DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Sistemas Inteligentes de Gestión

Guión de Prácticas de Minería de Datos

Práctica 3

Métodos de agrupamiento [Clustering]

© *Juan Carlos Cubero & Fernando Berzal*

FICHEROS DE DATOS

```
Datos de empleados.sav
iris.csv
```

ENTREGA DE LA PRÁCTICA

```
Clustering.doc
DatosApplet
      3ClustersSinOutliers{.data|.sav}
      3ClustersConOutliers(.data|.sav}
      ClusteringDifícil.data
      Clustering_DatosApplet.knime.zip
      Clustering_DatosApplet.spo
      centroides_*.pmml
Iris
      Clustering_Iris.knime.zip
DatosEmpleados
      DatosEmpleados.muestra.csv
      Clustering_DatosEmpleados.knime.zip
```

K-Means

Para la realización de esta práctica, comenzaremos ejecutando el siguiente aplet, que nos permitirá comprobar el funcionamiento del algoritmo de las k medias (k-means):

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

Nuestro conjunto de datos, en este caso, estará formado por tuplas con dos variables de tipo numérico, representables por tanto en el plano. Cada una de las tuplas de nuestro conjunto de datos corresponderá a un punto.

Seleccionando la opción para mostrar el historial [*Show History*], inicialice [*Initialize*], pulse *Start* y ejecute paso a paso el algoritmo [Step] para ver cómo van cambiando los centroides y la asignación de los puntos a sus centroides más cercanos.

A continuación, cree con el WordPad un fichero llamado `Clustering.doc` en el que ir guardando los resultados obtenidos durante la realización de los ejercicios.

Ejercicios tipo C

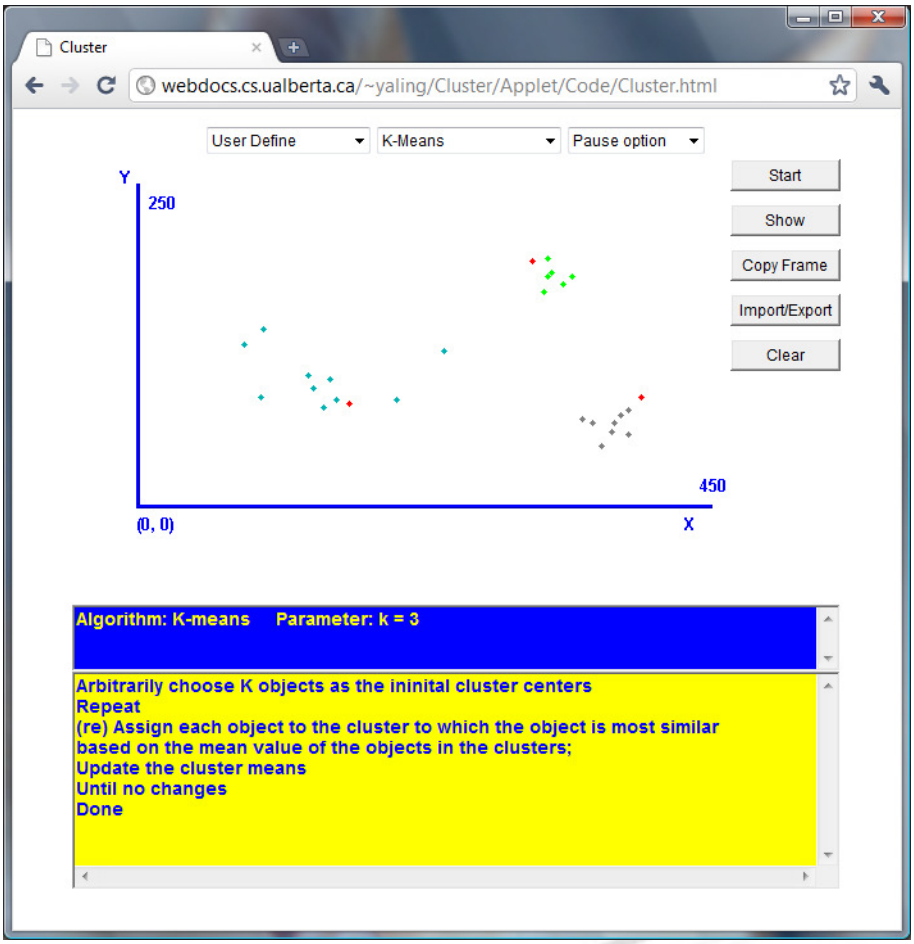
K-Means paso a paso

Ejecute ahora el siguiente applet:

http://www.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

Este applet nos permitirá crear nuestro propio conjunto de datos a golpe de ratón:

1. En la lista de conjuntos de datos [*Choose a Dataset*], escoja la opción correspondiente a conjuntos de datos definidos por el usuario [*User Define*].
2. Como algoritmo [*Choose an Algorithm*], elija *K-Means*.
3. Cree un conjunto de datos haciendo click sobre el área entre los ejes X e Y.
4. Pulse el botón Start para ejecutar el algoritmo de clustering.
5. Pulse el botón Show para visualizar los clusters obtenidos como resultado.



