Reconocimiento de patrones

Clase 10: K means

Para el día de hoy…

- K means

Antes de empezar

- Entregar en un documento pdf de una página los siguientes elementos para definir el proyecto final del curso:
  - Titulo
  - Miembros del equipo (máx. 3)
  - Descripción del problema ¿Qué no está sucediendo?
- ¿Quién está interesado?
- ¿Por qué es de interés para ustedes?
- ¿Qué previene que se haga?
- ¿Qué pasa si no se hace?
- ¿Qué se planea hacer? (en términos de lo existente y las operaciones básicas para transformar)
- ¿Cómo luce el éxito?

K means

- ¿Cuál es la idea?
- ¿Cuáles son las suposiciones?
- ¿Cuáles son las entradas?
- ¿Cuál es la salida?
- ¿Dadas las mismas entradas, la salida es siempre la misma?

¿Cuál es el algoritmo?

Un poco de explicación

- Dado un conjunto de datos $X = \{x_1, x_2, \dots, x_m\}$, suponemos la existencia de k grupos
- Los centros son aproximados por $y_1^{(0)}, \dots, y_k^{(0)}$
- Los grupos se encuentran de forma iterativa
- En cada paso todos los patrones se clasifican y cada centro se ajusta usando la media aritmética de las muestras del grupo hasta que entre dos iteraciones los grupos no cambien

El algoritmo

Output: $\{\boldsymbol{y}_j\}$, $1 \le j \le c$ – the final cluster centers.

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

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

it – the number of iterations needed for convergence.

Step 1. Initialization: set $\boldsymbol{y}_{j0} = \boldsymbol{x}_j$, $1 \le j \le c$ and it $= 0$.

Step 2. Classify $\{\boldsymbol{x}_i\}_{i=1}^{m}$ about the cluster centers $\{\boldsymbol{y}_{j0}\}_{j=1}^{c}$ using the minimum distance classifier. For $1 \le j \le c$ denote by $\{\boldsymbol{x}_{l_{ij}}\}_{i=1}^{m_j}$ the samples which cluster around $\boldsymbol{y}_{j0}$.

Step 3. For $1 \le j \le c$ obtain $\boldsymbol{y}_j$ which minimizes the performance index

$$I_j(z) = \sum_{i=1}^{m_j} \left\| z - \boldsymbol{x}_{l_{ij}} \right\|^2, \; z \in R^n \qquad (3.3.12)$$

Basic calculus implies

$$\boldsymbol{y}_j = \left(\sum_{i=1}^{m_j} \boldsymbol{x}_{l_{ij}} \right) / m_j \qquad (3.3.13)$$

i.e. $\boldsymbol{y}_j$ is the arithmetic mean of $\{\boldsymbol{x}_{l_{ij}}\}_{i=1}^{m_j}$. Set $it \leftarrow it + 1$.

Step 4. If

$$\boldsymbol{y}_j = \boldsymbol{y}_{j0}, \quad 1 \le j \le c \qquad (3.3.14)$$

output $\boldsymbol{y}_j, m_j, \{\boldsymbol{x}_{l_{ij}}\}_{i=1}^{m_j}$, $1 \le j \le c$; it and stop. Otherwise, if $it > N$ output 'number of iterations exceeded'; else set $\boldsymbol{y}_{j0} = \boldsymbol{y}_j$ and go to Step 2.

