Machine Learning — Repeat Exam

1	2	3	4	5	6	7	8	9	10	11	$\sum$
4	5	6	4	6	5	4	7	4	2	7	54

Do not write anything above this line

Name:

Student ID: Signature:

- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Pages 16-18 can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Do not unstaple the sheets!**
- Wherever answer boxes are provided, please write your answers in them.
- Please write your student ID (*Matrikelnummer*) on every sheet you hand in.
- **Only use a black or a blue pen (no pencils, red or green pens!).**
- You are allowed to use your A4 sheet of handwritten notes (two sides). **No other materials (e.g. books, cell phones, calculators) are allowed!**
- Exam duration - 120 minutes.
- This exam consists of 19 pages, 11 problems. You can earn 54 points.

Probability distributions

For your reference, we provide the following probability distribution.

- Univariate normal distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Bernoulli distribution

$$\mathrm{Bern}(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Decision Trees

Problem 1 [(2+2)=4 points] Assume you want to build a decision tree. Your data set consists of N samples, each with k features ($k \leq N$).

a) If the features are binary, what is the maximum possible number of leaf nodes and the maximum depth of your decision tree?

> Each feature can only be used once on a path from the root to the leaf.
>
> Maximum number of leaf nodes: $\min\left(2^k, N\right)$
>
> Maximum depth: k

b) If the features are continuous, what is the maximum possible number of leaf nodes and the maximum depth of your decision tree?

> Each feature can be used multiple times on a path from the root to the leaf.
>
> Maximum number of leaf nodes: N
>
> Maximum depth: $N - 1$

Regression

Problem 2 [(1+4)=5 points] We want to perform regression on a dataset consisting of N samples $\boldsymbol{x}_i \in \mathbb{R}^D$ with corresponding targets $y_i \in \mathbb{R}$ (represented compactly as $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} \in \mathbb{R}^N$).

Assume that we have fitted an L_2-regularized linear regression model and obtained the optimal weight vector $\boldsymbol{w}^* \in \mathbb{R}^D$ as

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \boldsymbol{w}^T \boldsymbol{w}$$

Note that there is no bias term.

Now, assume that we obtained a new data matrix $\boldsymbol{X}_{new}$ by scaling all samples by the same positive factor $a \in (0, \infty)$. That is, $\boldsymbol{X}_{new} = a\boldsymbol{X}$ (and respectively $\boldsymbol{x}_i^{new} = a\boldsymbol{x}_i$).

a) Find the weight vector $\boldsymbol{w}_{new}$ that will produce the same predictions on $\boldsymbol{X}_{new}$ as $\boldsymbol{w}^*$ produces on $\boldsymbol{X}$.

> Predictions of a linear regression model are generated as $\hat{y} = \boldsymbol{w}^T \boldsymbol{x}$.
>
> This means that we need to ensure that $\boldsymbol{w}^{*T} \boldsymbol{x}_i = \boldsymbol{w}_{new}^T \boldsymbol{x}_i^{new}$ or equivalently $\boldsymbol{w}^{*T} \boldsymbol{x}_i = \boldsymbol{w}_{new}^T a \boldsymbol{x}_i$. Solving for $\boldsymbol{w}_{new}$ we get $\boldsymbol{w}_{new} = \frac{\boldsymbol{w}^*}{a}$

b) Find the regularization factor $\lambda_{new} \in \mathbb{R}$, such that the solution $\boldsymbol{w}_{new}^*$ of the new L_2-regularized linear regression problem

$$\boldsymbol{w}_{new}^* = \arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_i^{new} - y_i)^2 + \frac{\lambda_{new}}{2} \boldsymbol{w}^T \boldsymbol{w}$$

will produce the same predictions on $\boldsymbol{X}_{new}$ as $\boldsymbol{w}^*$ produces on $\boldsymbol{X}$.

Provide a mathematical justification for your answer.

The closed form solution for $\boldsymbol{w}^*$ on the original data $\boldsymbol{X}$ is

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

The closed form solution for $\boldsymbol{w}^*_{new}$ on the new data $\boldsymbol{X}_{new}$ is

$$\begin{aligned}\boldsymbol{w}^*_{new} &= (\boldsymbol{X}_{new}^T\boldsymbol{X}_{new} + \lambda_{new}\boldsymbol{I})^{-1}\boldsymbol{X}_{new}^T\boldsymbol{y}\\ &= a(a^2\boldsymbol{X}^T\boldsymbol{X} + \lambda_{new}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

From the solution in (a), we want to ensure that

$$\begin{aligned}\boldsymbol{w}^*_{new} &\stackrel{!}{=} \frac{1}{a}\boldsymbol{w}^*\\ a(a^2\boldsymbol{X}^T\boldsymbol{X} + \lambda_{new}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} &\stackrel{!}{=} \frac{1}{a}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\ a(a^2\boldsymbol{X}^T\boldsymbol{X} + a^2\frac{\lambda_{new}}{a^2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} &\stackrel{!}{=} \frac{1}{a}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\ \frac{1}{a}(\boldsymbol{X}^T\boldsymbol{X} + \frac{\lambda_{new}}{a^2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} &\stackrel{!}{=} \frac{1}{a}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\ \frac{\lambda_{new}}{a^2} &\stackrel{!}{=} \lambda\\ \lambda_{new} &\stackrel{!}{=} a^2\lambda\end{aligned}$$

