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

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

We assume that all tosses $x_1, x_2, \ldots$ 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 θ.

$$P(D|\theta) = \prod_{i=1}^{7} P(x_i|\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta$$
$$= \theta^{\#\text{heads}} \cdot (1-\theta)^{\#\text{tails}}$$
$$= \theta^5 \cdot (1-\theta)^2$$

(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

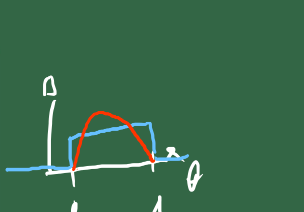
$$P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \hat{\theta}).$$

$$\log P(D|\theta) = 5 \cdot \log \theta + 2 \log(1-\theta) \qquad \text{(concave)}$$
$$\frac{\partial}{\partial \theta} \log P(D|\theta) = \frac{5}{\theta} - \frac{2}{1-\theta} \overset{!}{=} 0$$
$$\Rightarrow \hat{\theta} = \frac{5}{7}$$

$$P(x_8 = \text{head}, x_9 = \text{head}) = \hat{\theta} \cdot \hat{\theta} = \frac{25}{49}$$

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

$$p(\theta) = \begin{cases} 1 & \text{if} \quad 0 \le \theta \le 1 \\ 0 & \text{else.} \end{cases} \qquad \int p(\theta)\, d\theta = 1 \quad p(\theta) \ge 0$$



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

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

$$p(\theta|D) = \frac{P(D|\theta)p(\theta)}{\int P(D|\theta)p(\theta)d\theta} = \frac{\theta^5 \cdot (1-\theta)^2}{\int_0^1 \theta^5 (1-\theta)^2 d\theta} = 168 \cdot \theta^5 \cdot (1-\theta)^2$$

$$\int_0^1 \theta^5(1-\theta)^2 d\theta = \frac{1}{168}$$

$$\int P(x_8 = \text{head}, x_9 = \text{head} | \theta) \cdot 168 \cdot \theta^5 (1-\theta)^2 d\theta$$
$$= \int_0^1 \theta^2 \cdot 168 \cdot \theta^5 (1-\theta)^2 d\theta = \frac{7}{15}$$

Exercise 2: Principal Component Analysis

We consider an unsupervised dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$, 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.

$$\frac{1}{N}\sum_\ell e^T x_\ell = e^T\left(\frac{1}{N}\sum_\ell x_\ell\right) = e^T \bar{x}$$

$$\frac{1}{N}\sum_k \left(e^T x_k - \frac{1}{N}\sum_\ell e^T x_\ell\right)^2 =$$
$$= \frac{1}{N}\sum_k e^T x_k x_k^T e - 2 e^T x_k \bar{x}^T e + e^T \bar{x}\bar{x}^T e$$
$$= e^T\left(\frac{1}{N}\sum_k (x_k - \bar{x})(x_k - \bar{x})^T\right) e$$
$$= e^T C e$$

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

$$\mathcal{L}(e, \lambda) = e^T C e - \lambda(\|e\|^2 - 1)$$
$$\frac{\partial \mathcal{L}}{\partial e} = 2Ce - 2\lambda e \overset{!}{=} 0$$
$$\Leftrightarrow Ce = \lambda e$$

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

$$e^T C e = \lambda \cdot e^T e = \lambda \|e\|^2 = \lambda$$

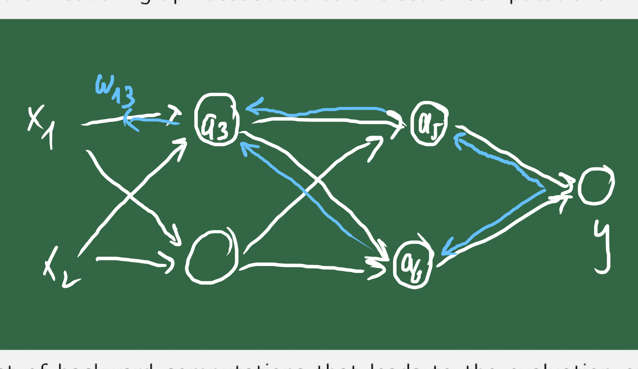
Exercise 3: Neural Networks

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:

$$z_3 = x_1 \cdot w_{13} + x_2 \cdot w_{23} \qquad z_5 = a_3 \cdot w_{35} + a_4 \cdot w_{45} \qquad y = a_5 + a_6$$
$$a_3 = \tanh(z_3) \qquad a_5 = \tanh(z_5)$$

$$z_4 = x_1 \cdot w_{14} + x_2 \cdot w_{24} \qquad z_6 = a_3 \cdot w_{36} + a_4 \cdot w_{46}$$
$$a_4 = \tanh(z_4) \qquad a_6 = \tanh(z_6)$$

(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: $\tanh'(t) = 1 - (\tanh(t))^2$.

$$\delta_5 = 1 - a_5^2 = \frac{\partial y}{\partial z_5}$$
$$\delta_6 = 1 - a_6^2$$
$$\delta_3 = (w_{35}\,\delta_5 + w_{36}\,\delta_6) \cdot (1 - a_3) = \frac{\partial y}{\partial z_3}$$
$$\frac{\partial y}{\partial w_{13}} = x_1 \delta_3$$

Exercise 4: Support Vector Machines

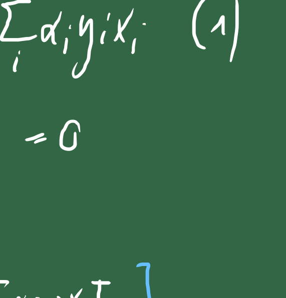
The primal program for the linear hard margin SVM is

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

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.

$$\mathcal{L}(w, \theta, \alpha_1, \alpha_2, \ldots, \alpha_N) = \frac{1}{2}\|w\|^2 + \sum_i \alpha_i (1 - y_i(w^T x_i + \theta)) \qquad (2)$$
$$\max_\alpha \min_{w, \theta} \mathcal{L}(w, \theta, \vec{\alpha}) \quad \text{s.t. } \alpha_i \ge 0 \ \forall i$$
$$\exists (w, \theta)\ \forall i\ y_i(w^T x_i + \theta) > 1 \Rightarrow \text{strong duality}$$



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

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \alpha_i y_i x_i \overset{!}{=} 0 \Rightarrow w = \sum_i \alpha_i y_i x_i \quad (1)$$
$$\frac{\partial \mathcal{L}}{\partial \theta} = -\sum_i \alpha_i y_i \overset{!}{=} 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

(1) into (2):
$$\mathcal{L}(\theta, \alpha) = \frac{1}{2}\left[\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j\right] + \sum_i \alpha_i - \left[\sum_i \alpha_i y_i \sum_j \alpha_j y_j x_j^T x_i\right]$$
$$\Rightarrow \mathcal{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$\text{s.t. } \forall i,\ \alpha_i \ge 0$$
$$\sum_i \alpha_i y_i = 0$$

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:

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

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.

$$\sum_i \sum_j c_i c_j \langle x_i, x_j \rangle^2 = \sum_i \sum_j c_i c_j \left(\sum_{k=1}^d x_{ik} \cdot x_{jk}\right)^2$$
$$= \sum_i \sum_j c_i c_j \left(\sum_k x_{ik} x_{jk}\right)\left(\sum_\ell x_{i\ell} x_{j\ell}\right)$$
$$= \sum_i \sum_j \sum_k \sum_\ell c_i c_j x_{ik} x_{jk} x_{i\ell} x_{j\ell}$$
$$= \sum_i \sum_j \sum_k \sum_\ell c_i x_{ik} x_{i\ell} \cdot c_j x_{jk} x_{j\ell}$$
$$= \sum_k \sum_\ell \left(\sum_i c_i x_{ik} x_{i\ell}\right)\left(\sum_j c_j x_{jk} x_{j\ell}\right) = \sum_k \sum_\ell \left(\sum_i c_i x_{ik} x_{i\ell}\right)^2 \ge 0$$

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