**Definir brevemente:**

* **Clasificación**
  + Agrupación de patrones/observaciones/ítems en grupos homogéneos.
* **Agrupamiento no exclusivo**
  + Tipo de clasificación en el que un punto puede pertenecer a más de un grupo.
* **Método de enlace simple**
  + En un modelo de interpretación de los datos como grafos, donde cada vértice del grafo y la semejanza se representa con un enlace. El método de enlace simple es aquel en el que se busca las partes conexas del grafo y finaliza cuando todos los puntos están conectados.
* **Agrupamiento particional**
  + Tipo de agrupamiento en la que el espacio de ítems se divide en conjuntos disjuntos de puntos.
* **Centroide**
  + Es el punto que representa un grupo, suele ser calculado como la media de todos los puntos del grupo aunque existen algoritmos que tienen otra definición de centroide de grupo.

**Describir los enfoques y medidas existentes para la validación de los resultados de un agrupamiento.**

Existen dos enfoques principales para la validación de los resultados, el enfoque supervisado y el enfoque no supervisado. El enfoque no supervisado recoge aquellas medidas en las que se pone enfásis en la semejanza de los puntos dentro del conjunto y la distancia con los puntos de los demás grupos, esto es debido a que no se conoce la estructura de los grupos, en cambio el enfoque supervisado interpreta la validación como un problema de clasificación, dado que se conoce a que grupo pertenece cada uno.

**Diferencias entre clústering jerárquico y el particional. ¿Cuándo se usa uno u otro?**

El clústering particional separa todo el espacio de las observaciones obtenidas en espacios disjuntos.

En cambio el clustering jerárquico permite que un grupo en un nivel se subdivida en varios grupos en el nivel inferior.

**Define brevemente las técnicas de validación de clústering no supervisadas.**

Las técnicas de validación no supervisadas intentan decir la bondad de una agrupación a partir de la semejanza de los puntos dentro de un grupo y de su distancia con los puntos de los demás grupos. Pertenecen a esta categoría la suma de errores cuadráticos.

**Describe brevemente las ideas básicas sobre clustering difuso. ¿Cuándo es conveniente aplicarlo?**

En el clustering difuso, cada observación tiene un grado de pertencia a cada grupo. Cada grado de pertenencia toma valor dentro del intervalo [0,1]. Puesto que hay un grado de pertenencia por cada valor, se puede considerar que por cada observación hay un vector de pertenecías donde se recogen dichos grados de pertenencia.

Es conveniente utilizarlo en aquellos casos en los que exista, solapamiento de conjuntos.

**Indique justificando la respuesta, qué tipo de técnicas de detección de outliers utilizaría en cada uno de los siguientes casos:**

**En un experimento se suministra un producto alimenticio a 40 animales y se observa el incremento de peso después de cuatro semanas. Antes de proceder a un análisis de los resultados se quiere detectar si existen valores anormalente bajos o altos.**

Dado que el número de observaciones es bajo, se puede determinar con un gráfico aquellos valores que cumplan la condición de ser anormalmente bajos o altos.

**Se dispone de un conjunto de 30 coches de los que se toman los datos siguientes: aceleración de 0 a 100, peso, cilindrada y número de litros consumidos en 1 hora a 100 km/h. Se quiere analizar si hay combinaciones de estas variables anormales.**

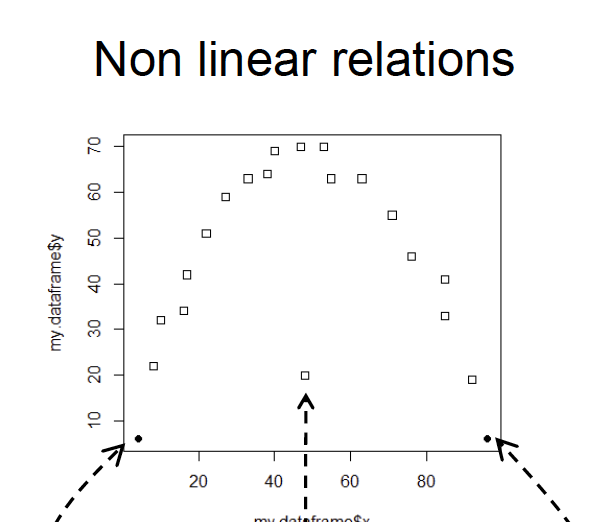
En este caso, al contrario que en el caso anterior no podemos recurrir a un gráfico puesto que el número de variables a considerar es mayor que el representable.