Alternative solution

$$\begin{aligned}\frac{\boldsymbol{w}^*}{a} &= \frac{1}{a}\arg\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}\\ &= \frac{1}{a}\arg\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}(\frac{\boldsymbol{w}^T}{a}a\boldsymbol{x}_i - y_i)^2 + \frac{a^2\lambda}{2}\frac{\boldsymbol{w}^T}{a}\frac{\boldsymbol{w}}{a}\\ &= \frac{a}{a}\arg\min_{\boldsymbol{w}_{new}=\frac{\boldsymbol{w}}{a}} \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_{new}^T\boldsymbol{x}_i^{new} - y_i)^2 + \frac{a^2\lambda}{2}\boldsymbol{w}_{new}^T\boldsymbol{w}_{new}\\ &\stackrel{!}{=} \boldsymbol{w}^*_{new} = \arg\min_{\boldsymbol{w}_{new}} \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_{new}^T\boldsymbol{x}_i^{new} - y_i)^2 + \frac{\lambda_{new}}{2}\boldsymbol{w}_{new}^T\boldsymbol{w}_{new}\end{aligned}$$

Classification

Problem 3 [(1+2+3)=6 points] We would like to perform binary classification on multivariate binary data. That is, the data points $\boldsymbol{x}_i \in \{0,1\}^D$ are binary vectors of length D, and each sample belongs to one of two classes $y_i \in \{1,2\}$.

Consider the following generative classification model. We place a categorical prior on y

$$p(y=1) = \pi_1 \qquad p(y=2) = \pi_2.$$

The class-conditional distributions are products of independent Bernoulli distributions

$$p(\boldsymbol{x} \mid y=1, \boldsymbol{\alpha}) = \prod_{j=1}^{D} \mathrm{Bern}(x_j \mid \alpha_j),$$
$$p(\boldsymbol{x} \mid y=2, \boldsymbol{\beta}) = \prod_{j=1}^{D} \mathrm{Bern}(x_j \mid \beta_j),$$

where $\boldsymbol{\alpha} \in [0,1]^D$ and $\boldsymbol{\beta} \in [0,1]^D$ are the respective parameter vectors for both classes.
That is, each component x_j is distributed as $x_j \sim \mathrm{Bern}(\alpha_j)$ if $y=1$ or $x_j \sim \mathrm{Bern}(\beta_j)$ if $y=2$.

a) Write down the expression for the posterior distribution $p(y \mid \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

Using the Bayes formula

$$\begin{aligned} p(y \mid \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \frac{p(\boldsymbol{x} \mid y, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(y \mid \boldsymbol{\pi})}{p(\boldsymbol{x} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})} \\ &= \frac{p(\boldsymbol{x} \mid y, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(y \mid \boldsymbol{\pi})}{p(\boldsymbol{x} \mid y=1, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(y=1 \mid \boldsymbol{\pi}) + p(\boldsymbol{x} \mid y=2, \boldsymbol{\alpha}, \boldsymbol{\pi}) p(y=2 \mid \boldsymbol{\pi})} \end{aligned}$$

b) Assume that $D=3$, $\boldsymbol{\alpha} = [1/3, 1/3, 3/4]$, $\boldsymbol{\beta} = [2/3, 1/2, 1/2]$, $\pi_1 = 1/3$ and $\pi_2 = 2/3$.

Write down a data point $\boldsymbol{x}_1 \in \{0,1\}^3$ that will be classified as class 1 by our model. Additionally, compute the posterior probability $p(y=1 \mid \boldsymbol{x}_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

Consider $\boldsymbol{x}_1 = (0,0,1)^T$. The probability that $\boldsymbol{x}_1$ belongs to class 1 is

$$\begin{aligned} p(y=1 \mid \boldsymbol{x}_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \frac{p(\boldsymbol{x}_1 \mid y=1, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(y=1 \mid \boldsymbol{\pi})}{p(\boldsymbol{x}_1 \mid y=1, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(y=1 \mid \boldsymbol{\pi}) + p(\boldsymbol{x}_1 \mid y=2, \boldsymbol{\alpha}, \boldsymbol{\pi}) p(y=2 \mid \boldsymbol{\pi})} \\ &= \frac{p(\boldsymbol{x}_1 \mid y=1, \boldsymbol{\alpha}, \boldsymbol{\beta}) \pi_1}{p(\boldsymbol{x}_1 \mid y=1, \boldsymbol{\alpha}, \boldsymbol{\beta}) \pi_1 + p(\boldsymbol{x}_1 \mid y=2, \boldsymbol{\alpha}, \boldsymbol{\pi}) \pi_2} \\ &= \frac{(1-\alpha_1) \cdot (1-\alpha_2) \cdot \alpha_3 \cdot \pi_1}{(1-\alpha_1) \cdot (1-\alpha_2) \cdot \alpha_3 \cdot \pi_1 + (1-\beta_1) \cdot (1-\beta_2) \cdot \beta_3 \cdot \pi_2} \\ &= \frac{2/3 \cdot 2/3 \cdot 3/4 \cdot 1/3}{2/3 \cdot 2/3 \cdot 3/4 \cdot 1/3 + 1/3 \cdot 1/2 \cdot 1/2 \cdot 2/3} \\ &= \frac{1/9}{1/9 + 1/18} \\ &= \frac{2}{3} \end{aligned}$$

