

Guía de lectura

Introduction to Statistical Learning. Springer, 2013

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Desarrolle los siguientes puntos basándose en la lectura.

Capítulo 8, punto 8.1: árboles de decisión

- Dado un árbol de decisión, entienda el proceso de predicción de una clase o valor numérico para un nuevo registro.
- Explore el algoritmo de aprendizaje de árboles de decisión (tanto de clasificación como de regresión).
- ¿Por qué se dice que la construcción de árbol de decisión es un proceso *greedy* (avaro)?
- ¿Hay que normalizar los datos antes de entrenar un modelo de árboles de decisión?
- ¿Por qué las regiones de particionamiento del espacio de representación de las variables predictivas tiene forma rectangular?

Hasta aquí sería la discusión en la sesión del jueves 3 de Noviembre.

- ¿Cuáles son los criterios de particionamiento que se utilizan para aprender árboles de decisión (tanto de regresión como de clasificación)? ¿Por qué no se utiliza el error de clasificación sino el Gini o el cross-entropy?
- ¿Cómo se estima la **complejidad** de un modelo de árbol de decisión? (1 línea)
- ¿Cómo se combate el overfitting en árboles de decisión?
- Que es mejor con respecto al modelo final: ¿dejar de crecer un árbol (pre-poda) o crecer un árbol hasta los límites de lo posible y luego podarlo (post-poda)?
- Dados estos datos, calcule la entropía y el gini de la variable objetivo, y el gini y las entropías condicionales para cada variable. Escoger la variable de particionamiento.

Genero	Edad	Compra Salchicha
H	Joven	SI
H	Joven	NO
M	Viejo	SI
H	Viejo	NO
M	Joven	SI
H	Joven	SI
H	Joven	NO
M	Viejo	SI
H	Joven	SI
M	Viejo	SI