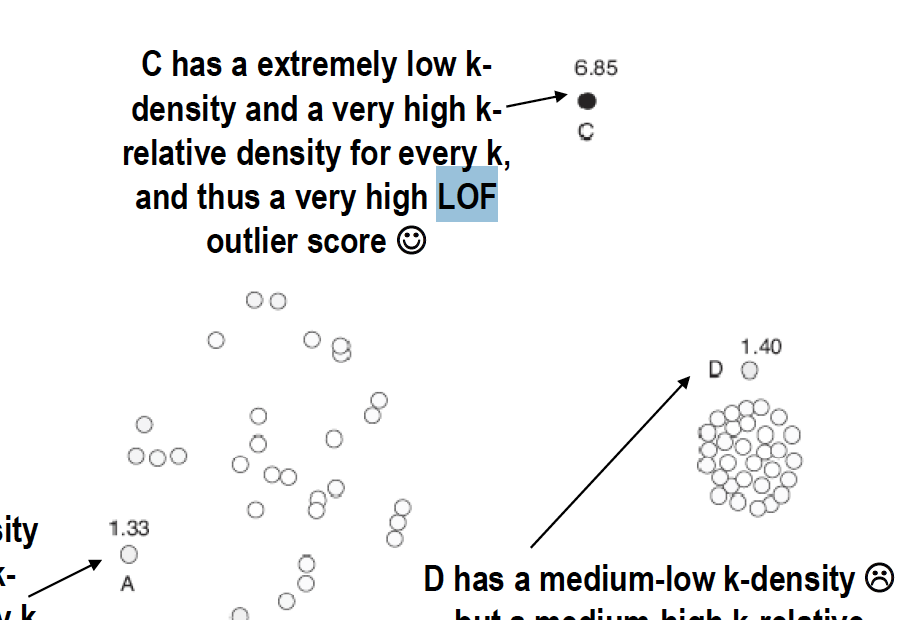
Dado que tampoco tenemos etiquetas para saber si algún caso es anómalo descartamos todos aquellos métodos supervisados y semisupervisados de detección de anomalías, para concentrarnos en los no supervisados. En esta categoría tenemos aquellos métodos basados en el algoritmo de los K vecinos más cercanos, que podrían resolver bien el problema puesto que todas las variables son numéricas.

**¿Cuántos outlier detecta el test estadístico de Grubb en el caso de una distribución normal univariante?**

Detecta únicamente una anomalía o outlier.

**Dibuje una nube de puntos en la que la distancia Mahalanobis no funcione adecuadamente y no detecte un outlier.**

****

**Dibuje una nube de puntos en la que el método de detección de outliers LOF (recuerde que era un método basado en la distancia y que tenía en cuenta la densidad de puntos) sí funcione adecuadamente y detecte un outlier cuando un método básico basado en distancia sin tener en cuenta la densidad no lo haría. **

El punto D es el que corresponde.

**Definir que son reglas de asociación y describir las diferencias entre las reglas de asociación clásicas, las reglas de asociación cuantitativas y las reglas de asociación multinivel.**

Las reglas de asociación relacionan una serie de elementos denominados antecedentes con otra serie de elementos o consecuentes. La aparición del antecedente en una transacción implica que suceda el consecuente.

Las reglas de asociación clásicas están definidas para conjuntos de datos categóricos en forma de lista pero no representa aquellos casos en los que las variable son numéricas o existe una jerarquía de conceptos. Para estos casos se desarrollaron las reglas de asociación cuantitativas y multinivel, respectivamente.

Las reglas de asociación cuantitativas, creadas para cubrir aquellos casos en las que las variables toman valores numéricos, dividen el dominio de la variable para representar el conocimiento, esta división se puede hacer “a priori” con el conocimiento de un experto o se puede programar el algoritmo para que lo aprenda por sí mismo.

Las reglas de asociación multinivel, añaden un concepto por cada nivel de la jerarquía de conceptos a la regla. De esa forma para niveles superiores de la jerarquía pueden obtener mejores valores de soporte y confianza y pueden ser más interpretables.

**Comentar brevemente 3 extensiones de reglas de asociación.**

**Describe el problema de la búsqueda de Reglas de asociación y las medidas clásicas de calidad de las mismas.**

Las reglas de asociación tienen por medidas clásicas el soporte y la confianza, definidas numéricamente como la probabilidad de encontrar el itemset en la base de datos en la transacción y la probabilidad de que se dé el antecedente y el consecuente en la base de datos sobre las apariciones del antecedente.

Por norma general se busca que ambos parámetros tenga los valores más cercanos a 1 posible pero estos casos pueden no ser los ideales, por ejemplo en el caso en el que antecedente se dé en prácticamente toda la base de datos, da igual en que zona de la base de datos se dé el consecuente que siempre dará un valor alto de confianza en la regla.

Por este problema se han desarrollado distintas medidas como el lift, Yules’Q o ratio de confianza entre otras

**¿Qué problemas presentan las medidas clásicas de claridad?¿Qué alternativas conoces, y cómo permiten paliar los problemas mencionados? ---**

**¿Qué son las reglas de asociación cuantitativas y cuándo es necesario considerarlas?¿Qué problemas adicionales plantea su búsqueda?** El problema adicional es el aprendizaje del intervalo.

**¿Qué son y qué ventajas y desventajas tienen los itemsets cerrados y los itemsets maximales?**

Los itemsets maximales son aquellos en los que cualquier superconjunto tienen menor medida de soporte, es por ello que son útiles para reducir el número de itemsets a utilizar en fases posteriores. Lo que no nos permite recuperar fácilmente este tipo itemset es el soporte de cada uno de sus subconjuntos.

Por otro lado los itemset cerrados son aquellos en los que cualquier superconjunto tiene un soporte distinto al actual, lo que nos permite tener siempre a nuestra disposición la medida del soporte pero a cambio de tener mayor número de itemsets.