Al mostrar los resultados, con *Show*, se muestra cada punto con un color que representa el cluster al que ha sido asignado cada punto. Los puntos en rojo corresponden a los centroides de cada agrupamiento.

Usando la opción *Import/Export*, podemos generar el conjunto de datos correspondiente a los puntos que hayamos introducido manualmente. Los pares de valores mostrados por el applet van separados por un espacio en blanco y se pueden importar fácilmente en otras herramientas.

Cree una nube de 8 puntos cualesquiera en el plano. Cuando haya generado su conjunto de datos, incluya sus pares de valores al fichero `Clustering.doc` y capture la pantalla de la nube de 8 puntos (con *Alt+ImprPant*). Inclúyala en el fichero `Clustering.doc` (con *Ctrl+V*).

A continuación, escoja aleatoriamente dos centroides (indíquelos en `Clustering.doc`) y realice una ejecución manual del algoritmo, incluyendo, para cada iteración, los centroides y la tabla de distancias de todos los puntos a los centroides (tantas filas como centroides y tantas columnas como puntos). Termine la ejecución en cuanto hayan transcurrido 5 iteraciones o no varíen los centroides.

El ajuste del modelo puede medirse a través de la suma total de distancias al cuadrado de los puntos a sus centroides (SSE). Calcule el ajuste del modelo determinado por el valor final de SSE.

DatosApplet

Utilizando el mismo applet del ejercicio anterior, construiremos los siguientes ficheros de datos, con al menos 40 puntos en cada uno de ellos:

- `3ClustersSinOutliers.data` (nube de puntos con 3 clusters sin outliers).
- `3ClustersConOutliers.data` (nube de puntos con 3 clusters bien definidos, más 5 valores anómalos o outliers que no pertenezcan a ningún cluster).

A los dos ficheros anteriores, añádales una primera línea de texto con el nombre de las variables (a saber, X, Y).

Capture la pantalla (con *Alt+ImprPant*) de las tres nubes de puntos generadas y cópielas en el fichero `Clustering.doc` (*Ctrl+V*).

MUY IMPORTANTE:
Cada alumno deberá trabajar con sus propios datos.
No se admitirán dos prácticas con los mismos conjuntos de datos.

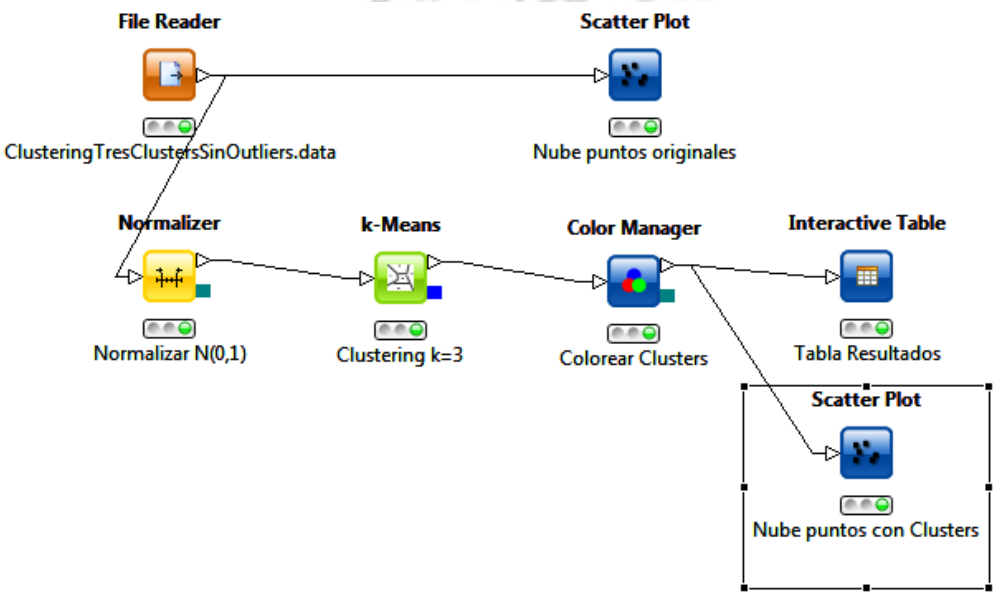
K-Means: DatosApplet @ KNIME

Cree un proyecto en KNIME llamado `Clustering_DatosApplet` con los siguientes nodos que nos permitan agrupar los datos de `3ClustersSinOutliers.data`:

