Reconocimiento Estadístico de Formas

Técnicas no Supervisadas: "clustering"

J. Salvador Sánchez

Planteamiento del Problema

- Disponemos de un conjunto de muestras sin etiquetar y queremos agruparlas en clases. Hay dos posibilidades:
  - Técnicas paramétricas.
  - Técnicas de agrupamiento (*clustering*).

Técnicas Paramétricas

- Suponemos que se conoce la estructura de probabilidades del problema:
  - Existen c clases.
  - Las probabilidades a priori de cada clase son $P(\omega_j), j = 1, ..., c$.
  - Las funciones densidad de probabilidad son $p(x|\omega_j, \theta_j)$.

3

Técnicas de Agrupamiento

- Se aplican cuando no se conoce la forma de las densidades de probabilidad o cuando el aplicar un método paramétrico es muy complejo.
- Se define una función que nos indique, para cada posible partición, "lo bien agrupados" que están las muestras.
- Tenemos que definir una medida de similitud (o disimilitud) entre prototipos.

4

Técnicas de Agrupamiento

5

Medidas de Similitud

- La medida más obvia es la *distancia euclídea*:
  - Diremos que dos muestras pertenecen a una misma clase si la distancia euclídea entre ellas es menor que una cierta distancia umbral.

$$\theta_1 = \theta_2 \Leftrightarrow d(x_1, x_2) \leq d_u$$

6

Medidas de Similitud

- Sin embargo, este procedimiento puede llevar a diversos problemas:
  - Si la distancia umbral es demasiado grande, la mayor parte de las muestras quedarán en un mismo agrupamiento (clase).
  - Si la distancia umbral es demasiado pequeña, podemos obtener demasiados agrupamientos: cada prototipo podría llegar a definir una clase distinta.

7

Función Criterio de Agrupamiento

- Supongamos que tenemos un conjunto X de n muestras $x_1, ..., x_n$ que queremos repartir en c agrupamientos disjuntos $X_1, ..., X_c$.
- Definiremos una función criterio que mida la calidad del agrupamiento:
  - La función más utilizada es la *suma de los errores al cuadrado*.

8

Suma de los Errores al Cuadrado

- Sea n_i el número de muestras en X_i y sea m_i la media de este conjunto. Entonces,

$$J_e = \sum_{i=1}^{c} \sum_{x \in X_i} \|x - m_i\|^2$$

donde J_e es el error cuadrático total en el que se incurre si representamos las muestras de cada conjunto por sus medias.

9

Suma de los Errores al Cuadrado

- Es una función criterio que da buenos resultados cuando los conjuntos forman nubes compactas y están alejados los unos de los otros.
- Sin embargo, falla cuando existe una gran diferencia de puntos entre los conjuntos.

10

Generalización

- El criterio del error cuadrático puede escribirse como,

$$J_e = \frac{1}{2}\sum_{i=1}^{c} n_i s_i \quad \textit{donde} \quad s_i = \frac{1}{n_i^2}\sum_{x\in X_i}\sum_{x'\in X_j}\left\|x-x'\right\|^2$$

  donde *s* es la distancia cuadrada media de entre los prototipos de un conjunto.

- Así, se puede sustituir *s* por la *media*, la *mediana* o la *distancia máxima* entre los puntos de un conjunto.

11

Definición del Problema

- Una vez establecida la función criterio, el problema del agrupamiento queda bien definido:

  "Encontrar la partición del conjunto de muestras que hace que la función criterio sea mínima"

- Esto podría resolverse por enumeración, pero el número de posibles particiones crece de manera exponencial con el número de muestras ⇒ en la práctica, se emplean *métodos aproximados*.

12

Algoritmo de las Distancias Encadenadas (Chain-Map)

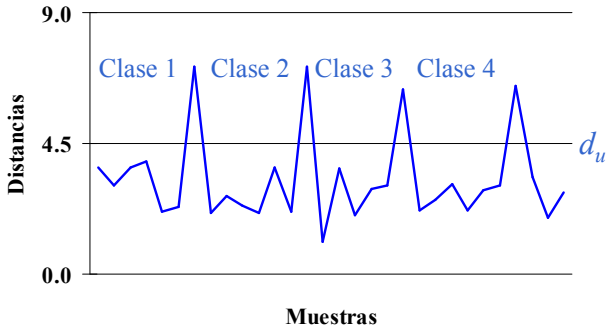
- Seleccionar una muestra al azar: x_i.
- Ordenar las muestras a partir de x_i según la sucesión:

  $$x_i(0), x_i(1), x_i(2), ..., x_i(k), x_i(k+1), ..., x_i(n-1)$$

  donde $x_i(j)$ es la muestra más próxima a $x_i(j-1)$.
- Calcular la sucesión de distancias $d_1, d_2, ..., d_{n-1}$ donde $d_j = \| x_i(j) - x_i(j-1) \|$.
- Cuando encontremos un $d_j > d_u$, crear una clase con las $x_i(j-1)$ muestras.