c) Consider the case when $D = 2$, $\pi_1 = \pi_2 = 1/2$, and $\boldsymbol{\alpha} \in [0,1]^2$ and $\boldsymbol{\beta} \in [0,1]^2$ are known and fixed. Show that the resulting classification rule can be represented as a linear function of $\boldsymbol{x}$. That is, find $\boldsymbol{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$, such that

$$\{\boldsymbol{x} \in \{0,1\}^2 : \boldsymbol{w}^T\boldsymbol{x} + b > 0\} = \{\boldsymbol{x} \in \{0,1\}^2 : p(y = 1 \mid \boldsymbol{x}) > p(y = 2 \mid \boldsymbol{x})\}$$

$$\begin{aligned}
& p(y = 1 \mid \boldsymbol{x}) > p(y = 2 \mid \boldsymbol{x}) \\
\iff & p(\boldsymbol{x} \mid y = 1)p(y = 1) > p(\boldsymbol{x} \mid y = 2)p(y = 2) \\
\iff & p(\boldsymbol{x} \mid y = 1) > p(\boldsymbol{x} \mid y = 2) \\
\iff & \log p(\boldsymbol{x} \mid y = 1) > \log p(\boldsymbol{x} \mid y = 2) \\
\iff & \log p(x_1 \mid y = 1) + \log p(x_2 \mid y = 1) > \log p(x_1 \mid y = 2) + \log p(x_2 \mid y = 2) \\
\iff & x_1 \log \alpha_1 + (1 - x_1) \log(1 - \alpha_1) + x_2 \log \alpha_2 + (1 - x_2) \log(1 - \alpha_2) \\
& - x_1 \log \beta_1 - (1 - x_1) \log(1 - \beta_1) - x_2 \log \beta_2 - (1 - x_2) \log(1 - \beta_2) > 0 \\
\iff & (\log \alpha_1 - \log(1 - \alpha_1) - \log \beta_1 + \log(1 - \beta_1))x_1 \\
& + (\log \alpha_2 - \log(1 - \alpha_2) - \log \beta_2 + \log(1 - \beta_2))x_2 \\
& + (\log(1 - \alpha_1) + \log(1 - \alpha_2) - \log(1 - \beta_1) - \log(1 - \beta_2)) > 0
\end{aligned}$$

We can equivalently represent this inequality as $\boldsymbol{w}^T\boldsymbol{x} + b > 0$, where

$$\boldsymbol{w} = \begin{pmatrix} \log \alpha_1 - \log(1 - \alpha_1) - \log \beta_1 + \log(1 - \beta_1) \\ \log \alpha_2 - \log(1 - \alpha_2) - \log \beta_2 + \log(1 - \beta_2) \end{pmatrix} \in \mathbb{R}^2$$

and

$$b = (\log(1 - \alpha_1) + \log(1 - \alpha_2) - \log(1 - \beta_1) - \log(1 - \beta_2)) \in \mathbb{R}$$

Kernels

Problem 4 [(4)=4 points] Prove or disprove whether the following operations on sets $A, B \subseteq \mathcal{X}$, where $\mathcal{X}$ is a finite set, define a valid kernel.

a) $k(A, B) = |A \times B|$, where $A \times B = \{(a, b) : a \in A, b \in B\}$ denotes the cartesian product and $|S|$ denotes the cardinality of set S, i.e. the number of elements in S.

Yes, using cardinality as a feature map $|A \times B| = |A| \cdot |B|$ obviously defines a kernel.

b) $k(A, B) = |A \cap B|$

Since the set $\mathcal{X}$ is finite we can denote its members as x_i, with $i \in \{1, 2, ...|\mathcal{X}|\}$, and define the membership vector $\boldsymbol{a}$, with

$$a_i = \begin{cases} 1 & \text{if } x_i \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Using this as a feature map we can write

$$|A \cap B| = \boldsymbol{a}^T \boldsymbol{b},$$

showing that $|A \cap B|$ is a kernel.

c) $k(A, B) = |A \cup B|$

Looking at the determinant of a Gram matrix for two sets $A, B \subseteq \mathcal{X}$ we have

$$\det \begin{pmatrix} |A| & |A \cup B| \\ |A \cup B| & |B| \end{pmatrix} = |A||B| - |A \cup B|^2,$$

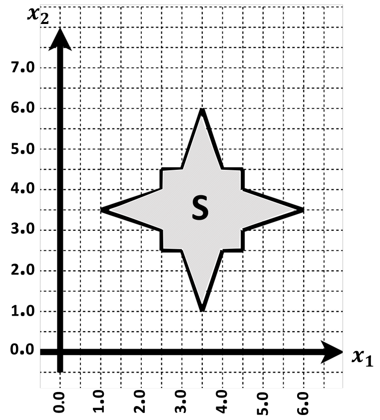
which is negative if $A \neq B$. Hence, the Gram matrix is not positive semidefinite and according to Mercer's theorem $|A \cup B|$ is not a kernel.

