Reconocimiento de patrones

Clase 9: Agrupamiento

Para el día de hoy…

- Agrupamiento
- Medidas de similitud
- Agrupamiento de umbral basado en ordenamiento
- Algoritmo Max-min

La tarea

Una de las tareas centrales en clasificadores basados en distancia mínima es

Determinar los prototipos o centros de las clases

Medida de similitud

- Una medida de similitud $\delta(x, y)$ para dos patrones x e y puede ser definida tal que
$$\lim \delta(x, y) = 0 \; x \to y$$
- Por ejemplo, si los patrones están en $\mathbb{R}^n$ y definimos
$$\delta(x, y) = ||x - y||$$
- Si x es un patrones que se espera sea normalmente distribuido entonces
$$\delta(x, \mu) = ||x - \mu||_C$$
- Donde μ es la media de la población, C su covarianza y
$$||x - \mu||_C = (x - \mu)^T C^{-1} (x - \mu)$$
- Es la distancia de Mahalanobis

El siguiente paso

- Obtener el procedimiento de agrupamiento que
  - Creará los grupos
  - Asignará cada patrón a su grupo
- El algoritmo puede estar basado en heurísticas
- Normalmente incluye la optimización de alguna medida de desempeño. Para patrones en $\mathbb{R}^n$

$$I = \sum_{(i=1)}^{m} \sum_{x \in \mathrm{C}_i} ||x - \mu_i||^2$$

- Donde $C_i, 1 \leq i \leq m$ denota varios grupos y μ denota los centros de los grupos. Normalmente $\mu_i = \frac{1}{N} \sum_{j=1}^{N_i} x_j^{(i)}$ y $x_j^{(i)}$ son los patrones pertenecientes a C_i
- ¿Algún problema en minimizar I?

Una medida de desempeño alternativa

- Tal vez preferiríamos que para dos patrones x e y pertenecezcan a la misma clase si su distancia es menor a un umbral

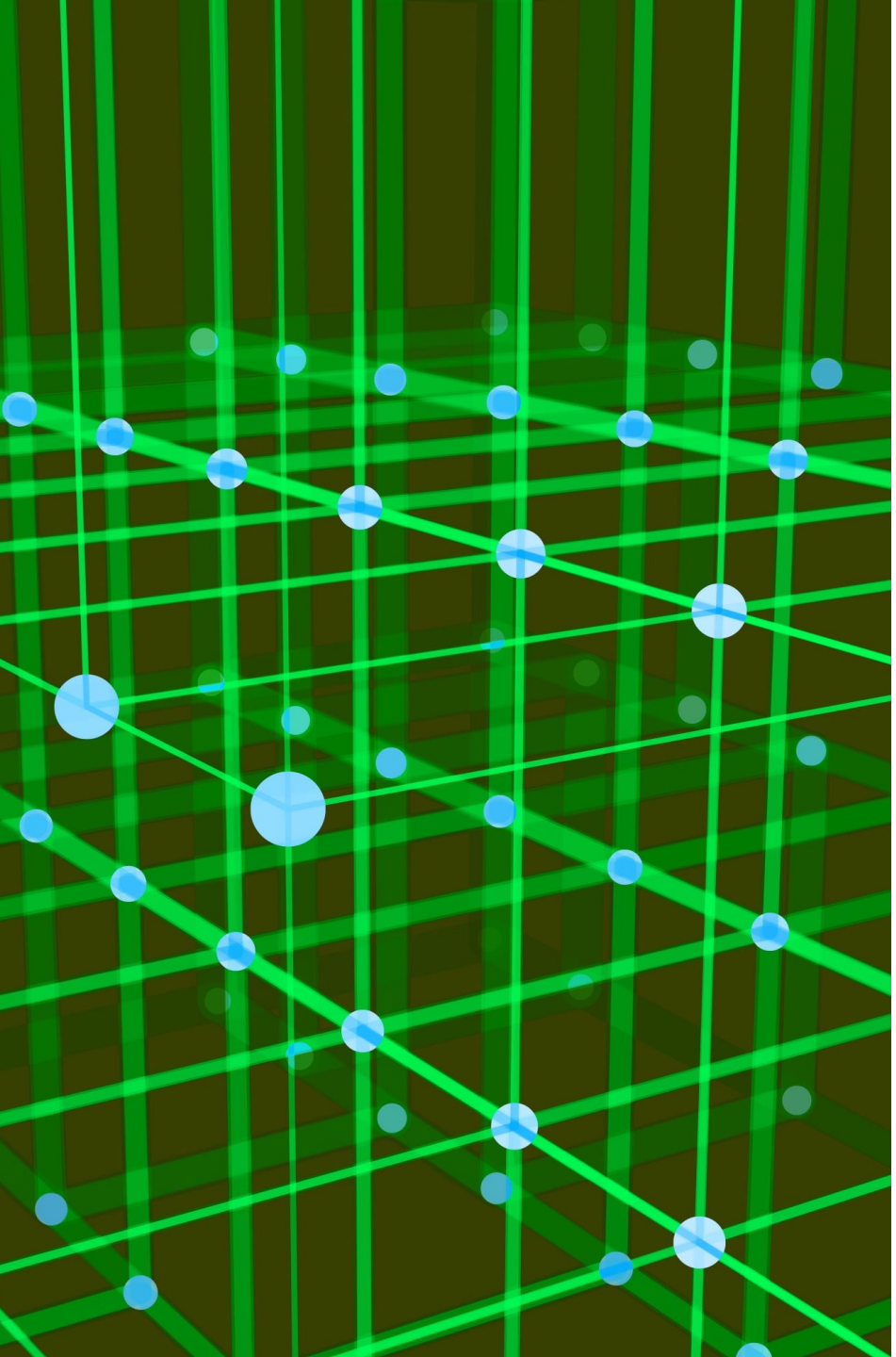
$$I' = \sum_{i=1}^{m} \left(\frac{1}{N_i} \sum_{x,y \in C_i} ||x - y||^2 \right)$$

