Estádistica Inferencial

Alejandro Sánchez Yalí

1. Preliminares

Para comenzar a trabajar con el aprendizaje computacional, necesitamos desarrollar algunas habilidades básicas. Todo el aprendizaje computacional está relacionado con la extracción de información de los datos. Así que comenzaremos por aprender habilidades prácticas para almacenar, manipular y pre-procesar datos. Además, el aprendizaje computacional por lo general trabaja con grandes cantidades de datos, los cuales podemos pensar como tablas numéricas, en donde las filas corresponden a ejemplos y las columnas a atributos. El Álgebra Lineal nos da un conjunto de técnicas para trabajar con datos representados como tablas numéricas. No nos extenderemos demasiado en la teoría del Álgebra Lineal, por ahora solo nos enfocaremos en las operaciones básicas de matrices y su implementación.

2. Datos, modelos y aprendizaje

Existen tres componentes principales en un sistema de aprendizaje computacional: datos, modelos y aprendizaje. La pregunta principal del aprendizaje computacional es ¿Qué entendemos por un buen modelo? La palabra modelo tiene muchas sutilezas, y lo estaremos revisando varias veces en este capítulo. No es fácil definir objetivamente que se quiere decir por «bueno». Unos de los principios que guía al aprendizaje estadístico es que los buenos modelos se deben tener un buen desempeño sobre datos que nunca han visto. Esto requiere que se definan algunas métricas de desempeño, tales como la exactitud o la distancia a los datos reales, así como también averiguar formas o estrategias para mejorar estas métricas de desempeño.

2.1. Datos como vectores

Asumimos que nuestros datos pueden ser leídos por un computador, y representados adecuadamente en una tabla numérica, en donde cada fila de la tabla representa un ejemplo o instancia particular, y cada columna una característica particular. En años recientes, el aprendizaje computacional ha sido aplicado a muchos tipos de datos que no son presentados en tablas numéricas, por ejemplo: secuencias genómicas, texto e imágenes de una página web y grafos de una red social.

2.2. Funciones como modelos

Un *predictor* es una función que cuando recibe una ejemplo (en nuestro caso, un vector de características), produce una *predicción*. Por hora, vamos a considerar la *predicción* como un solo número real. Esto puede ser escrito como:

$$f : \mathbb{R}^d \to \mathbb{R}, \tag{2.1}$$

donde el vector de entrada $\boldsymbol{x}$ es d-dimensional (tiene d características), y la función f es aplicada sobre él (escrito como $f(x)$) regresa un valor real. En estos apuntes no vamos a considerar el caso general de todas las funciones, que podría estar involucradas, en su lugar, solo vamos a considerar el caso especial de la funciones lineales:

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0 \tag{2.2}$$

para valores desconocidos de $\boldsymbol{\theta}$ y θ_0.

2.3. Modelos como distribuciones de probabilidad

En lugar de considerar un predictor como una simple función, nosotros podemos considerar los predictores como modelos probabilísticos, es decir; modelos que describen la distribución de funciones posibles.

2.4. Hipótesis de la clase de funciones

Dado un número n de ejemplos $x_i \in \mathbb{R}^d$ y es el correspondiente valor escalar $y_n \in \mathbb{R}$. Nosotros consideramos la tarea de aprendizaje supervisado, sobre los datos $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)$. Dado este conjunto de dato, deseamos estimar el predictor $f(\cdot, \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$, parametrizado por $\boldsymbol{\theta}$. El objetivo es encontrar un buen parámetro $\boldsymbol{\theta}^*$ tal que la predictor resultante se ajuste bien a los datos, esto es,

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}^*) \approx y_i \text{ para todo } i = 1, \ldots, n. \tag{2.3}$$

En esta sección, nosotros usamos la notación $\hat{y}_i = f(x_i, \theta^*)$ para representar la predicción del predictor.

Ejemplo 2.1. Vamos a introducir el problema de regresión por mínimos cuadrados ordinarios para ilustrar el proceso de minimización del riesgo empírico. Cuando la etiqueta y_n es un valor real, un elección popular de la clase de funciones para los predictores es el conjunto de todas las funciones afines. Aquí definimos una notación más compacta para representar una función afín, haciendo $x_i^{(0)} = 1$

para x_i, es decir, $x_i = [1, x_{i,1}, x_{i,2}, \ldots, x_{i,d}]^\top$. El vector de vectores correspondientes es $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_d]^\top$, lo que nos permite escribir el predictor como una función lineal

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}_i. \tag{2.4}$$

Referencias

M.P. Deisenroth, A.A. Faisal, and C.S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. ISBN 9781108470049.