



UNIVERSITAS
Miguel Hernández

Tema 5. Árboles de Decisión. Ejemplos.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones.

1. Carga y Preparación de los Datos

- Descargar el archivo `clasificacion.xlsx` que contiene los datos sobre los que deseamos emplear la técnica de clasificación.
- Leer el archivo Excel en un `DataFrame` de `pandas`.
- Visualizar las primeras filas del conjunto de datos.
- Definir las variables independientes X_1 , X_2 , X_3 y la variable dependiente Y .

2. Obtención del modelo.

- Divide los datos en un conjunto test del 30% y uno de entrenamiento del 70%.
- Obtén mediante CART un árbol de decisión para clasificar los individuos de dicha base de datos.

3. Evaluación del modelo

- Obtén los pronósticos de la muestra de prueba según el modelo anterior.
- Obtén la tabla de confusión.
- Obtén el informe de clasificación del modelo, utilizando la librería sklearn.

4. Interpretación.

- Obtén un diagrama del árbol de decisión que explique el proceso de clasificación.

5. Parámetros.

En la función `DecisionTreeClassifier` pueden configurarse distintos parámetros para configurar el árbol obtenido:

- `max_depth`: la profundidad máxima del árbol (valor entero).
- `min_samples_split`: el número mínimo de muestras necesarias para dividir un nodo (valor entero o porcentaje).
- `min_samples_leaf`: el número mínimo de muestras necesarias en una hoja (valor entero o porcentaje).
- `max_leaf_nodes`: limita el número máximo de hojas en el árbol (valor entero).

Experimenta con distintas configuraciones hasta obtener el árbol que consideres más relevante.

6. Árboles de decisión aplicados a regresión.

- Realiza una regresión sobre el fichero 'regresion.xlsx' mediante la técnica CART.
- Obtén el error cuadrático medio y el R^2 del modelo.
- Obtén un gráfico de los valores observados frente a los valores predichos.



UNIVERSITAS
Miguel Hernández