Dr. Wigand Rathmann, Jan Krause, Ehsan Waiezi Winter term 2022/23, 01.03.2023

# Exam
# Mathematics of Learning

## Question 1: Principal Component Analysis (8 points)

Let input data $x^1 = \begin{pmatrix} 2 \\ 7 \end{pmatrix}$, $x^2 = \begin{pmatrix} -6 \\ 3 \end{pmatrix}$, $x^3 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$, $x^4 = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$ be given.
Compute for all their respective centered data points the first principal component, i.e., the principal component with the largest eigenvalue.

*Hint:* Use $\frac{1}{N}$ as a factor in the formula for the covariance matrix computation.

## Question 2: K-Means Clustering (4 + 2 + 1 = 7 points)

a) (4 points)

Consider data $X := \left\{ \begin{pmatrix} -2 \\ 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} \begin{pmatrix} -2 \\ -1 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \end{pmatrix} \begin{pmatrix} 4 \\ -1 \end{pmatrix} \right\}$ and initial cluster means $m_1 = \begin{pmatrix} -4 \\ 0 \end{pmatrix}$ and $m_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Calculate two iterations of the 2-means algorithm. Give all clusters and corresponding cluster means as result.
*Hint:* You can use graphical figures to save some calculations.

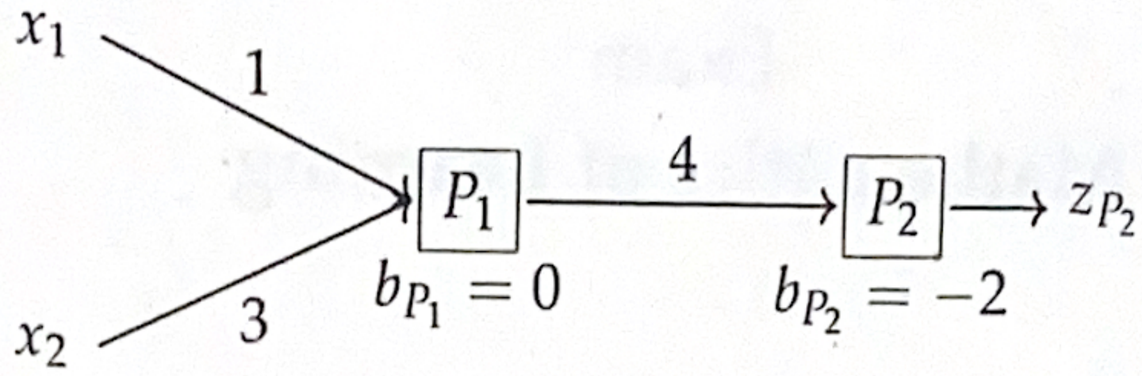
b) (2 points)

State the formula for the *clustering energy*. Furthermore, give a counter example to disprove the following statement: The $k$-means algorithm always terminates with a global minimum of the clustering energy.

c) (1 point)

Prove the following statement: If the set of data points is linearly independent, it is not possible that the $k$-means algorithm returns two non-empty clusters with the same mean.

**Question 3: Neural Network** (4 + 2 + 2 = 8 points)
Consider the following network with 2 neurons $P_1$, $P_2$




with initial weights (as denoted in the graph)

$$w_{P_1x_1} = 1,\ w_{P_1x_2} = 3,\ w_{P_2P_1} = 4,$$

initial biases (as denoted in the graph) $b_{P_1} = 0,\ b_{P_2} = -2$, and activation functions

$$\psi_{P_1}(t) = \frac{1}{1+3^{-t}},\ \psi_{P_2}(t) = t^2.$$

Let

$$\theta = (w_{P_1x_1}, w_{P_1x_2}, w_{P_2P_1}, b_{P_1}, b_{P_2})$$

and let $f_\theta(x) = z_{P_2} \in \mathbb{R}$ denote the output of the network using parameters $\theta$ and input $x = (x_1, x_2)^T \in \mathbb{R}^2$. Consider the loss function $C(\theta; x, y) = \frac{1}{2}||f_\theta(x) - y||^2$ for a given training pair $(x, y)$.

a) (4 points)

Perform one forward pass ~~and compute the loss~~ for $x = \begin{pmatrix} -5 \\ 2 \end{pmatrix}$. Furthermore, explain the idea and purpose of backpropagation.

b) (2 points)
Now, perform one update step for the weights and biases following the Stochastic Gradient Descent algorithm for a batch that contains the training samples $(x_1, y_1)$ and $(x_2, y_2)$. Their gradients with respect to $\theta$ are given by

$$\nabla C(\theta; x_1, y_1) = (3.0, 1.0, -1.6, -0.8, -1.0)^T$$

and

$$\nabla C(\theta; x_2, y_2) = (-1.0, 0.6, 4.0, -2.6, -1.4)^T,$$

and the step-size is fixed to $\eta = 0.1$.

c) (2 points)
Consider the common activation function *Tangens hyperbolicus*:

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1.$$

Show that we have the following formula for the derivative:

$$\tanh'(x) = 1 - \tanh(x)^2.$$

## Question 4: Algorithmic Strategies (1 + 2 + 2 = 5 points)

a) (1 point)
State two possibilities to reduce the generalization error of a neural network.

b) (2 points)
Let $N, p \in \mathbb{N}$, labeled data points $(x_1, y_1), ..., (x_N, y_N) \in \mathbb{R}^p \times \{-1, 1\}$ and a penalty parameter $C \in \mathbb{R}_{>0}$ be given. A version of the SVM classification optimization problem is the following:

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}, \xi \geq 0} ||\beta||_2^2 + C \sum_{i=1}^{N} \xi_i,$$
$$\text{s.t. } \xi_i \geq 1 - y_i(x_i^T \beta + \beta_0).$$

We denote by $\hat{\beta}, \hat{\beta}_0, \hat{\xi}$ an optimal solution of the optimization problem.

Consider the following setting with scaled data: we replace $x_i$ by $\lambda x_i$ for all $i = 1, ..., N$ and the penalty parameter $C$ by $\frac{C}{\lambda^2}$.

Show that $\frac{\hat{\beta}}{\lambda}, \hat{\beta}_0, \hat{\xi}$ solves the resulting SVM classification optimization problem.

c) (2 points)
Let $N, p \in \mathbb{N}$ and $X \in \mathbb{R}^{N \times p}$ (not necessarily with full column rank), and $0 \neq Y \in \mathbb{R}^N$. Then the corresponding linear regression problem

$$\min_{\beta \in \mathbb{R}^p} ||X\beta - Y||_2^2$$

minimizes the squared error.

Prove the following statement:
If a column of $X$ can be expressed as a linear combination of other columns of $X$, then the column can be removed without changing the minimal squared error.