Theory of Machine Learning

Exercise sheet 2 — Session 2

Exercise I (Estimation of true risk) 💻. Assume that $\mathcal{X} := \mathbb{R}$ and $\mathcal{Y} := \{0, 1\}$. Let the space of predictors $\mathcal{H} := \{f_t : \mathcal{X} \to \mathcal{Y} \text{ s.t. } \forall x \in \mathbb{R}, f(x) = \mathbb{1}_{x \geq t}\}$.

We want to estimate numerically the expected risk of a given predictor $\mathcal{R}(f_t)$, where the data is normally distributed $X \sim \mathcal{N}(0, 1)$ and the label is $Y := \mathbb{1}_{X \geq 0}$.

1. Sample $N \in \mathbb{N}^\star$ data points from $\mathcal{N}(0, 1)$ using `numpy.random.normal()`.

2. Given $t \in \mathbb{R}$, compute the Monte-Carlo estimation of $\mathcal{R}(f_t)$ using the previously sampled points. (Hint: empirical risk)

3. Plot the Monte-Carlo estimates of $\mathcal{R}(f_t)$ for various values of t as the number of points N increases. Generate a separate curve for each value of t and display all curves on the same figure for comparison.

Exercise II (Closed-form of true risk) ✎. Assume that $\mathcal{X} := \mathbb{R}$ and $\mathcal{Y} := \mathbb{R}$. We want to compute the expected risk of the identity predictor $g(x) := x$, for $x \in \mathbb{R}$, using the squared loss function. The expected risk $\mathcal{R}(g)$ is defined as $\mathcal{R}(g) := \mathbb{E}_X\left[(g(X) - f^*(X))^2\right]$, where the data is normally distributed $X \sim \mathcal{N}\left(0, \sigma^2\right)$ $(\sigma > 0)$ and the labels are determined by a fixed function $f^*(x) = 0$, for $x \in \mathbb{R}$.

1. Show that
$$\mathcal{R}(g) = \mathbb{E}_X\left[X^2\right] .$$

2. Prove the following formula:
$$\mathrm{Var}\,(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 ,$$
where X is a random variable with finite second moment and $\mathrm{Var}\,(X) := \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]$.

3. Compute $\mathcal{R}(g)$ using Question 2.

4. How does the variance σ^2 influence the expected risk $\mathcal{R}(g)$?

Exercise III (Consistency and bias of empirical risk) ✎. Consider a fixed predictor $f : \mathcal{X} \to \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Given an *i.i.d.* (independent and identically distributed) data sample $S := \{X_i\}_{i=1}^N \sim p$ of size $N \in \mathbb{N}^\star$, we remind the definition of the empirical risk

$$\hat{R}_S(f) := \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), f^*(X_i)) ,$$

where $\mathbb{E}\left[|\ell(f(X_1), c(X_1))|\right] < +\infty$ and the labels areß determined by a fixed function $f^* : \mathcal{X} \to \mathcal{Y}$.

1. Show that
$$\mathbb{E}_S\left[\hat{R}_S(f)\right] = R(f) .$$

2. Show that the empirical risk $\hat{R}_S(f)$ converges (in probability) to the expected risk $R(f)$ using the law of large numbers.

3. Which version of the law of large numbers did you use?

Exercise IV (Bayes predictor for binary classification) ✎. In this exercise, we compute the expression of the Bayes predictor for binary classification with $0-1$ loss ($\mathcal{Y} = \{0, 1\}$). As in the lecture, we set $\eta(x) := \mathbb{P}(Y = 1 \,|\, X = x)$ for all $x \in \mathcal{X}$ and we let

$$f^\star(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

1. Let $f : \mathcal{X} \to \mathcal{Y}$ be a predictor. Show that

$$\mathbb{P}(f(X) \neq Y \,|\, X = x) = \eta(x) \cdot \mathbb{P}(f(X) = 0 \,|\, X = x) + (1 - \eta(x)) \cdot \mathbb{P}(f(X) = 1 \,|\, X = x) \,.$$

2. Deduce that

$$\mathbb{P}(f^\star(X) \neq Y \,|\, X = x) = \min(\eta(x), 1 - \eta(x)) \,.$$

3. Show that, for any predictor $f : \mathcal{X} \to \mathcal{Y}$,

$$\mathbb{P}(f(X) \neq Y \,|\, X = x) \geq \mathbb{P}(f^\star(X) \neq Y \,|\, X = x) \,.$$

4. Deduce that $f^\star$ is risk optimal, that is, for any predictor f,

$$\mathcal{R}(f^\star) \leq \mathcal{R}(f) \,.$$