- Un nodo para leer el fichero.
- Un nodo *Data Manipulation* > *Column* > *Normalizer* para normalizar los datos, ya que KNIME implementa k-Means sin normalizar previamente las variables. Con los datos obtenidos desde el applet no hay ningún problema, ya que ambas variables tienen magnitudes similares. Sin embargo, es bueno que nos acostumbremos a normalizar las variables.

  NOTA: Recordemos que, en SPSS, se hacía de la forma siguiente: *Analizar* > *Estadísticos Descriptivos* > *Descriptivos* > *Guardar valores tipificados como variables*.

- Un nodo *Mining* > *Clustering* > *k-Means* para realizar el clustering. Este nodo añadirá una columna *Cluster*, indicando el agrupamiento asignada a cada tupla de nuestro conjunto de datos. En su configuración, debemos indicar los atributos que se usarán para establecer los clusters. En nuestro ejemplo, usaremos ambos.
- Un nodo *DataViews* > *Property* > *Color Manager* para colorear los datos correspondientes a la columna *Cluster* del nodo anterior.
- Un nodo *DataViews* > *Interactive Table* para ver los resultados.
- Dos nodos *DataViews* > *Scatter Plot* para ver la nube de puntos original y la coloreada con los clusters obtenidos.



Para ver los centroides obtenidos, seleccione con la derecha el nodo *k-Means* y muestre los resultados (*View: Cluster View*).

Para escribir en un fichero los valores asociados a los centroides, seleccionamos un nodo *IO* > *Writer* > *PMML Writer* y lo enlazamos con el nodo *k-Means*. En la configuración del nodo PMML, debemos indicar el nombre del fichero de salida, `centroides_3ClustersSinOutliers.pmml` en nuestro caso.

El fichero PMML no es más que un fichero XML que se utiliza para describir modelos de minería de datos. Puede encontrar más información sobre el estándar PMML en:

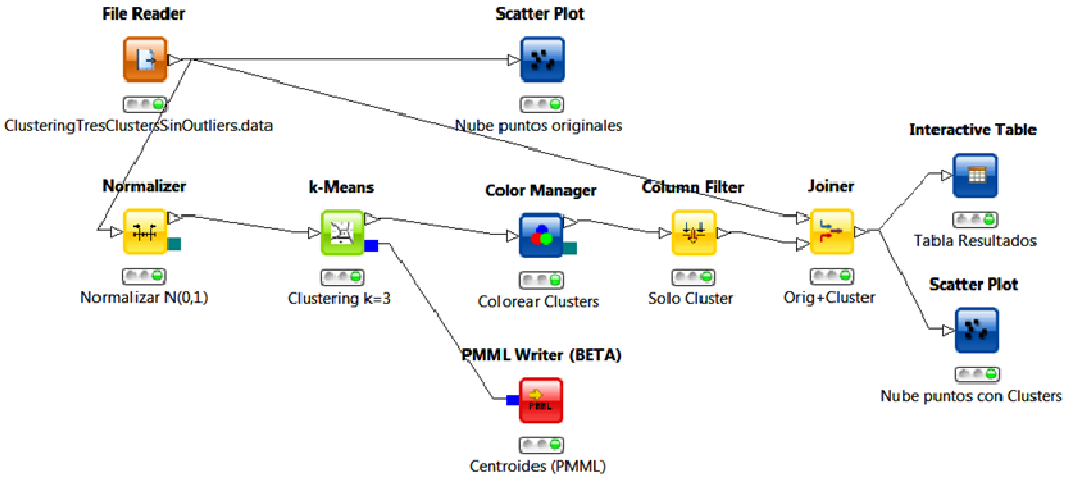
http://www.dmg.org/
http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language
http://www.frandzi.corex.es/PWP/frandzi/dm4pmml.html

Observe que, en nuestro proyecto actual, tanto la tabla como la nube de puntos nos muestran los datos normalizados. Si queremos ver los datos originales, basta hacer lo siguiente:

- De la salida de *k-Means*, mantenga únicamente el atributo *Cluster* y el identificador de fila (*RowID*). Esto lo podemos hacer con un nodo de tipo *Column Filter* aplicado sobre la salida de *Color Manager*.