Las preguntas

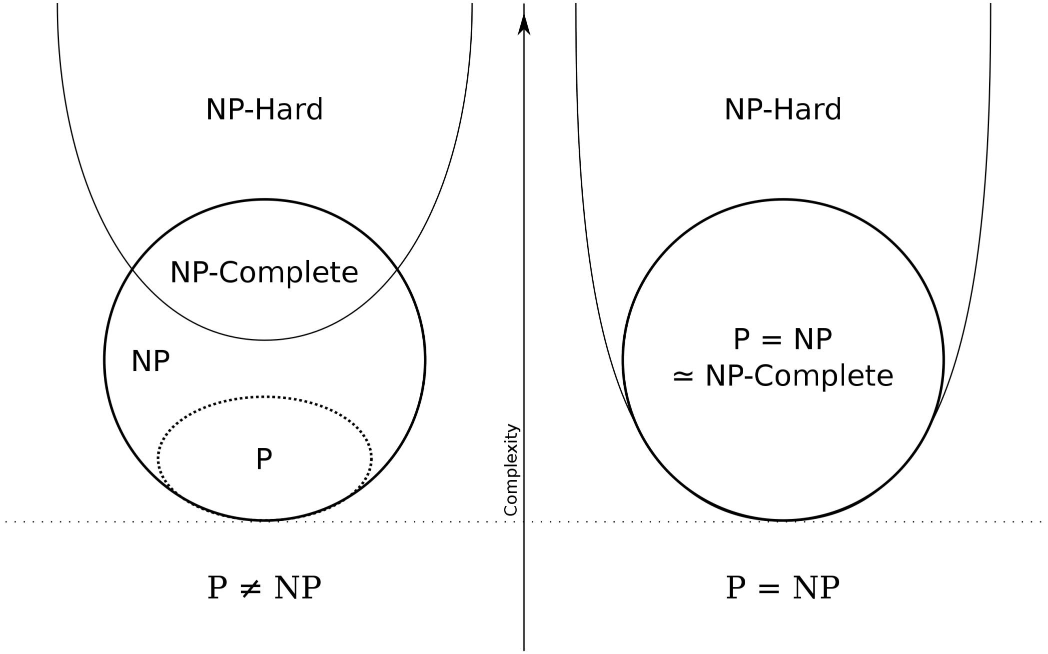
- ¿Es correcto?
- ¿Complejidad en tiempo?
- ¿Complejidad en espacio?
- ¿Ventajas y deventajas?

Un detour a complejidad computacional

- La elección óptima para k y y_{j0} así como las condiciones de convergencia del algoritmo no son conocidas
- La función objetivo de k-means es

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

- Este problema es NP-Hard



La clases de problemas

La clase P: contiene aquellos problemas que pueden resolverse en tiempo polinomial con una MT determinista.

- Buscar un número en un arreglo ordenado o desordenado
- Ordenar un arreglo
- Ruta más corta en un grafo
- Suma, multiplicación de polinomios

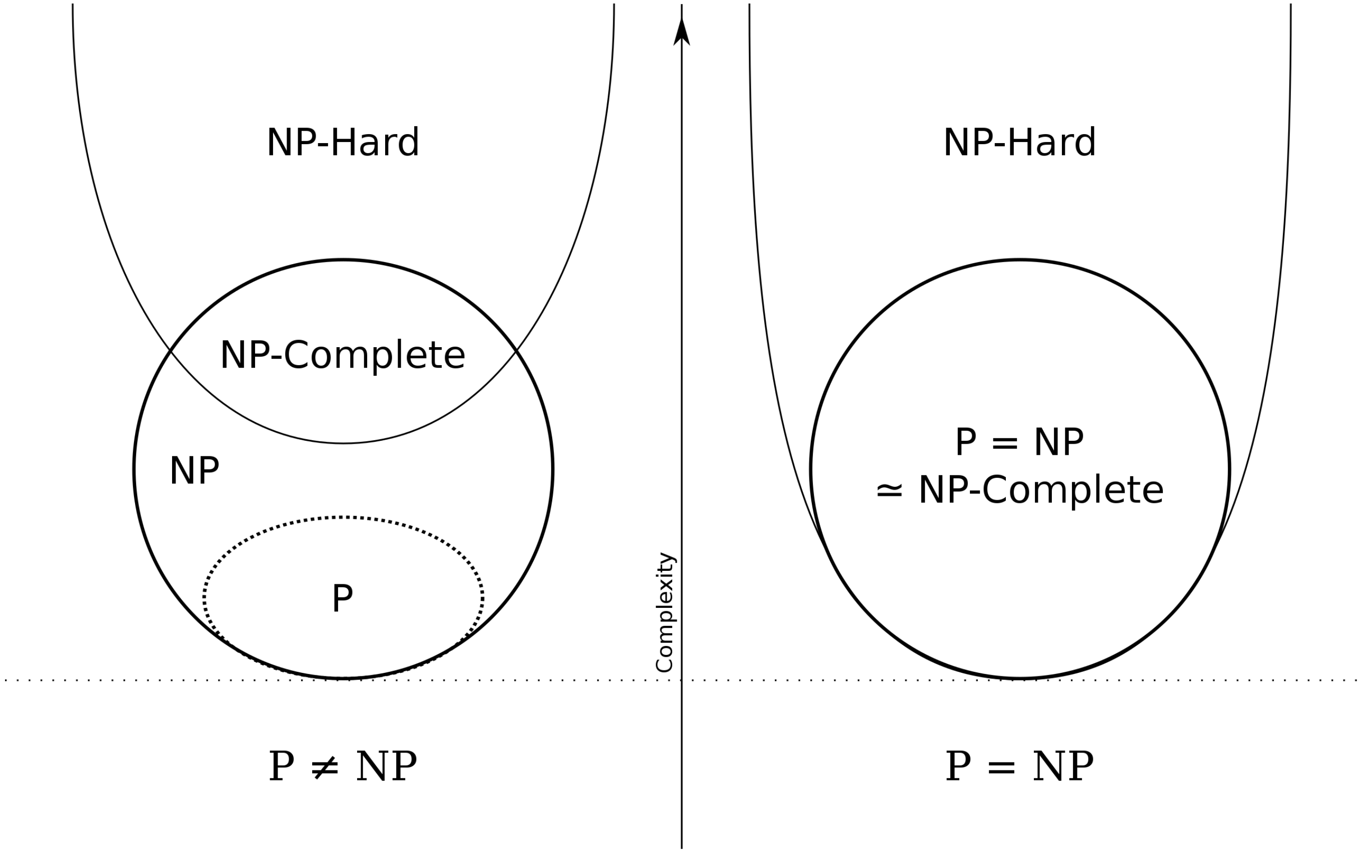
La clase NP: contiene aquellos problemas que pueden resolverse en tiempo polinomial con una MT no determinista

