de entrenamiento siguiendo distintas medidas de distancia/error:
  - Distancia euclídea:

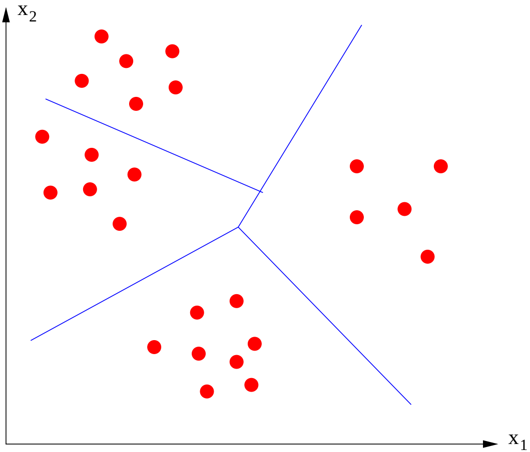
$$d(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_{r=1}^{n} (\vec{x}_i[r] - \vec{x}_j[r])^2} \tag{40}$$

  - Distancia Euclídea Ponderada:

$$d(\vec{x}_i, \vec{x}_j, \vec{w}) = \sqrt{\sum_{r=1}^{n} \vec{w}[r](\vec{x}_i[r] - \vec{x}_j[r])^2} \tag{41}$$

- Dos tipos:
  - Técnicas de particionamiento
  - Técnicas aglomerativas

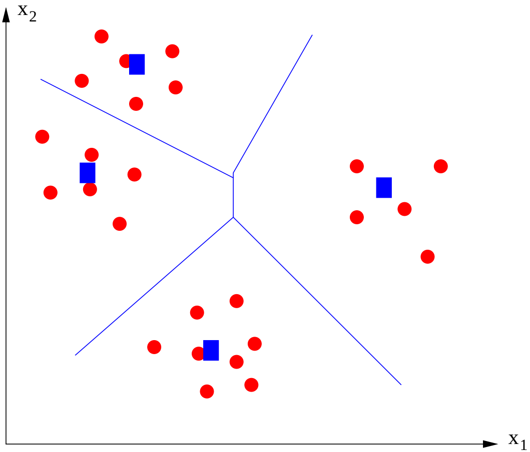
Ejemplo Técnicas de Particionamiento



Algoritmo de Lloyd o k-medias

- Realiza una discretización o particionamiento del espacio
- Genera regiones de Voronoi, que se definen mediante:
  - Una medida de distancia
  - Un conjunto de ejemplos representativos o prototipos
- Objetivo: generar el conjunto de prototipos que minimice una determinada medida de distorsión o error

Regiones de Voronoi



Algoritmo de Lloyd Generalizado (k-medias)

Algoritmo de Lloyd Generalizado (C_1, T, N)

1. Comenzar con un alfabeto inicial C_1. Sea $m = 1$.
2. Dado un alfabeto, C_m, ejecutar la Iteración de LLoyd para generar un nuevo alfabeto C_{m+1}.
3. Calcular la distorsión media para C_{m+1}.

$$D = \frac{1}{M} \sum_{j=1}^{M} \min_{\vec{y} \in C_{m+1}} (d(\vec{x}_j, \vec{y})), \tag{42}$$

4. Si ha cambiado en una pequeña cantidad solamente desde la iteración anterior, parar. Sino, hacer $m = m + 1$ e ir al paso 2.

Iteración de Lloyd (derivable de EM)

Iteración de Lloyd (C_m, T, N)

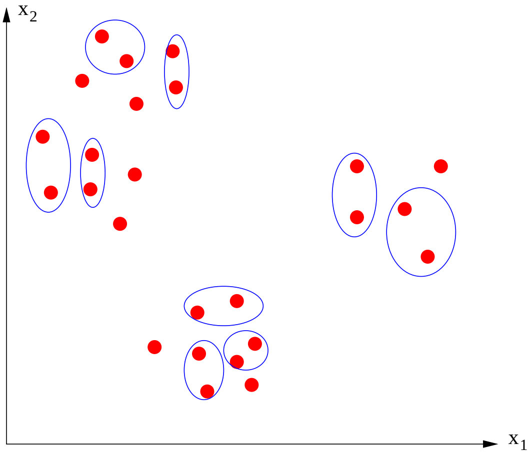
1. Dado un alfabeto $C_m = \{\vec{y}_i; i = 1, \ldots, N\}$ partir el conjunto de entrada T en particiones R_i usando la siguiente condición:

$$R_i = \{\vec{x} \in T : d(\vec{x}, \vec{y}_i) \leq d(\vec{x}, \vec{y}_j); \text{ para todo } j \neq i\}$$