- A continuación, combine el resultado del *Column Filter* con los datos originales mediante un nodo Joiner (eligiendo el *RowID* como *Join Column* e *Inner Join* como método de reunión).

Así quedaría nuestro proyecto KNIME tras realizar estas modificaciones:

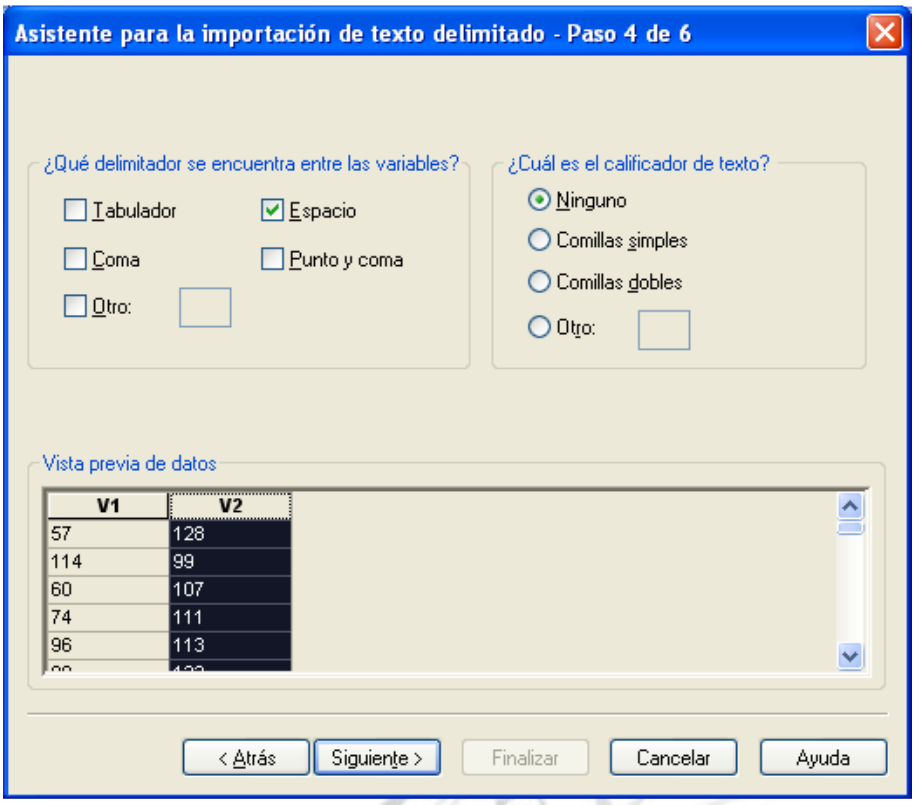


Abra ahora los nodos *Interactive Table* y *Scatter Plot*. Si pinchamos sobre un punto en la nube de puntos y con el ratón derecha seleccionamos *Hilite*, dicho punto se marca como naranja y, además, también se marca el dato correspondiente en la tabla interactiva.

Dentro del mismo proyecto KNIME, cree ahora un modelo análogo al anterior para el otro conjunto de datos, `3ClustersConOutliers.data`.

En el fichero `Clustering.doc`, introduzca después de cada nube de puntos los centroides obtenidos por K-Means y analice si los clusters generados se corresponden aproximadamente con los agrupamientos que cabría esperar.