Optimization

Problem 5 [(1+3+2)=6 points] Let f be the following convex function on $\mathbb{R}^2$:

$$f(x_1, x_2) = e^{x_1+x_2} - 5 \cdot \log(x_2)$$

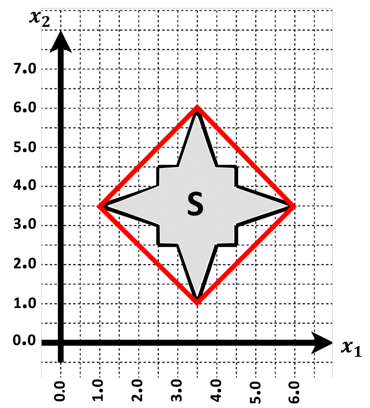
a) Consider the following shaded region $\boldsymbol{S}$ in $\mathbb{R}^2$. Is this region convex? Why?



It is not. Because we can choose two points in $\boldsymbol{S}$ where the line connecting the points does not completely resides in $\boldsymbol{S}$.

b) Find the maximizer (x_1^*, x_2^*) of f over the shaded region $\boldsymbol{S}$. For your computations, you can pick values from the following table. Justify your answer.

$e^{4.5} = 90.017$	$e^{5.0} = 148.41$	$e^{5.5} = 244.69$	$e^{6.5} = 665.14$
$e^{7.0} = 1096.63$	$e^{7.5} = 1808.04$	$e^{8.0} = 2980.95$	$e^{8.5} = 4914.76$
$e^{9.0} = 8103.08$	$e^{9.5} = 13359.726$	$e^{10.0} = 22026.46$	$e^{10.5} = 36315.50$
$\log(1.0) = 0$	$\log(2.5) = 0.9162$	$\log(3.0) = 1.0986$	$\log(3.5) = 1.2527$
$\log(4.0) = 1.3862$	$\log(4.5) = 1.5040$	$\log(5.0) = 1.6094$	$\log(6.0) = 1.7917$



According to the lecture, the maximum of a convex function over a set $\boldsymbol{S}$ lies on the vertices of the convex hull of $\boldsymbol{S}$. Therefore, we need to compute the value of f only on the four corners of the convex hull. Afterwards, we pick the corner at which the value of f is larger.

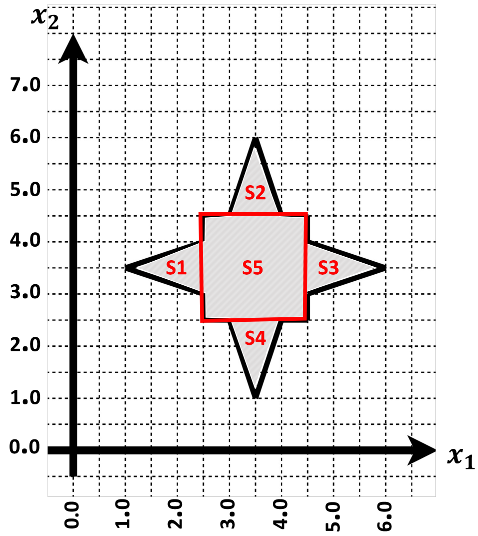
$$f(1, 3.5) = e^{4.5} - 5 \cdot \log(3.5) = 83.7533 \qquad f(3.5, 1.0) = e^{4.5} - 5 \cdot \log(1.0) = 90.01$$
$$f(6, 3.5) = e^{9.5} - 5 \cdot \log(3.5) = 13353.46 \qquad f(3.5, 6.0) = e^{9.5} - 5 \cdot \log(6.0) = 13350.76$$

Therefore, the maximum of f occurs at point $(6, 3.5)$, and the maximum of f is 13353.46.

c) Assume that we are given an algorithm `ConvOpt`$(f, \mathcal{X})$ that takes as input a convex function f and any <u>convex</u> region $\mathcal{X}$, and returns the <u>minimum</u> of f over $\mathcal{X}$.

Using the `ConvOpt` algorithm, how would you find the global <u>minimum</u> of f over the shaded region $\boldsymbol{S}$?

We can partition the shaded region S to the following five convex regions. Afterwards, we run the `ConvOpt` algorithm separately for the 5 regions.



