Exercises for the course

# Machine Learning 2

Summer semester 2024

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

## Exercise Sheet 0

### Exercise 1: Maximum Likelihood vs. Bayes

An unfair coin is tossed seven times and the event (head or tail) is recorded at each iteration. The observed sequence of events is

$$\mathcal{D} = (x_1, x_2, \dots, x_7) = (\text{head}, \text{head}, \text{tail}, \text{tail}, \text{head}, \text{head}, \text{head}).$$

We assume that all tosses $x_1, x_2, \dots$ have been generated independently following the Bernoulli probability distribution

$$P(x \mid \theta) = \begin{cases} \theta & \text{if} \quad x = \text{head} \\ 1-\theta & \text{if} \quad x = \text{tail}, \end{cases}$$

where $\theta \in [0, 1]$ is an unknown parameter.

(a) *State* the likelihood function $P(\mathcal{D}|\theta)$, that depends on the parameter $\theta$.

(b) *Compute* the maximum likelihood solution $\hat{\theta}$, and *evaluate* for this parameter the probability that the next two tosses are "head", that is, evaluate

$$P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \hat{\theta}).$$

(c) We now adopt a Bayesian view on this problem, where we assume a prior distribution for the parameter $\theta$ defined as:

$$p(\theta) = \begin{cases} 1 & \text{if} \quad 0 \leq \theta \leq 1 \\ 0 & \text{else.} \end{cases}$$

*Compute* the posterior distribution $p(\theta|\mathcal{D})$, and *evaluate* the probability that the next two tosses are head, that is,

$$\int P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \theta)\, p(\theta|\mathcal{D})\, d\theta.$$

### Exercise 2: Principal Component Analysis

We consider an unsupervised dataset $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$, where $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{k=1}^N \boldsymbol{x}_k$ is the empirical mean. The principal component analysis problem consists of finding the vector $\boldsymbol{e} \in \mathbb{R}^d$ of norm 1 such that the data projected in this space has maximum variance, i.e. is a solution of the optimization problem

$$\max_{\boldsymbol{e} \in \mathbb{R}^d} \frac{1}{N} \sum_{k=1}^N (\boldsymbol{e}^\top \boldsymbol{x}_k - m)^2 \qquad \text{subject to} \quad \|\boldsymbol{e}\|^2 = 1$$

where $m = \frac{1}{N}\sum_{k=1}^N \boldsymbol{e}^\top \boldsymbol{x}_k$ is the mean of the projected data.

(a) *Show* that the problem can be rewritten as the quadratic program

$$\max_{\boldsymbol{e} \in \mathbb{R}^d} \boldsymbol{e}^\top C \boldsymbol{e} \qquad \text{subject to} \quad \|\boldsymbol{e}\|^2 = 1$$

where $C = \frac{1}{N}\sum_{k=1}^N (\boldsymbol{x}_k - \bar{\boldsymbol{x}}) \cdot (\boldsymbol{x}_k - \bar{\boldsymbol{x}})^\top$ is the empirical covariance matrix.

(b) *Show* using the method of Lagrange multipliers that the solution of the optimization problem above is an eigenvector of the matrix $C$.

(c) *Show* that, among all possible eigenvectors of $C$, the solution of the optimization problem above is the one with highest associated eigenvalue.

**Exercise 3: Neural Networks**

We consider a neural network that takes two inputs $x_1$ and $x_2$ and produces an output $y$ based on the following set of computations:

$$
\begin{array}{lll}
z_3 = x_1 \cdot w_{13} + x_2 \cdot w_{23} & z_5 = a_3 \cdot w_{35} + a_4 \cdot w_{45} & y = a_5 + a_6 \\
a_3 = \tanh(z_3) & a_5 = \tanh(z_5) & \\
z_4 = x_1 \cdot w_{14} + x_2 \cdot w_{24} & z_6 = a_3 \cdot w_{36} + a_4 \cdot w_{46} & \\
a_4 = \tanh(z_4) & a_6 = \tanh(z_6) &
\end{array}
$$

(a) *Draw* the neural network graph associated to this set of computations.

(b) *Write* the set of backward computations that leads to the evaluation of the partial derivative $\partial y / \partial w_{13}$. Your answer should avoid redundant computations. Hint: $\tanh'(t) = 1 - (\tanh(t))^2$.

**Exercise 4: Support Vector Machines**

The primal program for the linear hard margin SVM is

$$\min_{\boldsymbol{w},\theta} \|\boldsymbol{w}\|^2 \qquad \text{subject to} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + \theta) \geq 1, \quad \text{for } 1 \leq i \leq N,$$

where $\|.\|$ denotes the Euclidean norm, and the minimization is performed in $\boldsymbol{w} \in \mathbb{R}^d, \theta \in \mathbb{R}$, while the data $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ are regarded as fixed constants.

(a) *State* the Lagrangian dual of the constrained optimization problem above and determine when the Slater's conditions for strong duality are satisfied.

(b) *Show* that the Lagrange dual takes the form of a quadratic optimization problem w.r.t. the dual variables $\alpha_1, \ldots, \alpha_N$.

**Exercise 5: Kernels**

A kernel function $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ generalizes the linear scalar product between two vectors. The kernel must satisfy positive semi-definiteness, that is, for any sequence of data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and coefficients $c_1, \ldots, c_n \in \mathbb{R}$ the following inequality should hold:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$$

We consider the kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2$.

(a) *Show* that this kernel is positive semi-definite.

(b) *Show* that this kernel can be rewritten as a dot product $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}') \rangle$.