Algoritmo de agrupamiento de umbral dependiente de ordenamiento

- Consideremos un conjunto de patrones en $\mathbb{R}^n$

$$S = \{x_1, \dots, x_N\}$$

- La medida de similitud es la normal Euclidiana
- Existe un umbral t que dice si dos patrones perteneces al mismo grupo



El algoritmo

- $y_1 = x_1$
- $C_1 = C_1 \cup \{x_1\}$
- $k = 2$
- Desde $i = 2$ hasta N
  - Si $||x_i - y_j|| \geq t, 1 \leq j \leq k$
    - $y_k = x_i, C_k = C_k \cup \{x_i\}, k = k + 1$
  - Si no
    - $l = \arg\min_j ||x_i - y_j||$
    - $C_l = C_l \cup \{x_i\}$

Las preguntas

Ejercicio

- Considere los patrones $(1,1)^T, (2,3)^T, (2,1)^T, (4,3)^T, (3,2)^T, (3,4)^T$
- $t = 1.5$
- ¿A qué clase pertenece cada patrón de acuerdo al algoritmo de agrupamiento de umbral dependiente de ordenamiento?

Algoritmo max-min

Input: n – the problem's dimension.

m – the number of samples.

$X = \{x_i\}$, $1 \le i \le m$ – the given samples in R^n.

t – a threshold value which determines whether a new cluster should be created.

Output: k – the number of cluster centers found.

$\{y_j\}$, $1 \le j \le k$ – the cluster centers.

$\{m_j\}$, $1 \le j \le k$ – the cluster sizes.

$\{l_{ij}\}$, $1 \le i \le m_j$ – the indices of the original samples which belong to the j – th cluster, $1 \le j \le k$.

Step 1. Set $y_1 = x_1$, $y_2 = x_{j_0}$, $l_{11} = 1$, $l_{12} = j_0$ where

$$\|x_{j_0} - y_1\| = \max_{2 \le i \le m} \|x_i - y_1\|$$

Set $k = 2$, $a = \overline{\|y_i - y_j\|}$ (arithmetic mean), where $1 \le i, j \le k$, $i \ne j$ and $X' = X - \{y_1, y_2\}$.

Step 2. Find j_0, $1 \le j_0 \le k$ and $x_{i_0} \in X'$ such that

$$d = \|x_{i_0} - y_{j_0}\| = \max_{x_i \in X'} \min_{1 \le j \le k} \|x_i - y_j\|$$

If $d < ta$ (no more clusters) go to Step 4; otherwise go to Step 3.

Step 3. Set $k \leftarrow k+1$, $y_{k+1} = x_{i_0}$, $l_{k1} = i_0$, $X' \leftarrow X' - \{y_{k+1}\}$ and go to Step 2.

Step 4. Set $m_j = 1$, $1 \le j \le k$.

Step 5. For each $x_i \in X'$ find j: $1 \le j \le k$ for which

$$\|x_i - y_j\| = \min_{1 \le j \le k} \|x_i - y_j\|$$

and set $m_j \leftarrow m_j + 1$ and $l_{m_j j} = i$.

Step 6. For $1 \le j \le k$ replace y_j by $(x_{l_{1j}} + x_{l_{2j}} + \ldots + x_{l_{m_j j}})/m_j$.

Step 7. For $1 \le j \le k$ output $y_j, m_j, \{l_{ij}\}_{i=1}^{m_j}$ and stop.

A subroutine MMD which incorporates algorithm 3.3.1 is given in the appendix.

Método de distancia Max-Min

- Se supone que al menos existen dos grupos
- Determinar todos los centros de los grupos
- Mantener fijo un umbral t que determina si se debe crear un nuevo grupo
  - Sea $y_1, \dots, y_k$ los centros de los grupos existentes; a la media aritmética entre los centros; b el centro del patrón más probable a ser elegido como el centro de un nuevo grupo
  - Si $s = min||b - y_i||, 1 \leq i \leq k$ es menor que ta
    - Terminar fase
  - Si no
    - $y_{k+1} = b$
- Cada muestra se asigna su grupo más cercano
- Los centros de ajustan para que sean la media aritmética de las observaciones de cada grupo

Las preguntas

Para la otra vez…

- K-means