SVM

Problem 6 [(5)=5 points] Given the data points

$$\boldsymbol{x}_1 = (1,1,0,1)^T \qquad \boldsymbol{x}_2 = (1,1,1,0)^T \qquad \boldsymbol{x}_3 = (0,1,1,1)^T \qquad \boldsymbol{x}_4 = (0,0,1,1)^T$$

Prove or disprove whether the following combinations of labels $\boldsymbol{y}$ and dual variables $\boldsymbol{\alpha}$ are the optimal solutions of a soft-margin SVM with $C = 1$.

a) $\boldsymbol{y} = (-1,-1,1,1)^T$, $\boldsymbol{\alpha} = (0.6, 0.6, 1, 0)^T$

b) $\boldsymbol{y} = (-1,-1,1,1)^T$, $\boldsymbol{\alpha} = \left(\frac{2}{3}, \frac{2}{3}, \frac{4}{3}, 0\right)^T$

c) $\boldsymbol{y} = (-1,1,-1,1)^T$, $\boldsymbol{\alpha} = (1,1,1,1)^T$

a) No, $\sum_i \alpha_i y_i \neq 0$.

b) No, $\alpha_3 > C$. **1pt.**

c) This solution is feasible, so we need to check optimality. The dual problem of the soft-margin SVM is

$$g(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j.$$

To check whether this is an optimal solution we first look at the gradient, which is given by

$$(\boldsymbol{\nabla}_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}))_i = 1 - \sum_j y_i y_j \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j.$$

To calculate the gradient, we need the kernel matrix, with $K_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j$:

$$K = \begin{pmatrix} 3 & 2 & 2 & 1 \\ 2 & 3 & 2 & 1 \\ 2 & 2 & 3 & 2 \\ 1 & 1 & 2 & 2 \end{pmatrix}.$$

Using this we can directly calculate

$$\begin{aligned}
\partial_{\alpha_1} g(\boldsymbol{\alpha}) &= 1 - (3 - 2 + 2 - 1) = -1 \\
\partial_{\alpha_2} g(\boldsymbol{\alpha}) &= 1 - (-2 + 3 - 2 + 1) = 1 \\
\partial_{\alpha_3} g(\boldsymbol{\alpha}) &= 1 - (2 - 2 + 3 - 2) = 0 \\
\partial_{\alpha_4} g(\boldsymbol{\alpha}) &= 1 - (-1 + 1 - 2 + 2) = 1
\end{aligned}$$

Now let us consider the constraints. Since all $\alpha_i = 1 = C$ we cannot increase any α_i. However, we also can't decrease α_1 without decreasing either α_2 or α_4 due to the constraint $\sum_i \alpha_i y_i = 0$. The absolute values of these three derivatives are equally high, which means that decreasing these values would not increase $g(\boldsymbol{\alpha})$. Hence, we cannot increase $g(\boldsymbol{\alpha})$ any further and this is indeed an optimal solution.

Deep Learning

Problem 7 [(2+2)=4 points] You are trying to solve a regression task and you want to choose between two approaches:

1. A simple linear regression model.
2. A feed forward neural network $f_{\boldsymbol{W}}(\boldsymbol{x})$ with L hidden layers, where each hidden layer $l \in \{1, ..., L\}$ has a weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{D \times D}$ and a ReLU activation function. The output layer has a weight matrix $\boldsymbol{W}_{L+1} \in \mathbb{R}^{D \times 1}$ and no activation function.

In both models, there are no bias terms.

Your dataset $\mathcal{D}$ contains data points with nonnegative features $\boldsymbol{x}_i$ and the target y_i is continuous:

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \qquad \boldsymbol{x}_i \in \mathbb{R}^D_{\geq 0}, \qquad y_i \in \mathbb{R}$$

Let $\boldsymbol{w}^*_{LS} \in \mathbb{R}^D$ be the optimal weights for the linear regression model corresponding to a global minimum of the following least squares optimization problem:

$$\boldsymbol{w}^*_{LS} = \underset{\boldsymbol{w} \in \mathbb{R}^D}{\arg\min}\, \mathcal{L}_{LS}(\boldsymbol{w}) = \underset{\boldsymbol{w} \in \mathbb{R}^D}{\arg\min}\, \frac{1}{2} \sum_{i=1}^N (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2$$

Let $\boldsymbol{W}^*_{NN} = \{\boldsymbol{W}^*_1, \dots, \boldsymbol{W}^*_{L+1}\}$ be the optimal weights for the neural network corresponding to a global minimum of the following optimization problem:

$$\boldsymbol{W}^*_{NN} = \underset{\boldsymbol{W}}{\arg\min}\, \mathcal{L}_{NN}(\boldsymbol{W}) = \underset{\boldsymbol{W}}{\arg\min}\, \frac{1}{2} \sum_{i=1}^N (f_{\boldsymbol{W}}(\boldsymbol{x}_i) - y_i)^2$$

a) Assume that the optimal $\boldsymbol{W}^*_{NN}$ you obtain are non-negative.
What will be the relation $(<, \leq, =, \geq, >)$ between the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN})$ and the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$? Provide a mathematical argument to justify your answer.

