

Descripción de Introduction to Statistical Learning

Antonio Álvarez Caballero
analca3@correo.ugr.es

12 de junio de 2016

Índice

1. Breve descripción del curso	2
---------------------------------------	----------

1. Breve descripción del curso

En el año 2014, los profesores de la universidad de Stanford Trevor Hastie y Rob Tibshirani hicieron un curso online basado en su libro de texto *An Introduction to Statistical Learning with Applications in R*. Tanto este libro como el curso derivado es una gran iniciación a lo que llamamos *Aprendizaje Automático*, *Aprendizaje Estadístico*, o más globalmente *Machine Learning*. Es una rama de la inteligencia artificial cuyo objetivo es el desarrollo de técnicas que permitan a un ordenador *aprender*.

En el curso indicado se estudian durante 10 temas una buena iniciación a este ámbito. Veamos un poco de cada uno de los temas que contiene el curso.

1. Introducción. Se introduce al tema incluyendo ejemplos clásicos como el de Netflix.
2. Aprendizaje estadístico. Se dan unos pequeños conceptos de estadística descriptiva y regresión. Se ven modelos paramétricos y se introduce al análisis sesgo-varianza, a los problemas de clasificación y al clasificador más sencillo, el K vecinos más cercanos. En la práctica se hace una introducción a R.
3. Regresión lineal. Se recuerdan conceptos de regresión lineal e intervalos de confianza, además de test de hipótesis. Se explica la selección de modelos y se introduce la no linealidad. En la práctica se introduce la regresión lineal en R, ajustando varios modelos para entenderlo mejor.
4. Clasificación. Se introduce al problema de clasificación (binaria), regresión logística y máxima verosimilitud. También se introduce, aunque no conozco suficiente del tema, regresión logística multiclase, análisis discriminante y teorema de bayes, además del clasificador Naïve Bayes, el clasificador *estúpido*. En práctica se aplican regresión logística, análisis discriminante y K vecinos.
5. Métodos de muestreo. En este tema se ven técnicas para predecir el error fuera de la muestra y se explican técnicas de validación cruzada, además de bootstrap. En prácticas se ven ambas técnicas.
6. Selección de modelos lineales y regularización. En este tema se explica cómo seleccionar un modelo lineal, estimación de cotas, reducción de la dimensionalidad del problema y regularización de los datos, que sirve para minimizar el sobreajuste en la medida de lo posible. En prácticas se aplica todo esto a casos prácticos, incluidos los métodos LASSO, que son métodos de selección de características.
7. Saliéndonos de la linealidad. En este tema se ve una introducción a modelos no lineales.
8. Modelos basados en árboles. En este tema se introducen los árboles de decisión y los métodos *ensemble* o aditivos, donde se combina la potencia de muchos árboles para formar clasificadores (o regresores) de gran calidad. En prácticas se ajustan modelos de árbol y de *ensemble*.
9. Máquinas de vectores soporte. En este tema se explican las máquinas de vectores soporte: clasificadores lineales con máximo margen, las cuales pueden ofrecer la solución óptima para un clasificador lineal. En prácticas se ajustan varios de estos modelos, incluyendo los que tienen núcleo no lineal.
10. Aprendizaje no supervisado. Se ven técnicas de aprendizaje no supervisado, como el K-medias o clustering jerárquico. En prácticas se ajustan estos dos modelos.

Personalmente, conocía este curso, ya que he usado el libro asociado como referencia puntual en mi curso de Aprendizaje Automático de la UGR. En particular, mi curso de referencia principal ha sido uno similar, [Learning from Data](#)¹, de CalTech, pero la parte práctica de este curso es bastante completa y ayuda a entender los conceptos introducidos, además de hacer más sencillo su aprendizaje en R. El curso de CalTech no tiene parte práctica.

¹<https://work.caltech.edu/telecourse.html>