- Satisfacción de fórmulas boolean
- Problema de la mochila
- Problema del viajero

Problemas NP completos

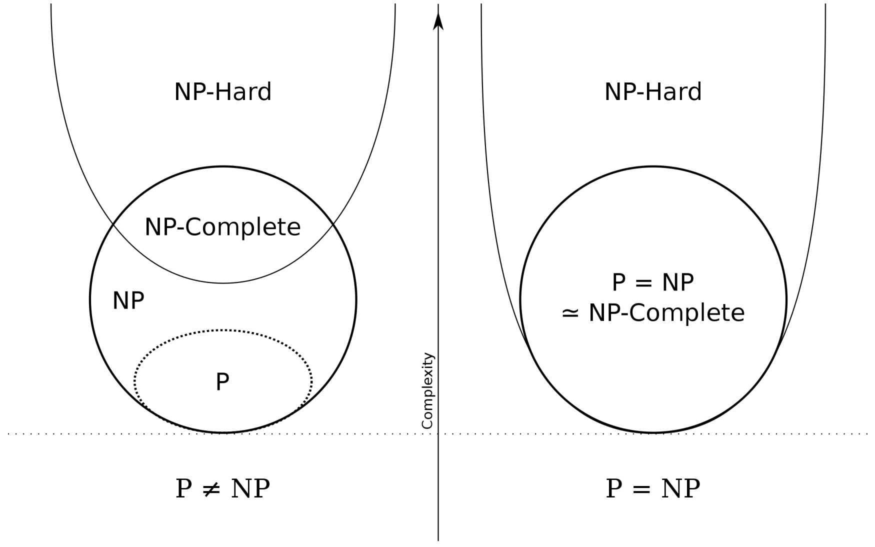
- Se dice que un problema es NP completo si las siguientes afirmaciones sobre un lenguaje L son verdaderas:

1. Pertenece a NP
2. Para todo lenguaje L′ perteneciente a NP existe una reducción en tiempo polinomial de L′ a L



Problemas NP-difíciles

- Algunos problemas L son tan difíciles que aunque podamos demostrar la condición 2 de la definición, no se puede demostrar la condición 1.
- En esos casos decimos que L es NP-difícil.
- Una demostración de que L es NP-difícil basta para demostrar que L es muy probable que requiera un tiempo exponencial o aún peor.



Para la otra vez…

- ISO Data