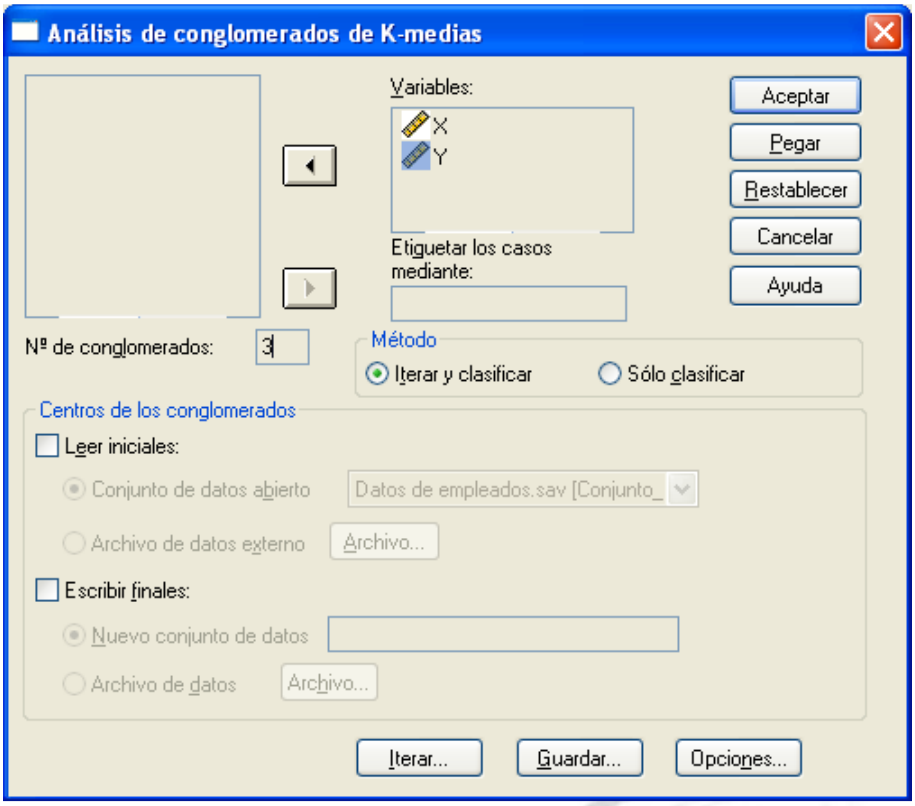
Ahora, lea los datos del fichero `3TresClustersSinOutliers.data` desde SPSS. Para ello, seleccione *Archivo* > *Leer datos de texto* e indique que en el fichero de datos utilizará espacios como delimitadores para separar los valores de las distintas variables:



En la siguiente pantalla del asistente, cambie el nombre de las variables V1 y V2 por X e Y, respectivamente. Para ambas variables, seleccione "*Numérica*" para especificar su tipo. Finalmente, guarde el fichero resultante en formato SPSS con el nombre `3ClustersSinOutliers.sav`.

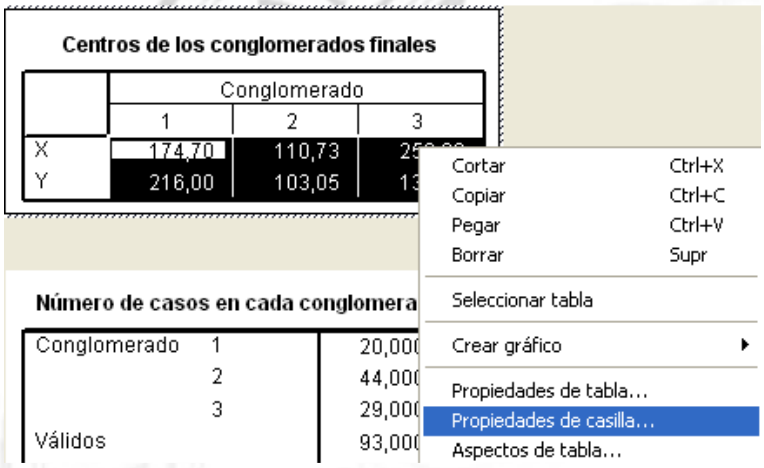
> IMPORTANTE: SPSS no normaliza las variables antes de ejecutar k-Means, por lo que habría que crear las variables normalizadas (*Analizar* > *Estadísticos Descriptivos* > *Descriptivos* > *Guardar valores tipificados como variables*) y usar k-Means sobre las estas nuevas variables, no sobre las originales.

A continuación, seleccione *Analizar* > *Clasificar* > *Conglomerado de K-medias*. Asegúrese de incluir las dos variables de nuestro conjunto de datos y escoja 3 como número de "conglomerados" (el valor del parámetro k):



Lance el "análisis de conglomerados" y, una vez ejecutado el algoritmo de las K Medias, aparecerán los centroides de los agrupamientos:

NOTA: Para que los valores asociados a los centroides aparezcan con decimales, deberá indicarlo en las opciones de las casillas:



Cuando trabaje con variables normalizados, debe tener en cuenta que los centroides también estarán normalizados. Para averiguar cuáles serían los valores originales (no normalizados) correspondientes a los centroides, debe deshacer la tipificación. Es decir, debe multiplicar el valor normalizado por la desviación típica y sumar la media aritmética (para cada una de las variables) .

Incluya en el fichero `Clustering.doc` las tablas correspondientes a los centroides, tipificados y no tipificados.

¿Salen los mismos resultados que los obtenidos con KNime? Razone la respuesta e incluya su razonamiento en el fichero `Clustering.doc`.