13

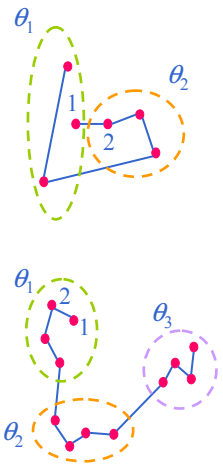
Algoritmo Chain-Map



14

Algoritmo Chain-Map

- La distancia umbral d_u es el parámetro más problemático:
  - Un umbral bajo dará lugar a clases ficticias.
  - Un umbral alto tenderá a agrupar en una misma clase muestras de clases distintas.
- La elección de la primera muestra no resulta demasiado sensible, salvo para ciertas distribuciones de clases.



15

Algoritmo Max-Min

- En cada paso, se calculan las distancias mínimas a las clases ya existentes, de manera que la máxima distancia de entre ellas (que corresponde a la muestra no agrupada que está más alejada de una clase concreta, la cual es a la vez la clase más próxima a ella) se compara con la distancia media entre las clases ya formadas para ver si es posible crear una nueva clase.

16

Algoritmo Max-Min

- Seleccionar una muestra al azar x_i, y crear una primera clase θ_1 con dicha muestra.
- Seleccionar la máxima distancia de las $n - 1$ muestras no agrupadas a x_i, y crear una segunda clase θ_2 con la muestra implicada x_j ($d(x_i, x_j)$ es máxima).
- Para cada una de las $n - 2$ muestras no agrupadas, x:
  - Calcular las distancias a los centroídes de θ_1 y θ_2: $d(x, z_1)$ y $d(x, z_2)$.
  - Tomar la mínima de las distancias entre $d(x, z_1)$ y $d(x, z_2)$, d_{min}.

17

Algoritmo Max-Min

- De las $n - 2$ distancias mínimas, seleccionar la distancia máxima, d_{max}.
- Si d_{max} es mayor que la distancia entre las dos clases formadas (es decir, $d_{max} > d(z_1, z_2) * f\ (0 < f < 1)$), crear una nueva clase θ_3 con la muestra correspondiente a esa distancia máxima.
- Para cada una de las $n - 3$ muestras no agrupadas, x:
  - Calcular las distancias a los centroídes de θ_1, θ_2 y θ_3: $d(x, z_1)$, $d(x, z_2)$ y $d(x, z_3)$.
  - Tomar la mínima de las distancias entre $d(x, z_1)$, $d(x, z_2)$ y $d(x, z_3)$, d_{min}.

18

Algoritmo Max-Min

- De las $n - 3$ distancias mínimas, seleccionar la distancia máxima, d_{max}.
- Si d_{max} es mayor que una fracción de la distancia media entre los prototipos de las clases formadas (es decir, si $d_{max} > 1/3 * [d(z_1, z_2) + d(z_1, z_3) + d(z_2, z_3)] * f$), crear una nueva clase θ_4 con la muestra correspondiente a esa distancia máxima.

19

Algoritmo Max-Min

- Repetir este proceso hasta que no se encuentre ninguna muestra que cumpla la condición d_{max} > distancia media entre clases * f.
- Asignar los elementos que queden por agrupar a la clase más cercana.

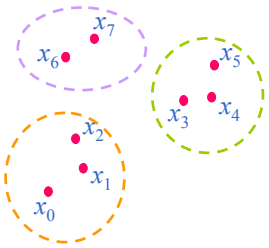
20

Algoritmo Max-Min

- Es muy sensible al parámetro *f*. Por tanto, en general, se prueba con diferentes valores de *f*.
- Combinación heurística de distancias mínimas y máximas.

21

Algoritmo Max-Min



- Muestra seleccionada al azar: x_4.
- La muestra más alejada de x_4 es x_0.
- La máxima de entre las mínimas distancias corresponde a $d(x_7, z_1)$.
- Tomando $f = 0'5$: $d(x_7, z_1) > 0'5 * d(z_1, z_2)$
- Creamos una tercera clase con x_7.
- Repetimos el proceso con las tres clases. Como ya no encontramos ninguna muestra que cumpla condición, finaliza el proceso de crear nuevas clases.
- Asignamos cada una de las muestras no agrupadas a la clase más próxima.

22

Algoritmo de las c-Medias

- Debemos conocer a priori el número de clases del problema, *c*.
- Es simple, pero muy eficiente.
- Obviamente, es muy sensible al parámetro *c*:
  - Un *c* superior al valor real dará lugar a clases ficticias.
  - Un *c* inferior al valor real producirá menos clases de las reales.

