Introducción © EDICIONES ROBLE, S.L.

Indice

Introducción

Introducción. Guilermo González Sánchez

I. Introducción al módulo

El dominio de las técnicas de inferencia y aprendizaje estadístico es una competencia básica del científico de datos. Los modelos estadísticos son útiles en muchas situaciones y son también la base de la comprensión de los datos, sus distribuciones y sus relaciones. Este módulo también servirá para afianzar o adquirir los conocimientos de probabilidad y estadística que se utilizan en algunos algoritmos de aprendizaje automático, aunque lo esencial de este se encuentra en las propiedades del algoritmo.

En este módulo, se recogen los fundamentos mínimos necesarios para ser capaz de aplicar métodos y técnicas estadísticas. El enfoque es aplicado, utilizando el lenguaje de programación estadístico R.

El módulo comienza con una unidad orientada a conocer los fundamentos de R como lenguaje de manipulación de datos, de manera similar a como se utilizó Python en la última unidad del módulo 1. Esto permite al estudiante poner en práctica y ampliar conocimientos cuando trabaje con datasets, que puede descargar de Internet.

La segunda unidad contempla el análisis exploratorio de datos, que se centra en entender la distribución de las variables, la calidad de los datos y las relaciones entre esos datos. En este análisis exploratorio, el uso de gráficos es eficaz si lo que se desea es obtener rápidamente una idea de los datos.

La tercera unidad trata de la inferencia estadística, es decir, generalizar hipótesis acerca de poblaciones a partir de muestras. Además de su interés como técnica de obtención de conocimiento, es la base de una gran cantidad de test estadísticos que se utilizan en muchas tareas de data science.

La cuarta unidad expone los modelos lineales de regresión: cómo crearlos, evaluarlos y seleccionar entre varios de ellos. Se explican nociones esenciales del diseño de modelos estadísticos, como el sesgo y la varianza, y cómo realizar la selección del modelo.

La quinta unidad está dedicada a tratar el problema de clasificación usando regresión logística, que ayudará a entender más adelante los modelos de redes neuronales. Además, se expone el algoritmo fundamental que usan mayoritariamente los algoritmos que son capaces de aprender datos: el gradiente descendente estocástico. Se introducen modelos avanzados para tipologías de datos particulares como Lasso o Ridge.

Finalmente, la sexta unidad introduce técnicas estadísticas avanzadas específicas de ciertos tipos de datos. Trata el problema de clasificación múltiple, regresión de enteros no negativos o de valores reales no negativos usando modelos lineales generalizados. También se da la definición y conceptos de serie temporal y los métodos de análisis lineal clásicos de Box-Jenkins: ARIMA y SARIMA.

II. Objetivos generales

Los objetivos generales del módulo que los alumnos alcanzarán tras su estudio son:

1. Entender los fundamentos de la programación con estructuras vectoriales y matriciales, que son la base de los lenguajes que utilizan los data scientists, en el entorno de la programación estadística.
2. Saber realizar análisis exploratorios sobre los datos utilizando estadísticas y técnicas gráficas eficaces.
3. Ser capaz de crear, evaluar y seleccionar modelos lineales para tipos de datos diversos.
4. Comprender el aprendizaje estadístico y las técnicas en las que se basa, así como la regresión logística como ejemplo que tiene relación con otros modelos de aprendizaje automático.
5. Conocer y saber seleccionar modelos de análisis y aprendizaje estadístico avanzados específicos de ciertas tipologías de datos.

El módulo se inicia exponiendo conceptos fundamentales, que luego serán desarrollados en el entorno de programación R y su IDE RStudio, de tal modo que todas las ideas que se ven aplicadas sean reproducibles más adelante.

A lo largo del manual se irán introduciendo llamadas a notebooks para poder consultar la información, las explicaciones y los ejemplos que se ubican en este formato.