Data Science para Economía y Negocios

ARBOLES DE CLASIFICACIÓN

JAVIER FERNÁNDEZ Y ESTEBAN LÓPEZ

Motivación

- Como hemos visto a lo largo del curso, el análisis de datos es útil en muchas disciplinas (Salud, PP, deportes, etc).

- En particular, en negocios, hemos visto que es una herramienta para agregar valor a las empresas facilitando la toma de decisiones.

Motivación

- Un objetivo de DS es predecir!
- La clase 10 revisamos como predecir sobre variables numéricas y continuas
- ¿Qué sucede con las variables categóricas?
- ¿Podemos predecir el valor de una variable que no es númerica?

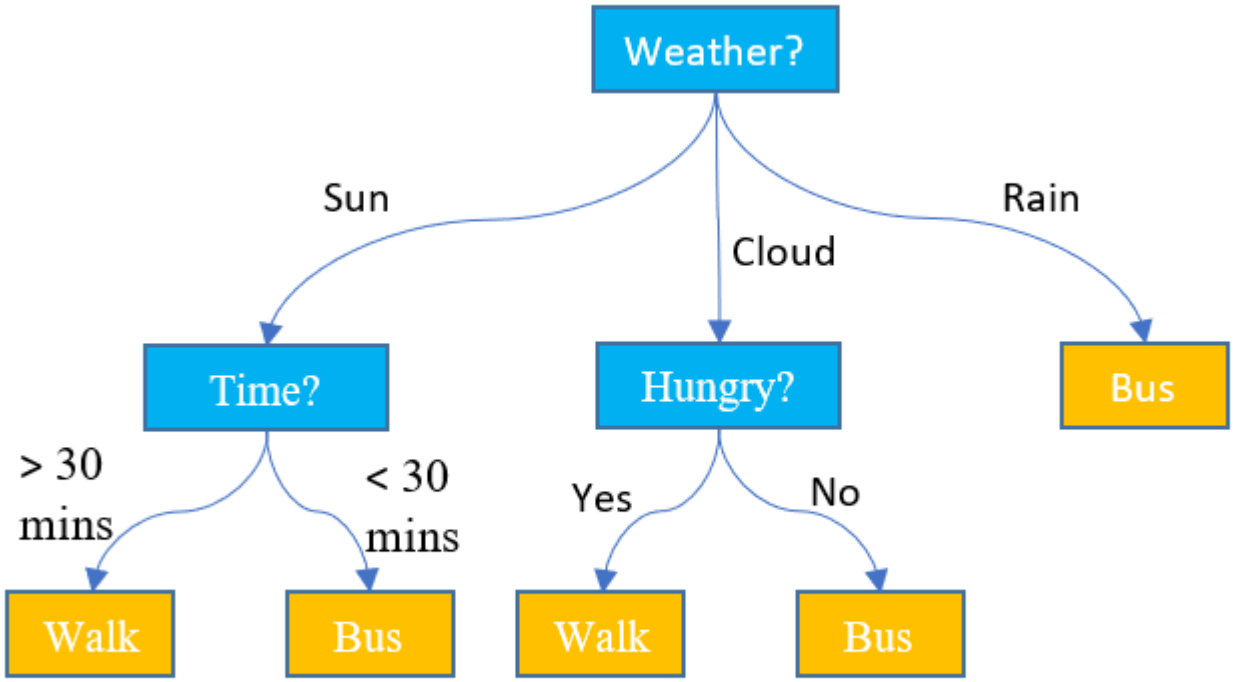
¿Existe técnicas para clasificar datos?

- Existen varias técnicas (L.Reg, NN, SVM, Neural Networks, etc) la que hoy revisaremos es los arboles de clasificación:

  - *Nos permiten clasificar los datos dependiendo de reglas de decisión sobre variables de la base de datos.*

  - *Se considera parte de los arboles de decisión, al igual que los arboles de regresión.*

Arboles de decisión



Fuente: displayr.com

Arboles de decisión

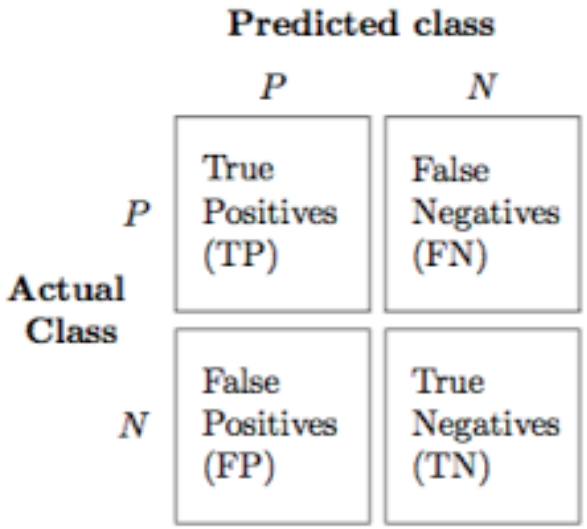
- Este método de nos permite clasificar los datos dependiendo de reglas de decisión sobre variables de la base de datos.
- Es decir, divide la muestra en subconjuntos dependiendo del valor de otras variables de la base de datos.
- Como es un método supervisado (conozco la variable sobre la cuál estoy prediciendo el modelo) puedo utilizar validación cruzada para comprobar la precisión del modelo.

Evaluar Clasificación

Precisión de la predicción:

$$\frac{diag(Matriz\ de\ Confusión)}{sum(A)}$$

Matriz de confusión:



Lab 12

- Utilizaremos los paquetes “**rpart**”, “**rpart.plot**” y **“caret”** para realizar análisis. Las funciones a utilizar son las siguientes:
  - rpart: permite utilizar arboles de decisión. Argumentos:
    - formula: variables del modelo.
    - data: datos.
    - method: método a utilizar. Para arboles de decisión usar “class”.
  - createDataPartition: Nos permite crear un index para set de entrenamiento y de prueba. Argumentos:
    - base de datos o columna por la cual se creará la partición.
    - times: número de particiones a crear.
    - p: proporción de la muestra.
    - list: TRUE or FALSE. Sí es true crea una lista, si es False un vector.
  - rpart.plot: plotea arboles de decisión. Necesita un objeto creado con rpart.