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

Exam
Mathematics of Learning
Solution sketches

Name ....

Student registration no. ....
(Matrikelnummer)

Signature ....

- **Do not turn over this page until instructed to do so by the examiner!**

- The time for completing the exam is **60 min**.
- The exam consists of 4 questions with a total of 28 points.
- Answers have to be readable and justified.
- Write in black or blue. Only in figures, colors are allowed.
- As auxiliary tool you may use one handwritten sheet of paper (DinA4, both sides). There are no other tools allowed (no books, no calculator, no phone . . .).

Good luck!

Q1 (8P)	Q2 (7P)	Q3 (8P)	Q4 (5P)	$\sum = 28$P

Grade:

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.

Solution Question 1:

1. (1P) Compute mean value
$$\overline{X} = \frac{1}{4}\sum_{i=1}^{4} x^i = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$$

2. (1P) Center data $y^i = x^i - \begin{pmatrix} -2 \\ 3 \end{pmatrix}$.

3. (2P) Compute covariance matrix
$$C = \frac{1}{4}\sum_{i=1}^{4} y^i (y^i)^T = \frac{1}{4}\left(\begin{pmatrix} 16 & 16 \\ 16 & 16 \end{pmatrix} + \begin{pmatrix} 16 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 16 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) = \begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix}$$
(Alternative: $Y = (y^1\ y^2\ y^3\ y^4)$, $C = \frac{1}{4}YY^T$)

4. (2P) Compute eigenvalues and eigenvectors. First, compute the roots of the characteristic polynomial of C:
$$\chi_C(\lambda) = (\lambda - 8)(\lambda - 8) - 16 = \lambda^2 - 16\lambda + 48.$$
Using the quadratic formula we can calculate the eigenvalues $\lambda = 12$ and $\lambda = 4$. Therefore, the largest eigenvalue is $\lambda = 12$ and since we want to compute one principal component it is enough to calculate the eigenvector for $\lambda = 12$.

5. (1P) This can be done with Gaussian elimination:
$$(12 \cdot \mathbb{1} - C)v = 0 \Leftrightarrow \begin{pmatrix} 4 & -4 \\ -4 & 4 \end{pmatrix} v = 0 \Leftrightarrow \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} v = 0.$$
This implies that the eigenvector is given by $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

6. (0P) $T = \begin{pmatrix} v \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

7. (1P) Compute the first principal components for each centered data point:
$$z^1 = T^T y^1 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 4 \end{pmatrix} = 8,$$
$$z^2 = T^T y^2 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} -4 \\ 0 \end{pmatrix} = -4,$$
$$z^3 = T^T y^3 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -4 \end{pmatrix} = -4,$$
$$z^4 = T^T y^4 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

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.

Solution.

a) (4P) In Iteration 1, we get

$C_1 := \{ \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \end{pmatrix} \}$

$C_2 := \{ \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ -1 \end{pmatrix} \}$

and the cluster means are $m_1 = \{ \begin{pmatrix} -2 \\ 0 \end{pmatrix} \}$, $m_2 = \{ \begin{pmatrix} 7/3 \\ 0 \end{pmatrix} \}$,

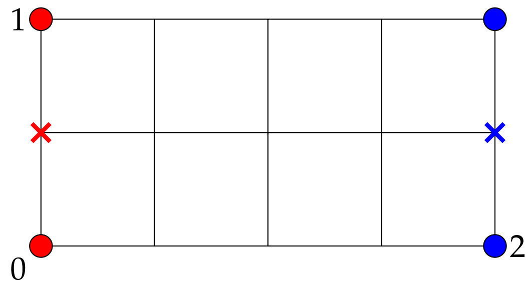
In iteration 2 we get

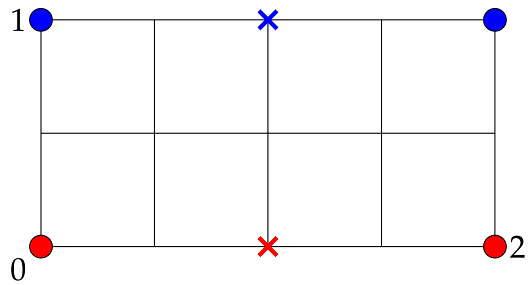
$C_1 := \{ \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \end{pmatrix} \}$

$C_2 := \{ \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ -1 \end{pmatrix} \}$

and the cluster means are $m_1 = \{ \begin{pmatrix} -5/3 \\ 0 \end{pmatrix} \}$, $m_2 = \{ \begin{pmatrix} 4 \\ 0 \end{pmatrix} \}$.

b) (2P) The clustering energy can be computed by $E(\underset{\sim}{C}, \underset{\sim}{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} ||x - m_k||^2$ with clustering $\underset{\sim}{C} := \{C_1, \ldots, C_K\}$ and centers $\underset{\sim}{m} := \{m_1, \ldots, m_K\}$. Look at the following counter example:

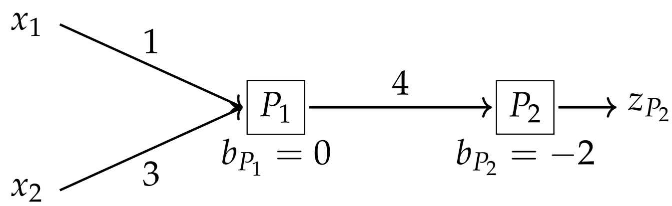




Both picture might depict a termination of 2-means, whereas the latter scenario has bigger clustering energy as the former one.

c) (1P) Linear independence implies unique linear combination of an arbitrary point of the set of vectors - so in particular cluster means. Two disjoint subsets of a linear independent set do not have the same mean, regardless of what the algorithm does.