23

Algoritmo de las c-Medias

- Para comprobar que el valor *c* haya sido seleccionado correctamente, podemos analizar la dispersión estadística de las clases formadas:
  - Cuando las dispersiones sean diferentes entre sí (que alguna de ellas sea muy elevada respecto a las demás), podemos pensar que hemos tomado un *c* bajo.
  - Si algunas distancias entre clases son demasiado pequeñas respecto a las demás, podemos sospechar que se ha utilizado un *c* alto.

24

Algoritmo de las c-Medias

- Seleccionar c muestras al azar, que constituirán los centroides de las c clases: $z_1(1)$, $z_2(1)$, ..., $z_c(1)$.
- Distribuir las n muestras entre las c clases, según la mínima distancia euclídea a los c centroides actuales.
- Actualizar los centroides de las c clases.
- Comprobar si se ha alcanzado una situación estable, es decir, si se cumple:

$$z_i(p+1) = z_i(p) \qquad \forall\, i = 1, 2, ..., c$$

25

Algoritmo de las c-Medias

- El objetivo en el cálculo de los nuevos centroides es minimizar la función criterio del error cuadrático para cada clase, es decir,

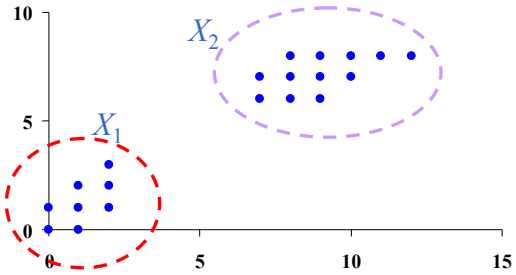
$$J_i = \sum_{x \in X_i(p)} \left\| x - z_i(p) \right\|^2 \quad i = 1, \ldots, c$$

- Esta función se minimiza con la media de $X_i(p)$:

$$z_i(p+1) = \frac{1}{n_i(p)} \sum_{x \in X_i(p)} x \quad i = 1, \ldots, c$$

26

Algoritmo de las c-Medias



- Seleccionamos al azar x_1 y x_2.
- El primer agrupamiento según la mínima distancia será:

  $X_1(1) = \{x_1, x_3\}$

  $X_2(1) = \{x_2, x_4, x_5, \ldots, x_{19}, x_{20}\}$

- Al actualizar los centroides y redistribuir las muestras:

  $X_1(2) = \{x_1, x_2, \ldots, x_7, x_8\}$ $\quad$ $X_2(2) = \{x_9, x_{10}, \ldots, x_{19}, x_{20}\}$

- Volvemos a actualizar los centroides. En este caso, al redistribuir las muestras, se producen los mismos agrupamientos. Por tanto, se detectaría una posición estable.

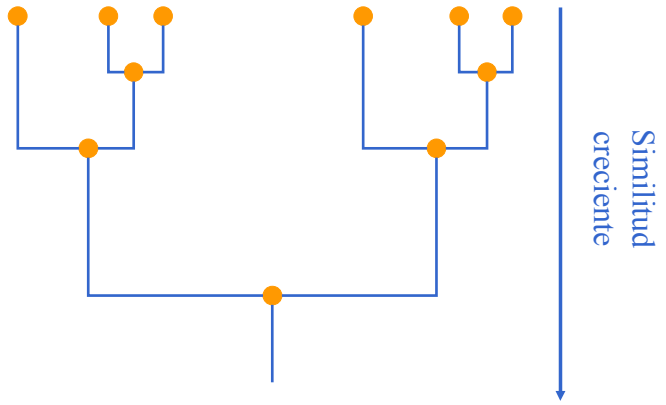
27

Agrupamiento Jerárquico

- Se comienza con una partición en la que cada muestra representa una clase distinta.
- Se van ajustando las clases en función de una cierta medida de similitud (en un orden creciente).
- De este modo, se obtiene un árbol de agrupamientos, llamado *dendograma*.

28

Agrupamiento Jerárquico



29

Agrupamiento Jerárquico

- Sea $c' = n$.
- Si $c' \leq c$, finalizar.
- Encontrar el par de clases que más se parezcan según la medida de similitud adoptada: X_i, X_j.
- Agrupar las clases X_i y X_j, y decrementar c'.
- Volver al Paso 2.

30

Agrupamiento Jerárquico

- Algunas de las funciones de similitud más usadas:

$$d_{min}(X_i, X_j) = \min_{x \in X_i, x' \in X_j} \|x - x'\|$$

$$d_{max}(X_i, X_j) = \max_{x \in X_i, x' \in X_j} \|x - x'\|$$

$$d_{prom}(X_i, X_j) = \frac{1}{n_i n_j} \sum_{x \in X_i} \sum_{x' \in X_j} \|x - x'\|$$

$$d_{med}(X_i, X_j) = \|m_i - m_j\|$$

31