Theory of Machine Learning

Exercise sheet 3 —Session 3

Exercise I (simple linear regression) ✎. Consider real-valued inputs and outputs ($\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$) with $X := (X_1, \dots, X_n)^\top \in \mathbb{R}^n$ the input vector and $Y := (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ the response vector. Let $\varphi(x) = (1, x)^\top$ and $\Phi \in \mathbb{R}^{n \times 2}$ the matrix of inputs with row i defined as $\Phi_{i,:} := \phi(X_i)^\top$. We set

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \quad \overline{XY} := \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\,, \quad \text{and} \quad \overline{X^2} := \frac{1}{n}\sum_{i=1}^{n} X_i^2\,.$$

1. Give the expression of $\Phi^\top\Phi$ using these notation.
2. Under which conditions is this matrix invertible?
3. Assume that $\Phi^\top\Phi$ is invertible. We want to minimize the empirical risk of $\theta \in \mathbb{R}^2$ defined as $\hat{\mathcal{R}}(\theta) = \frac{1}{n}\|Y - \Phi\theta\|^2$. Given that $\hat{\theta} := (\Phi^\top\Phi)^{-1}\Phi^\top Y$, express $\hat{\theta}$ in this specific case.

Exercise II (coding simple linear regression) 💻. The objective is to implement the previous exercise using numpy.

1. Generate a dataset $\{X_i, Y_i\}_{i=1}^n$ as follows:
   (a) Sample the $n \in \mathbb{N}^\star$ inputs as $X_i \sim \mathcal{U}([-5, 5])$ (*Hint: use* `numpy.random.uniform()`)
   (b) The labels are $Y_i = X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1)$ (*Hint: use* `numpy.random.normal()`)
2. Implement the least-square estimator $\hat{\theta}$ of exercise I as follows:
   (a) First construct the matrix of inputs $\Phi \in \mathbb{R}^{n \times 2}$ using the previously sampled inputs X_i.
   (b) Compute $\Phi^\top\Phi$.
   (c) Compute $\hat{\theta} := (\Phi^\top\Phi)^{-1}\Phi^\top Y$, where Y is the label vector sampled previously.

Exercise III (maximum likelihood estimation) ✎. In the fixed design setting, we can make further assumptions on the noise, for instance let us assume that the ε_is are i.i.d. Gaussian, with mean zero and variance σ^2. A typical approach in statistics is to look at the *likelihood* of observations, defined as the product of densities at the observations, then to maximize it with respect to the parameters.

1. What is the density of the random variable Y_i in this setting?
2. Write $\mathcal{L}(Y|\theta, \sigma^2)$ the product of densities evaluated at $Y_1, \dots, Y_n$.
3. Find $\tilde{\theta}$ which maximizes $\mathcal{L}(Y|\theta, \sigma^2)$. What is his relationship to $\hat{\theta}$? *(Hint: maximizing $\mathcal{L}$ is equivalent to maximizing* $\log \mathcal{L}$*)*