The Mathematics of Machine Learning
Homework Set 1

Due 23 February 2023 before 13:00
via Canvas

You are allowed to work on this homework in pairs. One person per pair submits the answers via Canvas. Make sure to put both names on the submission.

1 Theory Exercises

1. [4 pt]

   (a) [2 pt] What is the Bayes-optimal predictor f_{B} for binary classification with $Y \in \{-1, +1\}$ and the following cost-sensitive loss, which considers a false negative worse than a false positive?

$$\ell(Y, \hat{Y}) = \begin{cases} 0 & \text{if } \hat{Y} = Y, \\ 1 & \text{if } Y = -1 \text{ and } \hat{Y} = +1, \\ 10 & \text{if } Y = +1 \text{ and } \hat{Y} = -1. \end{cases}$$

   (b) [2 pt] For least-squares regression with the absolute error loss[1],

$$\ell(Y, \hat{Y}) = |Y - \hat{Y}|,$$

   the Bayes optimal predictor is such that $f_{\mathrm{B}}(X)$ is any median of Y under $P^*(Y|X)$. This follows from the following lemma:

   Lemma 1. *For any random variable Y with distribution P,*

$$\mathbb{E}[|Y - c|]$$

   is minimized in c by any median of P.

   Prove this lemma. You may use without proof that at least one median m always exists.
   Hint 1: The median of any distribution P is any point m such that

$$P(Y \leq m) \geq \frac{1}{2} \qquad \text{and} \qquad P(Y \geq m) \geq \frac{1}{2}.$$

[1] NB This is a common alternative to the squared error loss $\ell(Y, \hat{Y}) = (Y - \hat{Y})^2$ that we considered in the lecture. The absolute error is less sensitive to large errors, which may be an advantage if there may be outliers (extreme points with small probability).

Hint 2: By symmetry, it is sufficient to show that if $c < m$ and m is a median of P, then

$$\mathbb{E}[|Y - c|] \geq \mathbb{E}[|Y - m|].$$

(You do not have to prove this.)
Hint 3: Let $\mathbf{1}\{A\}$ be the indicator for any event A, which is 1 if A holds and 0 otherwise. Show that, if $c < m$, then

$$\begin{aligned}\mathbb{E}[|Y - c|] - \mathbb{E}[|Y - m|] = \mathbb{E}[(c - m)\mathbf{1}\{Y \leq c\}] \\ + \mathbb{E}[(2Y - m - c)\mathbf{1}\{c < Y < m\}] \\ + \mathbb{E}[(m - c)\mathbf{1}\{Y \geq m\}],\end{aligned}$$

and find a simpler lower bound on this expression using that $Y \geq c$ in the middle case.
Hint 4: Use the properties of the median to show that the lower bound from Hint 3 is non-negative.

2 Programming Exercise

The following programming exercise is to be implemented in Python, using a Jupyter notebook. As a starting point, you may use the notebook `Homework1-start.ipynb`, which is available from the course website.

2. [8 pt]

   (a) Simulate a training set of size $N = 200$ and a test set of size 10 000 by sampling from the following binary classification distribution with $X \in \mathbb{R}^2$ and $Y \in \{-1, +1\}$:

      i. Sample a Bernoulli random variable $Z \in \{-1, +1\}$ such that $\Pr(Z = 1) = 3/4$. The interpretation is that Z represents the unobserved true class.

      ii. Set $\mu_Z = \begin{pmatrix} +Z \\ -Z \end{pmatrix}$ and sample X from a normal distribution $\mathcal{N}(\mu_Z, I)$.

      iii. Sample Y such that $\Pr(Y = Z) = 4/5$, so we only observe a noisy label that might differ from Z.

   (b) Plot the training set in 2 dimensions, using different symbols or colors for the two classes.

   (c) Similar to Figure 2.4 in the book, plot the error = average 0/1-loss both on the training set and on the test set for the K-nearest neighbour classifier as a function of $K = N, \ldots, 1$.

   (d) Derive a way to compute the Bayes-optimal classifier f_{B} for the 0/1-loss, and add its errors on the training set and on the test set as horizontal lines to the plot.

Hint to calculate f_{B}: Let $g(x, y, z) = \Pr(Z = z) \Pr(Y = y|Z = z)\phi(x; \mu_z, I)$ denote the joint density of X, Y and Z, where $\phi(x; \mu, \Sigma)$ is the density of a multivariate Gaussian with mean μ and covariance matrix Σ. We need to determine whether $\Pr(Y = +1|X) \geq \Pr(Y = -1|X)$, which is equivalent to

$$\sum_{z \in \{-1,+1\}} g(x, +1, z) \geq \sum_{z \in \{-1,+1\}} g(x, -1, z).$$