uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

Modelos básicos para clasificación y regresión : árboles y k-vecinos más cercanos (KNN)

Esta clase tiene tres objetivos:

- Describir la *pipeline* básica de aprendizaje automático (la secuencia de procesos que generalmente se siguen en Machine Learning).
- Recordar algunos conceptos básicos: instancias, atributos y espacio de instancias.
- Describir dos de los algoritmos básicos de Machine Learning: árboles de decisión (y reglas) y k-nearest neighbors (k-vecinos más cercanos)

Sobre los árboles de decisión:

- El algoritmo básico que crea los árboles de decisión se explica a través de un ejemplo.
- Se muestra que el algoritmo es recursivo y básicamente equivale a elegir el mejor atributo en cada nodo, minimizando una medida llamada entropía, aunque también se podrían utilizar otras medidas como Gini, que tienen una forma similar. Dado que los atributos son evaluados de acuerdo con la calidad de las particiones que generan, cualquier medida que sea capaz de clasificar correctamente las particiones de datos también podría, en principio, utilizarse para evaluar atributos.
- Se muestra cómo extender el algoritmo básico a las tareas de regresión, a través de árboles de modelos y árboles de regresión. En este caso, la medida a minimizar es la varianza, en lugar de la entropía. Los árboles de modelos contienen modelos lineales en las hojas, mientras que los árboles de regresión son más simples, ya que únicamente contienen constantes en las hojas. Los modelos lineales se construyen a

partir de instancias que alcanzan esa hoja en particular, mientras que las constantes son el promedio de las instancias que alcanzan la hoja.

Sobre el algoritmo KNN:

• KNN se explica tanto para clasificación como para regresión. KNN es un algoritmo que no construye un modelo explícito a partir de los datos. Por el contrario, almacena el conjunto de datos de entrenamiento y la predicción se hace en el momento que llega una instancia de test. Para hacerlo, se computan las distancias entre la instancia de test y todas las instancias de entrenamiento. La clase de la instancia más cercana es la respuesta de KNN. Se pueden usar valores mayores que 1 para K. En ese caso, la predicción es la clase mayoritaria entre las k instancias de entrenamiento más cercanas. Para problemas de regresión, una estrategia simple es calcular el promedio de las k instancias de entrenamiento más cercanas.

• Se explican algunas de las ventajas de KNN (la principal es que no es necesario para construir un modelo explícito), y los inconvenientes (lentitud clasificando nuevas instancias, el problema con atributos irrelevantes, la maldición de la dimensionalidad y los problemas con datos ruidosos)

• Se muestra lo que significa el hiperparámetro k, su influencia en la clasificación y cómo seleccionarlo. En particular, k> 1 puede ser útil para problemas con ruido, como una forma de promediar o suavizar el ruido.

Material asociado:

Además de las diapositivas de clase y algunos ejercicios (consultar la guía del curso), existe un tutorial bastante completo sobre la librería de aprendizaje automático MLR y una práctica. Más específicamente, los tutoriales incluyen un breve repaso del lenguaje R, el uso directo de R para construir y evaluar árboles y KNN, y especialmente, el uso de MLR para entrenar, visualizar y evaluar los modelos básicos vistos en la clase.