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 18 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?

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

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 $\boldsymbol{X}$.

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.

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

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

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 $x_j \sim \operatorname{Bern}(\alpha_j)$ if $y=1$ or $x_j \sim \operatorname{Bern}(\beta_j)$ if $y=2$.

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

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 $p(y=1 \mid \boldsymbol{x}_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

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

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

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.

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

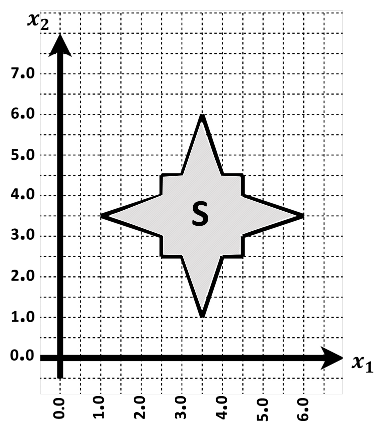
c) $k(A, B) = |A \cup B|$

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?



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.

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

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

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

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

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:

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

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 $\boldsymbol{W}_{NN}^* = \{\boldsymbol{W}_1^*, \ldots, \boldsymbol{W}_{L+1}^*\}$ 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.

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.

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 $\boldsymbol{X} \in \mathbb{R}^{4\times3}$.

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

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.

b) Project the data to two dimensions, i.e. write down the transformed data matrix $\boldsymbol{Y} \in \mathbb{R}^{4\times2}$

using the top-2 principal components you computed in (a). What fraction of variance of $\boldsymbol{X}$ is preserved by $\boldsymbol{Y}$?

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).

Clustering

Problem 9 [(4)=4 points] Let $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ 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 $K = 2$.

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

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

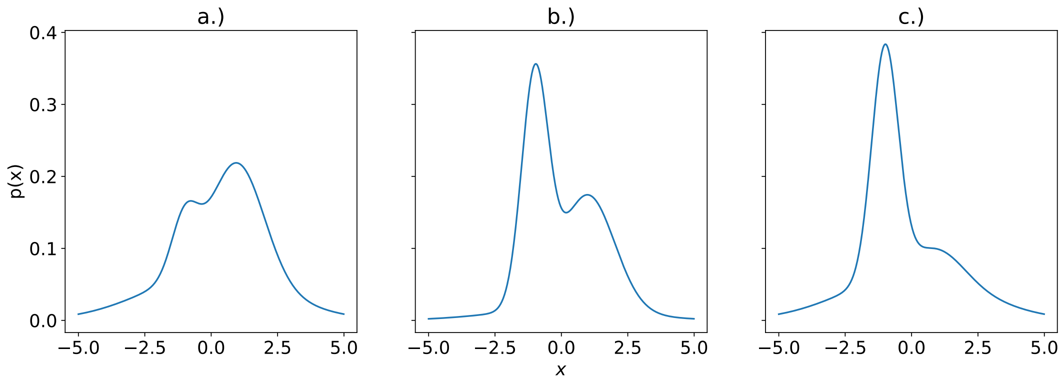
and three different vectors of mixing coefficients π defining categorical cluster priors.

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...	



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.

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^*.

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.

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 $\mathbb{KL}(q_3^*(z) \parallel p(z \mid x)) > 0$.