Regresión lineal - Parte 1
Clase 3

1 Introducción

En los problemas de regresión se busca una función f que mapea los inputs $x \in \mathbb{R}^D$ a valores de la función $f(x) \in \mathbb{R}$. Se asume un conjunto de entrenamiento x_n y observaciones con ruido $y_n = f(x_n) + \epsilon$, donde ϵ es una variable aleatoria que describe el ruido de medición o fenómenos no modelados, y lo consideraremos como distribuciones gaussianas con media cero. El objetivo es encontrar una función que modele los datos de entrenamiento y pueda generalizar la relación entre las variables independientes y la dependiente.
Encontrar una función de regresión requiere resolver una variedad de problemas, que incluyen:

- Elección del tipo de modelo y su parametrización
- Encontrar los parámetros óptimos
- Overfitting y model selection
- Relación entre la función de pérdida y los prior de los parámetros
- Modelización de la incertidumbre

2 Formulación del problema

Por el ruido en la observaciones, se adoptará una perspectiva probabilística, explicitando en el modelo el ruido con una función de verosimilitud.

$$p(y|x) = \mathcal{N}(y|f(x), \sigma^2) \tag{1}$$

Donde $x \in \mathbb{R}^D$ son los inputs e $y \in \mathbb{R}$ son los valores de la función (objetivos) con ruido. Considerando lo anterior, la relación funcional entre x e y queda:

$$y = f(x) + \epsilon \tag{2}$$

Donde $\epsilon \sim \mathcal{N}(0, \sigma^2)$ son variables de ruido i.i.d. El objetivo es encontrar una función similar a f que generó los datos y que generalice bien.
Nos enfocamos en modelos paramétricos, es decir, aquellos que dependen de parámetros θ y se encuentran aquellos θ^* que "funcionen bien" para modelar los datos. Se asume σ^2 conocida. En regresión lineal se considera el caso especial en que los parámetros aparecen linealmente en el modelo. Un ejemplo es

$$p(y|x, \theta) = \mathcal{N}(y|x^T, \theta) \Leftrightarrow y = x^T\theta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

donde $\theta \in \mathbb{R}^D$ son los parámetros que se buscan. Este tipo de funciones son lineas rectas que pasan por el origen. La parametrización elegida es $f(x) = x^T\theta$. La verosimilitud en 3 es la función de densidad de y evaluada en $x^T\theta$. Sin el ruido de las observaciones, la relación entre x e y sería determinística.

3 Estimación de parámetros

Considerando la formulación anterior y un conjunto de entrenamiento $\mathcal{D} := \{(x_1, y_1), \cdots, (x_N, y_N)\}$ que consiste en N inputs $x_n \in \mathbb{R}^D$ y sus respectivos targets $y_n \in \mathbb{R}$, $n = 1, \cdots, N$. Dados los inputs x_i, x_j, las variables y_i, y_j son condicionalmente independientes, quedando la función de verosimilitud:

$$p(\mathcal{Y}|\mathcal{X}, \theta) = p(y_1, \cdots, y_N|x_1, \cdots, x_N, \theta) \tag{4}$$

$$= \prod_{n=1}^{N} p(y_n|x_n, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n|x_n^T\theta, \sigma^2) \tag{5}$$

Donde $\mathcal{Y}$ y $\mathcal{X}$ representan los conjuntos de objetivos e inputs de entrenamiento, respectivamente.

3.1 Estimador de máxima verosimilitud

El método de máxima verosimilitud o *maximum likelihood* es muy utilizado para optimizar parámetros. Intuitivamente, maximizar la verosimilitud significa maximizar la distribución predictora de los datos de entrenamiento, dados los parámetros, matemáticamente:

$$\theta_{ML} = \underset{\theta}{argmax}\, p(\mathcal{Y}|\mathcal{X}, \theta) \tag{6}$$

Para encontrar los parámetros óptimos, se minimiza el opuesto del logaritmo de la verosimilitud (*negative log-likelihood).*

$$-\log p(\mathcal{Y}|\mathcal{X}, \theta) = -\log \prod_{n=1}^{N} p(y_n|x_n, \theta) = -\sum_{n=1}^{N} \log p(y_n|x_n, \theta) \tag{7}$$