Repita el mismo estudio para el otro conjunto de datos: `3ClustersConOutliers.data`.

Antes de terminar, asegúrese de guardar los resultados de su sesión completa en SPSS, que deberá entregar en un fichero que se llame `Clustering_DatosApplet.spo`.

NOTA:

Weka sí implementa k-Means normalizando previamente las variables. Además, usa una variante de k-medoids para poder trabajar también con atributos nominales. En este caso, los centroides se construyen usando la media aritmética de los puntos del cluster para cada variable numérica y la moda de los puntos del cluster para cada variable nominal.

Clustering jerárquico

Ejecute el applet disponible en http://metamerist.com/slhc/example40.htm

Este applet muestra un conjunto de datos de ejemplo sobre dos variables. Paso a paso, muestra cómo un algoritmo de clustering jerárquico aglomerativo va formando clusters (usando la distancia euclídea para medir la distancia entre dos puntos). Al principio se conectan los dos puntos más cercanos, formando un primer cluster con dos elementos. Progresivamente, se irán uniendo puntos a clusters y clusters entre sí, mostrando cada cluster de un color. Al finalizar el proceso, nos quedará un único cluster.

Ejecute varias veces el applet con distintos conjuntos de datos de partida.

Clustering jerárquico: DatosApplet @ SPSS

Abra el fichero `Clustering_DatosApplet.spo` que ya creamos en la sección anterior. Seguiremos trabajando en él sobre los datos que ya teníamos en `3ClustersSinOutliers.sav`.

Para lanzar un algoritmo de clustering jerárquico en SPSS, seleccione *Analizar* > *Clasificar* > *Conglomerados Jerárquicos*.

A continuación, asegúrese de escoge las siguientes opciones:

Incluya las dos variables (X, Y) a la hora de realizar el clustering y, pinchando en el botón *Gráficos...*, indique que desea crear un dendrograma.

IMPORTANTE: Como sucedía con el algoritmo de las K Medias, por defecto no se normalizan las variables, a no ser que se indique explícitamente mediante la opción *Estandarizar* de la ventana que aparece al pulsar sobre *Método...* Hágalo.

Haga doble click sobre el dendrograma obtenido en el visor de resultados. Se abrirá otra ventana, en la que se puede editar (en formato de texto) dicho dendrograma. Quite el texto necesario para que no se vea partido y cópielo en el fichero `Clustering.doc`.

Identifique de alguna forma los 3 clusters naturales que deberían salir. Para ello, puede usar colores diferentes (o algo más sofisticado si utiliza un procesador de textos menos rudimentario que el WordPad).

En el visor de resultados de SPSS (al igual que en Clementine y Weka), también aparece el número de ejemplos que han sido asignados a cada cluster. En cualquier estudio, habría que analizar estos datos detenidamente, ya que es posible que tuviésemos que dividir aquellos clusters que tuviesen demasiados ejemplos (aumentando k) , lo que sucede a menudo ante la presencia de outliers (el outlier sería un cluster y todos los demás datos quedarían englobados en otro). No obstante, hay que se cuidadoso con este tipo de cosas, ya que también pueden existir agrupamientos naturales que engloben a una parte significativa de nuestro conjunto de datos.

Detección de outliers: DatosApplet @ SPSS

Veamos ahora cómo detectar outliers. Si usamos un algoritmo de clustering jerárquico, un caso extremo o outlier se unirá en algún momento a aquel cluster que esté más cercano a él. Si bien que un cluster contenga muchos ejemplos no quiere decir que necesariamente contenga outliers, lo que sí es cierto es que un outlier se unirá siempre a un cluster en las fases finales del agrupamiento jerárquico: los outliers serán aquellos valores que, de forma aislada, se unen a algún cluster en las últimas iteraciones del algoritmo jerárquico.

Cargue desde SPSS el fichero `3ClustersConOutliers.sav` y vaya añadiendo su análisis al fichero `Clustering_DatosApplet.spo`.

Realice ahora un clustering jerárquico, copie el dendrograma en el fichero `Clustering.doc` e inspecciónelo para detectar los valores correspondientes a los outliers, indicando en el fichero cuáles son.

A continuación, elimine los registros corresondientes a los outliers del conjunto de datos (bien físicamente, o bien excluyéndolos del análisis) y vuelva a ejecutar el algoritmo de las k Medias con k=3. Guarde la tabla de centroides resultantes en `Clustering.doc` y compare los centroides obtenidos al incluir los outliers y al excluirlos del análisis. Analice ambos resultados (y no olvide incluir sus comentarios en `Clustering.doc`).