Note that for any non-negative $\boldsymbol{x}$ and any non-negative $\boldsymbol{W}$ it holds $\text{ReLU}(\boldsymbol{x}\boldsymbol{W}) = \boldsymbol{x}\boldsymbol{W}$.

Therefore, since our data points have non-negative features $\boldsymbol{x}_i$ and the optimal weights $\boldsymbol{W}^*_{NN}$ are non-negative, every ReLU layer is equivalent to a linear layer when plugging in the optimal weights. This means we can write

$$\begin{aligned} f_{\boldsymbol{W}^*_{NN}}(\boldsymbol{x}_i) &= ReLU(ReLU(ReLU(\boldsymbol{x}_i^T \boldsymbol{W}^*_1)\boldsymbol{W}^*_2) \cdots \boldsymbol{W}^*_L)\boldsymbol{W}^*_{L+1} \\ &= \boldsymbol{x}_i^T \boldsymbol{W}^*_1 \boldsymbol{W}^*_2 \cdots \boldsymbol{W}^*_{L+1} \\ &= \boldsymbol{x}_i^T \boldsymbol{w}^*_{NN} \end{aligned}$$

where we defined $\boldsymbol{w}^*_{NN} = \boldsymbol{W}^*_1 \boldsymbol{W}^*_2 \cdots \boldsymbol{W}^*_{L+1}$. From this we can see that the neural network with optimal weights behaves like a linear regression with a different set of weights $\boldsymbol{w}^*_{NN}$.

Note also that linear regression is a special case of the above neural network, i.e. for any weights $\boldsymbol{w}_{LS}$ you can find weights $\boldsymbol{W}_{NN}$ that produce the same output.

Given the above facts and since we the optimal weights correspond to a global minima we can conclude that $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN}) = \mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$ and the optimal weights found by solving the least squares optimization problem will be $\boldsymbol{w}^*_{LS} = \boldsymbol{w}^*_{NN}$.

b) In contrast to (a), now assume that the optimal weights $\boldsymbol{w}_{LS}^*$ you obtain are non-negative. What will be the relation $(<, \leq, =, \geq, >)$ between the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}_{LS}^*)$ and the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}_{NN}^*)$? Provide a mathematical argument to justify your answer.

As stated in (a) linear regression is a special case of the above neural network, i.e. for any weights $\boldsymbol{w}_{LS}$ you can find weights $\boldsymbol{W}_{NN}$ that produce the same output. That is, everything that can be learned with a linear regression can be learned equally well with a neural network.

However, the reverse direction doesn't hold, since in principle neural networks can learn more complicated functions compared to linear regression. Moreover, the given fact that $\boldsymbol{w}_{LS}^*$ are non-negative does not tell us anything about the optimal weights of the neural network $\boldsymbol{W}_{NN}^*$.

Therefore it holds $\mathcal{L}_{NN}(\boldsymbol{W}_{NN}^*) \leq \mathcal{L}_{LS}(\boldsymbol{w}_{LS}^*)$ since the neural network can potentially find a better fit for the data (e.g. by taking advantage of non-linearity).

Dimensionality Reduction

Problem 8 [(3+2+2)=7 points] You are given $N = 4$ data points: $\{\boldsymbol{x}_i\}_{i=1}^4, \boldsymbol{x}_i \in \mathbb{R}^3$, represented with the matrix $\boldsymbol{X} \in \mathbb{R}^{4\times 3}$.

$$\boldsymbol{X} = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 1 & -2 \\ 4 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix}$$

Hint: In this task the results of all (final and intermediate) computations happen to be integers.

a) Perform principal component analysis (PCA) of the data $\boldsymbol{X}$, i.e. find the principal components and their associated variances in the transformed coordinate system. Show your work.

First we center the data. The mean is $\bar{\boldsymbol{x}} = [2, 1, 1]$, thus we have

$$\boldsymbol{X}_c = \boldsymbol{X} - \bar{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \end{bmatrix}$$

Then we compute the covariance matrix.

$$\boldsymbol{\Sigma}_{X_c} = \frac{1}{N}\boldsymbol{X}_c^T\boldsymbol{X}_c = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Since $\boldsymbol{\Sigma}_{X_c}$ it is already in a diagonal form we can conclude that $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{X_c}$ and $\boldsymbol{\Gamma} = \boldsymbol{I_3}$, and that it holds $\boldsymbol{\Sigma}_{X_c} = \boldsymbol{\Gamma\Lambda\Gamma}^T$. The principal components are the canonical basis vectors.

b) Project the data to two dimensions, i.e. write down the transformed data matrix $\boldsymbol{Y} \in \mathbb{R}^{4\times 2}$ using the top-2 principal components you computed in (a). What fraction of variance of $\boldsymbol{X}$ is preserved by $\boldsymbol{Y}$?