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

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

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 θ 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 θ 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:

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

Solution Question 3:

a) (2P) We start computing the layers' outputs using a forward pass:

$$\begin{aligned}
a_{P_1} &= w_{P_1 x_1} \cdot x_1 + w_{P_1 x_2} \cdot x_2 + b_{P_1} = 1 \cdot (-5) + 3 \cdot 2 + 0 = 1 \\
z_{P_1} &= \psi_{P_1}(a_{P_1}) = \frac{1}{1+3^{-1}} = \frac{1}{1+\frac{1}{3}} = \frac{3}{4} \\
a_{P_2} &= w_{P_2 P_1} \cdot z_{P_1} + b_{P_2} = 4 \cdot \frac{3}{4} - 2 = 1 \\
z_{P_2} &= \psi_{P_2}(a_{P_2}) = 1^2 = 1
\end{aligned}$$

(2P) idea of backpropagation

b) We have to compute the averaged gradient $\overline{\nabla} C(\theta)$ (1P) and use it for the update step. The averaged gradient is given by

$$\overline{\nabla} C(\theta) = \frac{1}{2}(\nabla(\theta, x^1, y^1) + \nabla(\theta, x^2, y^2)) = \frac{1}{2}\begin{pmatrix} 2 \\ 1.6 \\ 2.4 \\ -3.4 \\ -2.4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.8 \\ 1.2 \\ -1.7 \\ -1.2 \end{pmatrix}.$$

The update is given by (1P):

$$\theta^{new} = \theta - \eta \overline{\nabla} C(\theta) = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 0 \\ -2 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 1 \\ 0.8 \\ 1.2 \\ -1.7 \\ -1.2 \end{pmatrix} = \begin{pmatrix} 0.9 \\ 2.92 \\ 3.88 \\ 0.17 \\ -1.88 \end{pmatrix}.$$

The updated parameters are therefore given by

$$w^{new}_{P_1 x_1} = 0.9 \quad w^{new}_{P_1 x_2} = 2.92 \quad w^{new}_{P_2 P_1} = 3.88$$
$$b^{new}_{P_1} = 0.17 \quad b^{new}_{P_2} = -1.88.$$

c) (2P) On the one hand, by application of quotient rule, we have: $\tanh'(x) = \frac{u'(x)v(x) - v'(x)u(x)}{(v(x))^2}$ with $u(x) = 2, v(x) = 1 + e^{-2x}$. The derivatives for the latter are $u'(x) = 0, v'(x) = -2e^{-2x}$, so $\tanh'(x) = \frac{4e^{-2x}}{(1+e^{-2x})^2}$. On the other hand we get $1 - \tanh^2(x) = 1 - (\frac{2}{1+e^{-2x}} - 1)^2 = 1 - (\frac{4}{(1+e^{-2x})^2} - \frac{4}{1+e^{-2x}} + 1) = \frac{4}{1+e^{-2x}} - \frac{4}{(1+e^{-2x})^2} = \frac{4e^{-2x}}{(1+e^{-2x})^2}$. Both together yields $\tanh'(x) = 1 - \tanh(x)^2$.

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 λx_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.

Solution: (1P) The solution stays feasible, since

$$1 - y_i(\lambda x_i^T \frac{1}{\lambda}\hat{\beta} + \hat{\beta}_0) = 1 - y_i(x_i^T \hat{\beta} + \hat{\beta}_0).$$

(1P) The solution is optimal, since the objective function evaluated at $\frac{1}{\lambda}\hat{\beta}$ is

$$\frac{1}{\lambda^2}||\hat{\beta}||_2^2 + \frac{C}{\lambda^2} \sum_{i=1}^{N} \hat{\xi}_i,$$

which is just the original objective function of the problem scaled by $\frac{1}{\lambda^2}$, and this is minimized by the original solution.

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.

Solution: (2P) True. Assume that $X_j = \sum_{k \neq j} \lambda_k X_k$, and β solve the regression problem, then $\hat{\beta}$ with $\hat{\beta}_k = \beta_k + \lambda_k \beta_j \, \forall k \neq j$ solves the regression problem (with the same value) with respect to the matrix that is obtained by deleting the $j-$th column from X. (Being always able to set β_j to zero is equivalent to be able to remove the corresponding column.)