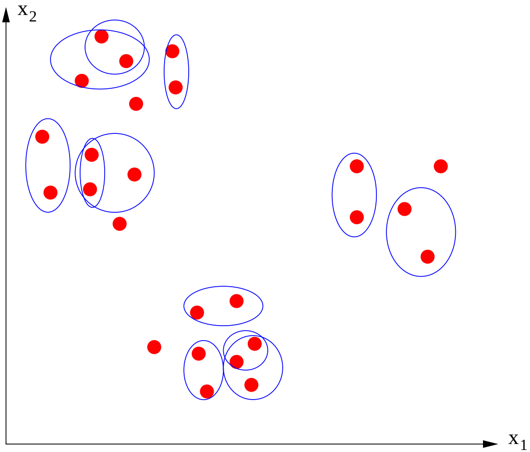
2. Calcular los centroides de cada partición para recalcular el alfabeto. Hacer $C_{m+1} = \{cent(R_i)\}$. Si se generó una celda vacía en el paso 1, se asignará un vector alternativo (en vez del cálculo del centroide) para esa celda.

$$cent(R) = \frac{1}{\|R\|} \sum_{\vec{x}_i \in R} \vec{x}_i \tag{43}$$

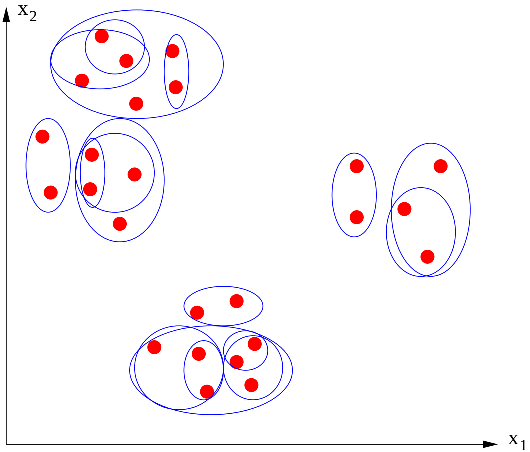
Ejemplo Técnicas Aglomerativas



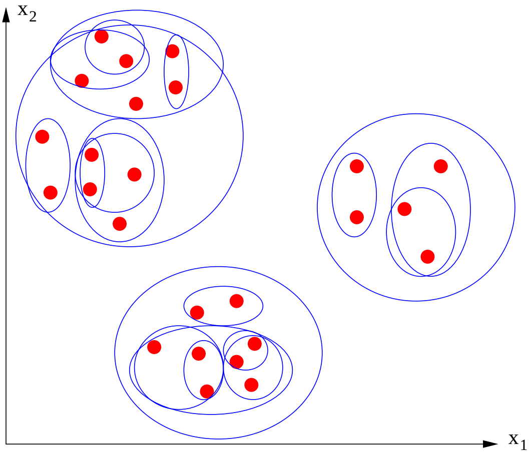
Ejemplo Técnicas Aglomerativas



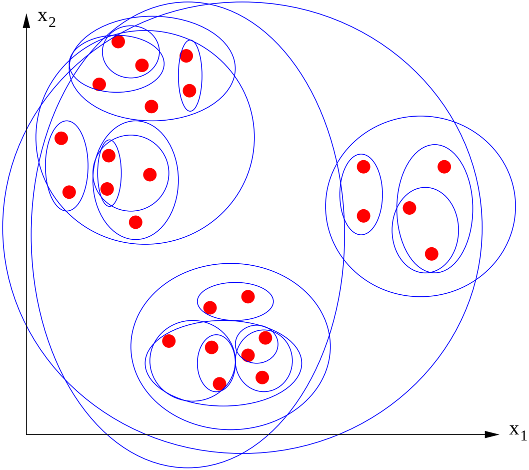
Ejemplo Técnicas Aglomerativas



Ejemplo Técnicas Aglomerativas



Ejemplo Técnicas Aglomerativas



Métodos Aglomerativos

- Se generan unos árboles denominados dendogramas
- Búsqueda hacia arriba:
    - Inicialmente, cada nodo representa un ejemplo.
    - Se repite N-1 veces (siendo N el número de ejemplos):
        - Se calcula la similitud entre todo par de ejemplos
        - Se agrupan los dos más cercanos
        - Se sustituyen los dos nodos por su representante o centroide o prototipo
- No sólo se generan grupos o clases, sino también una jerarquía entre ellos

Resumen

- Aprendizaje no supervisado o agrupación
- ¿Cuántos grupos hay?
- Métodos paramétricos y no paramétricos
- Papel de las medidas de distancia

Bibliografía

- Machine Learning, Tom Mitchell. McGraw Hill. 1997. Capítulo 6
- Vector Quantization and Signal Compression. Allen Gersho and Robert M. Gray. Kluwer Academic Publishers. 1992