The projection matrix is:

$$\boldsymbol{\Gamma}_{trunc} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

since we pick the first and the third principal vector corresponding to the two largest eigenvalues. Thus, we have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Gamma}_{trunc} = \begin{bmatrix} 2 & 1 \\ 0 & -3 \\ 2 & 1 \\ -4 & 1 \end{bmatrix}$$

We preserve $\frac{6+3}{6+2+3} = \frac{9}{11}$ of the variance.

c) Let $\boldsymbol{x}_5 \in \mathbb{R}^3$ be a new data point. Specify the vector $\boldsymbol{x}_5$ such that performing PCA on the data including the new data point $\{\boldsymbol{x}_i\}_{i=1}^5$ leads to exactly the same principal components as in (a).

Let $\boldsymbol{x}_5 = \bar{\boldsymbol{x}}$, i.e. the new data point equals the mean before including $\boldsymbol{x}_5$ to the dataset. Therefore, the new mean including $\boldsymbol{x}_5$ is equal to the old mean. We have:

$$\boldsymbol{X}_c = \boldsymbol{X} - \bar{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

which leads to the same $\boldsymbol{\Sigma}_{X_c}$ as in (a) up to a difference in the multiplicative constant, in (a) we had $\frac{1}{4}\boldsymbol{X}_c^T\boldsymbol{X}_c$ and here we have $\frac{1}{5}\boldsymbol{X}_c^T\boldsymbol{X}_c$. While the this difference leads to different eigenvalues, the eigenvectors and thus the principal components stay the same.

Clustering

Problem 9 [(4)=4 points] Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ be the centroids computed by the K-means algorithm. Prove that the set $\mathcal{X}_j$ of all points in $\mathbb{R}^D$ assigned during inference to the cluster j is a convex set.

$$\mathcal{X}_j := \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} \text{ would be assigned to centroid } \boldsymbol{\mu}_j \text{ by } K\text{-means}\}$$

Hint: start by thinking about the case with $K = 2$.

Consider the case with $K = 2$ clusters denoted as j and k.

The set of points assigned to cluster j and not to cluster k is

$$\begin{aligned} \mathcal{X}_{jk} &:= \{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2 < \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^D : (\boldsymbol{x} - \boldsymbol{\mu}_j)^T(\boldsymbol{x} - \boldsymbol{\mu}_j) < (\boldsymbol{x} - \boldsymbol{\mu}_k)^T(\boldsymbol{x} - \boldsymbol{\mu}_k)\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^D : 2(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\boldsymbol{x} + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k - \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j > 0\} \end{aligned}$$

is a half-space (all points to one side of a hyperplane). This set is obviously convex.

In case we have $K > 2$ clusters, the set of all points assigned to centroid j can be expressed as an intersection

$$\begin{aligned} \mathcal{X}_j &= \bigcap_{k \neq j} \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} \text{ is assigned to } j \text{ and not to } k\} \\ &= \bigcap_{k \neq j} \mathcal{X}_{jk} \end{aligned}$$

which is an intersection of convex sets, and therefore is convex.

Problem 10 [(2)=2 points] Given three 1-dimensional Gaussian distributions $\mathcal{N}(\mu_i, \sigma_i^2)$ with parameters

$$\mu_1 = 1, \qquad \mu_2 = -1, \qquad \mu_3 = 0,$$
$$\sigma_1 = 1, \qquad \sigma_2 = 0.5, \qquad \sigma_3 = 2.5$$

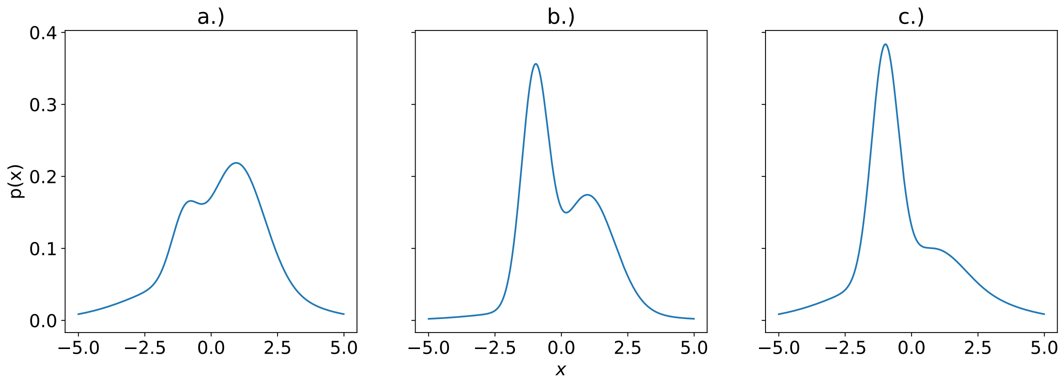
and three different vectors of mixing coefficients π defining categorical cluster priors.

Match the value of π in each row of the following table with one of the probability density functions

$$p(x) = \sum_{i=1}^{3} \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i)$$

of the resulting GMM showed below. Fill in the last column of the table, no argumentation required.

	π_1	π_2	π_3	PDF (a, b or c)
case 1	0.111...	0.444...	0.444...	
case 2	0.444...	0.111...	0.444...	
case 3	0.444...	0.444...	0.111...	



case 1 is c.)

case 2 is a.)

case 3 is b.)

