1.pdf

destherhada

Inteligencia Artificial II

3º Grado en Ingeniería Informática

Facultad de Informática
Universidad Complutense de Madrid

Aprendizaje supervisado: Hay que dar al sistema ejemplos ya clasificados. El objetivo del sistema es descubrir reglas de clasificación.

A la hora de *representar los individuos* tendremos un número de individuos n descrito por un conjunto de variables m. Los datos se representan en forma de matriz n x m donde las filas son los individuos y las columnas las variables. Las variables m son las dimensiones en las que representamos los individuos. Suele ser de ayuda poder inspeccionar los datos visualmente, para ello se utiliza el diagrama de dispersión (scatter plot). Los algoritmos de aprendizaje automático son sensibles a la forma en que los individuos están representados, por eso es importante representar los individuos y sus variables adecuadamente.

Aprendizaje no supervisado: No se le da ninguna información al sistema, tiene que descubrir patrones en el conjunto de entrenamiento que permita agrupar unos ejemplos de otros.

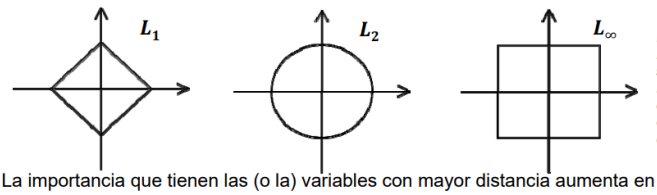
El objetivo del *aprendizaje no supervisado* es encontrar estructura en los datos proporcionados sin atender a ninguna categoría prefijada. Se busca estructura en los individuos (agrupamiento/ clustering) o en las variables (reducción de la dimensionalidad). La estructura nos permite ganar comprensión sobre los datos.

Las *técnicas de reducción de la dimensionalidad* se aplican cuando contamos con muchas variables en un problema. Se parte de un conjunto de variables m que se busca reducirlo a un conjunto p mucho menor de factores que conservan la máxima información inicial. Se obtienen como una combinación de las variables originales.

El objetivo de las *técnicas de agrupamiento o clustering* es agrupar los n individuos de nuestro conjunto de datos en una serie de grupos de forma que los individuos del mismo grupo sean lo más parecidos entre sí y de forma que los individuos de grupos diferentes sean lo más diferentes entre sí. De esta forma los grupos revelan cierta estructura de los individuos de nuestro conjunto de datos. Existen dos grandes grupos de algoritmos de agrupamiento o clustering:

Algoritmos de clustering jerárquico:
Cada individuo empieza siendo un cluster -> hay n clusters. Se repite el algoritmo hasta que todos los individuos formen un único cluster, agrupando los clusters más próximos en un único cluster. Necesitamos definir la distancia que se usa para medir la proximidad.

Manhattan $L_1(A,B) = \sum_{i=1}^{m} |x_{Ai} - x_{Bi}|$ Euclídea $L_2(A,B) = (\sum_{i=1}^{m} (x_{Ai} - x_{Bi})^2)^{1/2}$ Chebychev $L_\infty(A,B) = max_i |x_{Ai} - x_{Bi}|$



En estas imágenes se muestran los puntos que se encuentran a la misma distancia en un espacio bidimensional con cada una de estas distancias

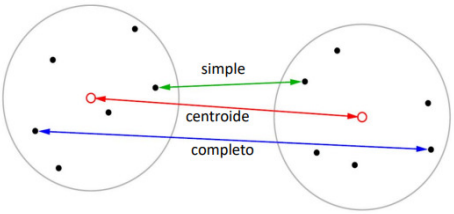
- La importancia que tienen las (o la) variables con mayor distancia aumenta en $L_1 < L_2 < L_\infty$
  - En L_1 damos igual valor a todas las diferencias, en L_2 el cuadrado hace que pesen más las diferencias grandes y en L_∞ solamente se tiene en cuenta la variable donde la diferencia es mayor

Centroide: Se toma la distancia entre los puntos medios (el vector medio) de los dos *clusters*

Enlace simple (*single linkage*): Se toma la distancia entre los puntos más próximos de los dos *clusters*

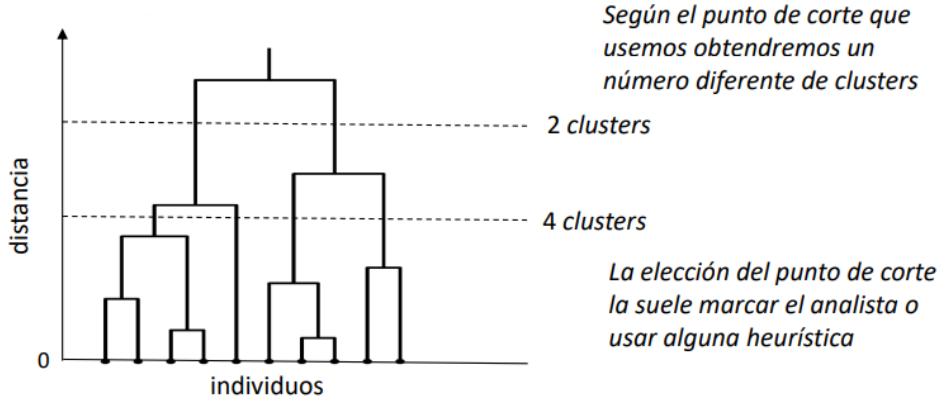
Enlace completo (*complete linkage*): Se toma la distancia entre los puntos más alejados de los dos *clusters*

Ejemplo con dos clusters en un espacio bidimensional



1. Crear la matriz de distancias D.
2. Agrupar individuos.
    a. Partición inicial -> cada individuo es un cluster.
    b. Calcular la siguiente partición usando la matriz D.
        i. Elegir los dos clusters más cercanos y agrupar los dos en un único cluster. Actualizar la matriz D.
    c. Repetir paso b hasta tener un único cluster con todos los individuos. Representar el dendograma.

El dendrograma es una representación bidimensional de la jerarquía inferida por el algoritmo de clustering jerárquico. En un eje ponemos los individuos y los clusters, el otro eje representa la distancia.



Según el punto de corte que usemos obtendremos un número diferente de clusters

La elección del punto de corte la suele marcar el analista o usar alguna heurística

Algoritmos de clustering basados en particiones:
Tienen como objetivo dividir los n individuos en un número de clusters k. Divide el espacio de representación m dimensional en k regiones. El algoritmo de k-medias es el más común y funciona de la siguiente manera:

1. Inicializar centroides a puntos aleatorios del espacio.
2. Asignar cada individuo al centroide más cercano
3. Actualizar la posición de los centroides al valor medio de las posiciones de los individuos asignados.

Se cicla hasta que la posición de los centroides no cambia.

A la hora de decidir el número de clusters k que vamos a utilizar podemos ir probando diferentes valores pero existen funciones para valorar la dispersión de los centroides.

$$\text{Índice Dunn} = \frac{\min(\textit{distancia interclúster})}{\max(\textit{distancia intraclúster})}$$

Cuanto mayor sea el número de clusters k, mejores serán los valores de Dunn y Davies-Bouldin. Para saber exactamente cuál es el número ideal se utiliza el diagrama del codo.