Trabajo sobre Preprocesamiento de Datos y Comparación de Modelos

Minería de datos

José Tomás Palma Méndez

Dept. of Information and Communication Engineering. University of Murcia
Contacting author: jtpalma@um.es

1. Ejercicios sobre preprocesamiento

Ejercicio 1. Abre con algún editor el fichero echocardiogram.data que podrás encontrar en la carpeta **DataSets** en los recursos del aula virtual.

1.a) ¿Cuáles crees que deben ser los tipos asociados a cada columna?
1.b) ¿Existen valores desconocidos? ¿Cómo están representados?
1.c) ¿Qué información crees que falta?

Ejercicio 2. Importa el fichero echocardiogram.data desde RStudio, sin modificar los parámetros.

2.a) Ejecuta el comando str(echocardiogram.data) y copia el resultado ¿Qué anomalías encuentras? Enuméralas.
2.b) La función complete.cases nos indica el número de filas completas que hay en el data.frame. Ejecútala sobre el conjunto de datos importados ¿Es el resultado esperado?
2.c) Prueba a volver a cargar los datos desde RStudio utilizando correctamente le parámetro na.strings. Una vez importados los datos ejecuta el comando str y compara los resultados con el caso anterior ¿Ves alguna anomalía?
2.d) ¿Cuáles son las filas con valores desconocidos?

Ejercicio 3.

3.a) Realiza las operaciones necesarias para asignar los tipos adecuados según la información contenida en el fichero `echocardiogram.names`.
3.b) Da los siguientes nombres a las columnas: Survival, StillAlive, AgeAttack, PericardEffu, FracShort, EPSS, LVDD, WMS, WMI, Mult, Name, Group, AliveAt1.
3.c) De acuerdo con la información suministrada, calcula los valores ausentes de la columna de clasificación (la úlitma).

3.d) Antes del proceso de imputación y en función de la distribución de NA, indicar si sería conveniente eliminar alguna instancia o atributo. Razona tu respuesta.

Ejercicio 4. Partiendo del conjunto de datos generado en el Ejercicio 3:

4.a) Genera algunas gráficas con las funciones del paquete VIM que permitan visualizar la distribución de los valores ausentes.
4.b) Utiliza la función impute() para imputar los valores ausentes en los atributos que contienen valores desconocidos. Utiliza los métodos median y mean y analiza los resultados ¿Han cambiado el tipo de los atributos?
4.c) Compara los resultados con los que se obtendrían con la función mice() y el método pmm y la imputación con la función kNN.
4.d) Genera cuatro ficheros con los resultados obtenidos que contengan sólo las columnas útiles para construir un clasificador. Los ficheros deberán tener los siguientes nombres: echo.medianImpute.csv, echo.meanImpute.csv, echo.pmmImpute.csv y echo.kNNImpute.csv.

2. Ejercicios sobre comparación de modelos

Ejercicio 5. Se quiere comprobar si existen diferencias significativas en el comportamiento de dos clasificadores: SVM y una análisis lineal discriminante (LDA). Para tal fin, disponemos de los resultados de la precisión para cada uno de los pliegues para ambos clasificadores (fichero ejercicio1.dat). También se dispone de los objetos generados por la función `train()` correspondientes a los dos clasificadores (ficheros SVMFit y LDAFit) así como las base de datos sobre la que se ha hecho el experimento (el último atributo es la clase que es un factor, el primero es real y el resto enteros). Realizar las comprobaciones necesarias para llegar a una conclusión.

Ejercicio 6. Se quiere comprobar si existen diferencias significativas en el comportamiento de cuatro regresores: RegA, RegB, RegC y RegD. Para tal fin, disponemos de los resultados del error cuadrático medio de cada regresor medido sobre nueve conjunto de datos (fichero ejercicio2.dat). Realizar las comprobaciones necesarias para llegar a una conclusión.