Variational Inference

Problem 11 [(3+1+1+2)=7 points] Consider the following latent variable probabilistic model

$$p(z) = \mathcal{N}(z \mid 0, 1)$$
$$p(x \mid z) = \mathcal{N}(x \mid z, 1)$$

We want to approximate the posterior distribution $p(z \mid x)$ using the following variational family

$$\mathcal{Q} = \{\mathcal{N}(z \mid \mu, 1) \text{ for } \mu \in \mathbb{R}\}$$

that includes all normal distributions with unit variance.

Questions (a), (b), (c) and (d) are all concerning this setup.

Hint: Variance of $p(z \mid x)$ *is equal to* 0.5.

a) Write down the closed-form expression for ELBO $\mathcal{L}(q)$ and simplify it. You can ignore all the terms constant in μ.

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log p(x, z) - \log q(z \mid \mu)\right] \\
&= \mathbb{E}_q\left[\log p(x \mid z) + \log p(z) - \log q(z \mid \mu)\right] \\
&= \mathbb{E}_q\left[-\frac{1}{2}(x - z)^2 - \frac{1}{2}z^2 + \frac{1}{2}(z - \mu)^2\right] + \text{const.} \\
&= \mathbb{E}_q\left[zx - \frac{1}{2}z^2 - z\mu + \frac{1}{2}\mu^2\right] + \text{const.} \\
&= \mathbb{E}_q\left[z\right]x - \frac{1}{2}\mathbb{E}_q\left[z^2\right] - \mathbb{E}_q\left[z\right]\mu + \frac{1}{2}\mu^2 + \text{const.} \\
&= \mu x - \frac{1}{2}(\mu^2 + 1^2) - \mu^2 + \frac{1}{2}\mu^2 + \text{const.} \\
&= \mu x - \frac{1}{2}(\mu^2 + 1^2) - \mu^2 + \frac{1}{2}\mu^2 + \text{const.} \\
&= \mu x - \mu^2 + \text{const.}
\end{aligned}$$

b) Find the optimal variational distribution $q^* \in \mathcal{Q}$ that maximizes the ELBO

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q)$$

i.e. find the mean μ^* of the optimal variational distribution q^*.

We already know the ELBO: just compute the derivative w.r.t. μ and set it to zero to obtain the optimal μ^*.

$$\begin{aligned}
\mathcal{L}(\mu) &= \mu x - \mu^2 + \text{const.} \\
\frac{\partial}{\partial \mu}\mathcal{L}(\mu) &= x - 2\mu \stackrel{!}{=} 0 \\
\iff \mu^* &= \frac{x}{2}
\end{aligned}$$

c) Assume that the optimal q^* (i.e., the optimal μ^*) from question (b) is given. Which of the following statements is true?

(1) $\mathbb{KL}(q(z \mid \mu^*) \parallel p(z \mid x)) < 0$

(2) $\mathbb{KL}(q(z \mid \mu^*) \parallel p(z \mid x)) = 0$

(3) $\mathbb{KL}(q(z \mid \mu^*) \parallel p(z \mid x)) > 0$

Justify your answer.

KL-divergence is non-negative, so (1) is impossible. KL-divergence is equal to zero if and only if two distribution are equal. We know that variance of $p(z \mid x)$ is 0.5 (see hint), and the variance of $q(z \mid \mu)$ for any μ is equal to 1 (by definition). Hence, $q(z \mid \mu^*) \neq p(z \mid x)$ and therefore KL-divergence is greater than zero (option (3)).

d) For each of the conditions (1), (2), (3) from question (c) above, provide a parametric variational family $\mathcal{Q}_i$, such that the optimal q_i^* from each family would fulfill the respective condition, or explain why it's impossible.

That is, provide $\mathcal{Q}_1$, such that for $q_1^* = \arg\max_{q \in \mathcal{Q}_1} \mathcal{L}(q)$ we have $\mathbb{KL}(q_1^*(z) \parallel p(z \mid x)) < 0$, for $q_2^* = \arg\max_{q \in \mathcal{Q}_2} \mathcal{L}(q)$ we have $\mathbb{KL}(q_2^*(z) \parallel p(z \mid x)) = 0$, and for $q_3^* = \arg\max_{q \in \mathcal{Q}_3} \mathcal{L}(q)$ we have $\mathbb{KL}(q_3^*(z) \parallel p(z \mid x)) > 0$.

(1) Is impossible since KL is always nonnegative.

(2) The likelihood $p(x \mid z)$ and the prior $p(z)$ are both normal distributions, hence the posterior $p(z \mid x)$ is also a normal distribution. Therefore, the posterior is contained in the set of all normal distributions $\mathcal{Q}_2 = \{\mathcal{N}(\mu, \sigma^2) \text{ for } \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}\}$.

(3) $\mathcal{Q}_3 = \{\mathcal{N}(\mu, 1) \text{ for } \mu \in \mathbb{R}\}$ works (see answer to (c)).