> CUIDADO: Cada vez que se elimina un registro, SPSS cambia todos los identificadores. Si vamos a eliminar los registros con identificadores 35 y 40, por ejemplo, y borramos primero el 35, el segundo registro que queríamos borrar tendrá ahora el identificador 39. Para evitar errores, lo mejor es eliminar los registros de mayor a menor orden.

Ejercicios tipo B

K-Means: Iris @ KNIME

Copie el fichero `iris.csv` (disponible en la web de DECSAI) y cree un proyecto en KNIME llamado `Clustering_Iris`.

Este conjunto de datos, muy conocido, contiene las longitudes y anchuras de pétalos y sépalos de 150 muestras correspondientes a tres variedades de plantas de tipo Iris (lirios), además de un quinto atributo *class* que indica el tipo concreto de Iris (Setosa, Virgina o Versicolour) . Más información sobre este conjunto de datos en http://archive.ics.uci.edu/ml/datasets/Iris

Para ver si las características morfológicas de las muestras determinan el tipo de Iris, utilizaremos un algoritmo de agrupamiento como el de las K Medias.

En su proyecto de KNIME, lea los valores del fichero `iris.csv`, elimine el atributo *class* y normalice los atributos numéricos. A continuación, utilice k-Means con k=3 e indique en `Clustering.doc` cuáles son los centroides que ha encontrado.

Finalmente, construya una tabla con los datos originales, incluido *class*, y añada el atributo *Cluster* que indique el cluster al que ha sido asignada cada muestra de iris. En una tabla interactiva, puede seleccionar la cabecera de una columna con el ratón y arrastrarla al sitio que se quiera, de tal forma que puede colocar *class* y *Cluster* en columnas adyacentes.

Para ver en cuántos registros se ha "acertado" al agrupar los datos, utilice un nodo *Mining* > *Scoring* > *Scorer*, filtrando antes todas las columnas excepto *class* y *Cluster*. Este nodo cuenta el número de registros que contienen cada uno de los valores de *class*, cruzándolos con todos los de *Cluster*. Exporte la tabla resultante del nodo *Scorer* (con *File* > *Export as PNG*) e incluya el gráfico en el `Clustering.doc` junto con sus comentarios sobre los resultados obtenidos.

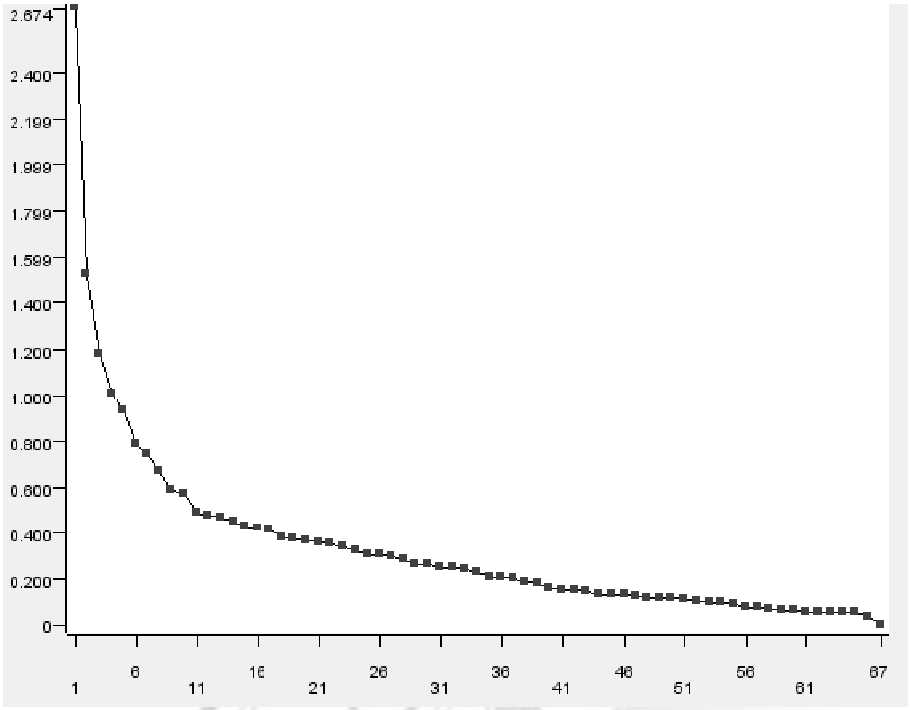
Clustering jerárquico: Iris @ KNIME

Añada un nodo *Mining* > *Hierarchical Clustering* a su proyecto KNIME. Conéctelo a sus datos de entrada (quitando previamente la columna *class* y normalizando variables). Configure el nodo para que que se utilice la distancia euclídea y *single linkage*. Ejecute el algoritmo de clustering y copie el dendrograma en el fichero `Clustering.doc` (*View* > *Dendrogram/Distance View* y, sobre la ventana que aparece, *File* >*Export as PNG*). Analice los resultados e indique los grupos que se aprecian en el dendrograma.