Como el ruido es Gaussiano, se tiene que:

$$\log p(y_n|x_n, \theta) = -\frac{1}{2\sigma^2}(y_n - x_n^T\theta)^2 + \mathcal{C} \tag{8}$$

donde $\mathcal{C}$ incluye términos que no dependen de θ. Al reemplazar esta expresión, e ignorando los términos constantes, queda:

$$\mathcal{L}(\theta) := \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^T\theta)^2 \tag{9}$$

$$= \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) = \frac{1}{2\sigma^2}||y - X\theta||^2 \tag{10}$$

Donde la matriz $X := [x_1, \cdots, x_N]^T \in \mathbb{R}^{N\times D}$ son los datos de entrenamiento y el vector $y := [y_1, \cdots, y_N]^T \in \mathbb{R}^N$ los targets. Derivando respecto de los parámetros se tiene:

$$\frac{d\mathcal{L}}{d\theta} = \frac{d}{d\theta}\Big(\frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta)\Big) \tag{11a}$$

$$= \frac{1}{2\sigma^2}\frac{d}{d\theta}\Big(y^T y - 2y^T X\theta + \theta^T X^T X\theta\Big) \tag{11b}$$

$$= \frac{1}{\sigma^2}\Big(-y^T X + \theta^T X^T X\Big) \in \mathbb{R}^{1\times D} \tag{11c}$$

Los parámetros óptimos θ_{ML} resuelven $\frac{d\mathcal{L}}{d\theta} = 0^T$, obteniéndose:

$$\frac{d\mathcal{L}}{d\theta} = 0^T \Leftrightarrow \theta_{ML}^T X^T X = y^T X \tag{12a}$$

$$\Leftrightarrow \theta_{ML}^T = y^T X(X^T X)^{-1} \tag{12b}$$

$$\Leftrightarrow \theta_{ML} = (X^T X)^{-1} X^T y \tag{12c}$$

3.2 Estimación MAP

En el caso de ML no se asume nada sobre los parámetros antes de buscar los valores óptimos. MAP propone asumir una distribución prior $p(\theta)$, que restringe los valores que pueden tomar, antes de haber visto los datos. Por ejemplo, un prior gaussiano $p(\theta) \sim \mathcal{N}(0, 1)$ confina θ al intervalo $[-2, 2]$ con alta probabilidad. Una vez que se tiene el dataset, en lugar de maximizar la verosimilitud, se maximiza el posterior de $p(\theta|\mathcal{X}, \mathcal{Y})$, que aplicando Bayes queda:

$$p(\theta|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \theta)p(\theta)}{p(\mathcal{Y}|\mathcal{X})} \tag{13}$$

A partir del cual, optimizando respecto de los parámetros, se llega a:

$$\theta_{MAP} = \Big(X^T X + \frac{\sigma^2}{b^2} I\Big)^{-1} X^T y$$

Donde σ^2 es la varianza del error y b^2 del prior de los parámetros.

3.3 Maxima verosimilitud como proyección ortogonal

Considerando un caso sencillo de regresión lineal, donde $f : \mathbb{R} \to \mathbb{R}$ pasa por el origen. El parámetro θ determina la pendiente de la línea.

$$y = x\theta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Considerando un conjunto de datos unidimensional $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ se tiene el estimador óptimo:

$$\theta_{ML} = (X^T X)^{-1} XTy = \frac{X^T y}{X^T X} \in \mathbb{R} \tag{15}$$

Con $X = [x_1, \cdots, x_N] \in \mathbb{R}^N, y = [y_1, \cdots, y_N] \in \mathbb{R}^N$ Lo que significa que de los inputs de entrenamiento X se obtienen las reconstrucciones óptimas de los objetivos de entrenamiento de la siguiente manera:

$$X\theta_{ML} = X(X^T X)^{-1} XTy = \frac{XX^T}{X^T X} y \tag{16}$$

De donde se ve que θ_{ML} hace una proyección ortogonal de y sobre el subespacio de unidimensional generado por X, siendo $\frac{XX^T}{X^T X}$ la matriz de proyección, θ_{ML} las coordenadas de la proyección del subespacio de $\mathbb{R}^N$ generado por X y $X\theta_{ML}$ como la proyección ortogonal de y en ese subespacio, como se muestra en las siguientes figuras.