Clustering

¿Qué se entiende por clustering?

* Procedimiento de agrupación de una serie de vectores de acuerdo con un criterio.
  + Sus criterios son por lo general distancia o similitud.
* ¿Para qué sirve?
  + Biología para clasificar animales y plantas
  + Medicina para identificar enfermedades
  + Marketing para identificar personas con hábitos de compras similares.
  + Teoría de la señal sirven para eliminar ruido
  + Biometría para identificar al locutor o de caras
* ¿Qué tipos de clustering existen?
  + Clustering Jerárquico
  + Clustering Particional

Las técnicas de clustering se aplican cuando no existe una clase para predecir, pero las instancias naturalmente se dividen en grupos.

* Estos grupos reflejan algún mecanismo que actúa en el dominio de las instancias, un mecanismo que hace que algunos casos sean más parecidos entre sí que los restantes.
* Clustering requiere naturalmente de diferentes técnicas a los métodos de aprendizaje de clasificación que hemos estudiado hasta ahora.

Hay diferentes formas de expresar el resultado de la agrupación

* Exclusivos (instancia pertenece a un solo grupo)
* Superpuestos (instancia puede pertenecer a varios grupos)
* Probabilísticos (instancia pertenece a cada grupo con una cierta probabilidad)
* Jerárquicos (Una división aproximada de casos en grupos en el nivel superior, siendo reﬁnados en los niveles inferiores hasta llegar a las instancias individuales)

En realidad, la elección entre estas posibilidades debe ser dictada por la naturaleza de los mecanismos que se cree que subyacen en el fenómeno de agrupamiento en particular.

Sin embargo, debido a que estos mecanismos son raramente conocidos la existencia misma de los grupos es, después de todo, algo que estamos tratando de descubrir- y por razones pragmáticas también, la elección suele estar dictada por las herramientas de agrupación que están disponibles.

K-Means

* Técnica de agrupación clásica.

1. Se especifica de antemano la cantidad de grupos que se desean obtener: Parámetro K
2. K puntos son elegidos al azar como centros de grupo
3. Todas las distancias son asignadas a su centro más cercano de acuerdo con la distancia euclidiana
4. Se calcula el centro de gravedad, o media (means) de las instancias de cada grupo.
5. Estos centroides se toman como nuevos valores de centro en sus respectivos grupos
6. Todo el proceso se repite con los neuvos centros de los grupos
7. La iteración continua hasta que los mismos puntos se asignen a cada grupo de bucles consecutivos.

* Es un método simple y eficaz
  + Fácilmente se demuestra que el centro del cluster **minimiza la distancia al cuadrado total** de cada punto del cluster a su centro
  + Una vez que la iteración se ha estabilizado, a cada punto se le asigna su **centro más cercano,** por lo que el efecto general es minimizar la distancia al cuadrado total de todos los puntos a sus centros
* Sin embargo, **la distancia mínima hallada es local,** no hay ninguna garantía que sea global.
  + Los grupos finales son bastante sensibles a los centros iniciales
  + Arreglos de datos completamente diferentes pueden originarse debido a pequeños cambios en la elección aleatoria inicial.
* Para aumentar la probabilidad de encontrar una solución global, a menudo se ejecuta el algoritmo varias veces con diferentes opciones iniciales y se elije el mejor resultado final, el que tiene la distancia al cuadrado total más corta.
* Es fácil imaginar situaciones en las que k-means no encuentra un buen agrupamiento.
  + Considere por ejemplo cuatro instancias dispuestas en los vértices de un rectángulo en el espacio de dos dimensiones.
  + Hay dos grupos naturales, formados mediante la agrupación de los dos vértices en cada extremo de un lado corto.
  + Pero supongamos que los dos centros iniciales pasan a caer en los puntos medios de los lados largos.
  + Esto forma una conﬁguración estable. Los dos grupos contienen cada uno las dos instancias en cada extremo de un lado largo, no importa cuán grande es la diferencia entre el largo y los lados cortos.
* k-means puede mejorar drásticamente mediante una cuidadosa selección de los centros iniciales, a menudo llamados semillas (seed)
* En lugar de comenzar con un conjunto arbitrario de semillas, que aquí hay un mejor procedimiento
  + Elija la semilla inicial al azar de todo el espacio, con una distribución de probabilidad uniforme.
  + A continuación, elija la segunda semilla con una probabilidad que es proporcional al cuadrado de la distancia de la primera.
  + Proceder, en cada etapa de la elección siguiente con una semilla con una probabilidad proporcional al cuadrado de la distancia desde la semilla más cercana que haya sido elegida.