Para comprobar si los grupos marcados en el dendrograma coinciden con los obtenidos por *k-Means* o con los datos originales, puede seleccionar una zona del dendrograma con el ratón y seleccionar *Hilite*. Si tenemos abierta la tabla con los datos originales, se marcarán automáticamente los correspondientes registros y podremos comprobar si coinciden los valores de *class* (*versicolor, setosa, virginica*).

Añada ahora otro nodo de clustering jerárquico, pero esta vez utilice la distancia euclídea con *average linkage*. Cope el dendrograma en el fichero `Clustering.doc`, compárelo con el dendrograma anterior y analice los resultados.

En la ventana en la que se muestra el dendrograma, hay una pestaña *Distance* que muestra los valores de SSE (la suma de las distancias al cuadrado de cada punto al centroide de su cluster). Esto puede resultar bastante útil para determinar el valor más adecuado de k. Cuanto mayor sea k, menor será el valor de SSE, pero habrá un punto a partir del cual la ganancia no será significativa. Este será el punto que determinará el valor más adecuado para k.



Exporte el gráfico anterior para los dos clustering jerárquicos que ha realizado y proponga cuál sería su recomendación sobre el valor k (que posiblemente estaría entre 4 y 6 para la figura que se muestra arriba).

> NOTA: SPSS ofrece otro método de clustering ("en dos fases"). Éste no es más que una variante del método BIRCH en la que se obtiene automáticamente una estimación del valor k usando precisamente el gráfico anterior. Así pues, el resultado de este método es directamente es una tabla de centroides, pero sin tener que indicar a priori el valor de k (ni estimarlo manualmente).

Ejercicios tipo A

Datos de Empleados

Cargue los datos de empleados en SPSS. Como son bastantes datos, el clustering jerárquico puede ser muy ineficiente, por lo que seleccionaremos un 40% de la muestra aplicando un filtro con *Datos* > *Seleccionar Casos* > *Muestra Aleatoria*. Guarde los datos en formato `csv`, en el fichero `DatosEmpleados.muestra.csv`.

Cree un nuevo proyecto en KNIME, llamado `Clustering_DatosEmpleados` para leer los datos de nuestra muestra de empleados y lanzar un algoritmo de clustering jerárquico. No olvide realizar todas las operaciones de preprocesamiento que estime necesarias, como, por ejemplo, seleccionar las variables que se utilizarán en el método de agrupamiento. Copie el dendrograma resultante y el gráfico de distancias SSE en el fichero `Clustering.doc`. Estime un valor adecuado para k.

A cotinuación, utilice el método de las K Medias utilizando el valor estimado para k mediante el algoritmo de clustering jerárquico. Guarde la tabla resultante de centroides en `Clustering.doc` e interprétela en términos de los k perfiles o prototipos de clientes que haya obtenido.

DatosApplet

Cree con el applet utilizado al principio de este guión una nube de puntos que corresponda a tres clusters cuya forma sea difícilmente detectable por el método de las K Medias (por ejemplo, clusters no esféricos o de distinta densidad). Guarde los puntos generados en el fichero `ClusteringDificil.data`.

Utilice KNIME para realizar un agrupamiento basado en un nodo de tipo *Weka* > *Cluster Algorithms* > *SimpleKMeans,* configurado para un valor de k=3. Muestre el ajuste final del modelo, indicando el valor SSE (este dato está disponible en la salida del nodo Weka, etiquetado como *Within cluster sum of squared error*).

Estime ahora un valor de k con un diagrama SSE (pestaña *Distance view* del nodo de clustering jerárquico) y vuelva a lanzar el algoritmo de las K Medias. Indique de nuevo el valor SSE y compárelo con el anterior. Incluya los gráficos obtenidos y la discusión que se estime oportuna en el fichero `Clustering.doc`.

EVALUACIÓN DE LAS PRÁCTICAS

El fichero `Clustering.doc` ha de incluir todas sus respuestas a las distintas preguntas que aparecen en este guión.

PD: No olvide incluir también sus proyectos KNIME y el fichero correspondiente a su sesión en SPSS.