Ciencia de Datos en inteligencia de Negocios
Lectura: Clustering.

Clustering.

Clustering es el nombre que se le da un algoritmo de agrupamiento de datos en función de un criterio establecido. Los criterios más usados son medidas de similitud o distancia. Un algoritmo de clustering se utiliza para encontrar patrones en conjuntos de datos donde no se tienen establecidos relacionen entre datos.

Los algoritmos de clustering son métodos de clasificación no supervisada, es decir, no se tiene a priori la definición de posibles grupos dentro de los datos. Estos algoritmos se consideran como exploratorios para determinar posibles relaciones desconocidas.

Hay que considerar dos puntos importantes que normalmente tienden a confundirse:

1.- El índice de similitud indica una forma de medir que tan parecido es un dato, un objeto, variable con otra. Una información similar se obtiene al determinar la distancia entre los dos entes que se están comparando.

2.- El algoritmo de clustering se basa en una heurística que hace un agrupamiento entre datos que son parecidos según el índice de similitud o distancia que se esté utilizando.

En estos dos puntos se debe de considerar también que las relaciones en un conjunto de datos pueden cambiar en función del índice de similitud y el algoritmo de agrupamiento elegido.

Hierarchical clustering (Agrupamiento Jerárquico)

En minería de datos, el agrupamiento jerárquico es un método de análisis de grupos puntuales, el cual busca construir una jerarquía de grupos. Este método es aplicado exclusivamente a datos numéricos. Las estrategias para agrupamiento jerárquico pueden clasificarse en dos tipos:

- **Aglomerativas:** Este es un acercamiento ascendente: cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía.
- **Divisivas:** Este es un acercamiento descendente: todas las observaciones comienzan en un grupo, y se realizan divisiones mientras uno baja en la jerarquía.

El agrupamiento o división de los datos se basa en heurísticas establecidas.

Como ya se sabe, los índices de similitud son medidas de parecido entre datos puntuales. Pero, en orden de decidir qué grupos deberían ser combinados (para aglomerativo), o cuando un grupo debería ser dividido (para divisivo), una medida de similitud entre conjuntos es requerida. En la mayoría de los métodos de agrupamiento jerárquico, esto es logrado mediante uso de una métrica apropiada (una medida de distancia entre pares de observaciones), y un criterio de enlace el cual especifica la similitud de conjuntos como una función de las distancias dos a dos entre observaciones en los conjuntos.

Ciencia de Datos en inteligencia de Negocios
Lectura: Clustering.

Los intentos de establecer una medida de similitud entre conjuntos se han propuesto criterios de enlace. El criterio de enlace determina la distancia entre conjuntos de observaciones como una función de las distancias entre observaciones dos a dos. Algunos criterios de enlace entre dos conjuntos de observaciones *U* y *V* usados son:

Nombre	Ecuación	
Enlace simple o agrupamiento por mínimo (The Nearest Point Algorithm)	$d(u,v) = \min(dist(u(i), v(j)))$	
Enlace completo o agrupamiento por máximo (The Farthest Point Algorithm o Voor Hees Algorithm)	$d(u,v) = \max(dist(u(i), v(j)))$	
Enlace por media o promedio (UPGMA)	$d(u,v) = \sum_i \sum_j \frac{d(u(i), v(j))}{\lvert u\rvert * \lvert v\rvert}$	
Enlace ponderado o weighted (WPGMA)	$d(u,v) = \frac{d(s,v) + d(t,v)}{2}$	
Enlace por centroide (UPGMC)	$d(u,v) = \lVert c_u, c_v\rVert_2$	c_u, c_v son los centroides de los conjuntos que se están aglomerando

El agrupamiento jerárquico tiene la ventaja de que cualquier medida de distancia puede ser usada. Los datos originales no son utilizados para el aglomeramiento, solo se usa una matriz de distancia.