Data Analytics and Machine Learning
Department of Informatics
Technical University of Munich

Eexam
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Maschinelles Lernen

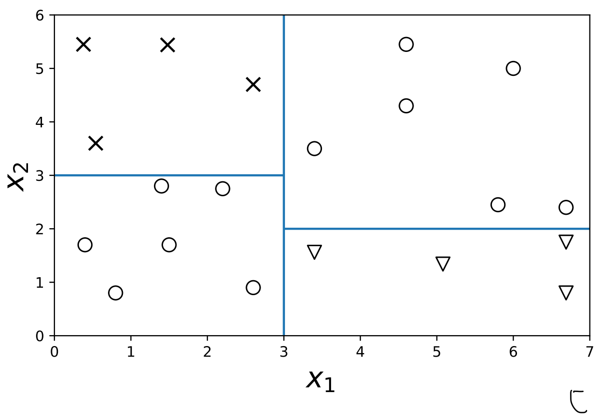
Exam:	IN2064 / Endterm	**Date:**	Wednesday 13$^{\text{th}}$ April, 2022
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	08:00 – 10:00

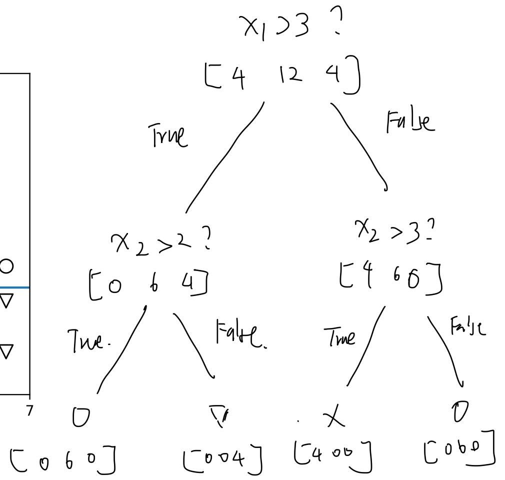
Working instructions

- This graded exercise consists of **48 pages** with a total of **11** problems and four versions of each problem.
  Please make sure now that you received a complete copy of the graded exercise.
- Use the problem versions specified in your personalized submission sheet on TUMExam. Different problems may have different versions: e.g. Problem 1 (Version A), Problem 5 (Version C), etc. If you solve the wrong version you get **zero** points.
- The total amount of achievable credits in this graded exercise is 82.
- This document is copyrighted and it is **illegal** for you to distribute it or upload it to any third-party websites.
- Do **not** submit the problem descriptions (this document) to TUMexam
- You can ignore the “student sticker” box above.

Problem 1: Decision Trees (Version A) (8 credits)

You are given a dataset with points from three different classes and want to classify them using a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols [×, ∘, ▽]) and the decision boundaries of a decision tree of depth 2. Each decision boundary corresponds to a specific decision node.





0 1 2

a) Draw the corresponding decision tree. Make sure that you include the feature (x_1 or x_2) and threshold of each inner node (i.e. decision node). Also, for each inner node and leaf node, indicate the number of samples of each class that pass the node.

$i_G(t_1) = 1 - \sum \pi_{ci}^2 = 1 - \left(\frac{1}{5}^2 + \frac{3}{5}^2 + \frac{1}{5}^2 \right) = \frac{14}{25}\ \frac{56}{100}$

$i_G(t_{2L}) = 1 - \sum \pi_{ci}^2 = 1 - \left(\frac{6}{10}^2 + \frac{4}{10}^2 \right) = \frac{48}{100}$

0 1 2

b) Compute the Gini index of each node of your decision tree.

$i_G(t_{2R}) = 1 - \sum \pi_{ci}^2 = 1 - (- -) = \frac{48}{100}$

$i_G(t_{3LL}) = 1 - \sum \pi_{ci} = 1 - 1 = 0$

0 1 2

c) Compute the misclassification rate of each node of your decision tree.

$i_E(t_1) = 1 - \max p = \frac{2}{5}$ $i_E(t_3) = 0$

$i_E(t_{2L}) = 1 - \max p = \frac{2}{5} = i_E(t_{2R})$

0 1

d) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the Gini index as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem b).

Yes

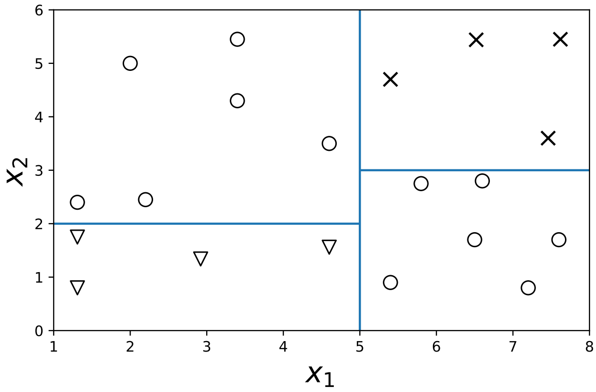
0 1

e) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the misclassification rate as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem c).

No

Problem 1: Decision Trees (Version B) **(8 credits)**

You are given a dataset with points from three different classes and want to classify them using a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols [$\times, \circ, \triangledown$]) and the decision boundaries of a decision tree of depth 2.



a) Draw the corresponding decision tree. Make sure that you include the feature (x_1 or x_2) and threshold of each inner node (i.e. decision node). Also, for each inner node and leaf node, indicate the number of samples of each class that pass the node.

0
1
2

b) Compute the Gini index of each node of your decision tree.

0
1
2

c) Compute the misclassification rate of each node of your decision tree.

0
1
2

d) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the Gini index as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem b).

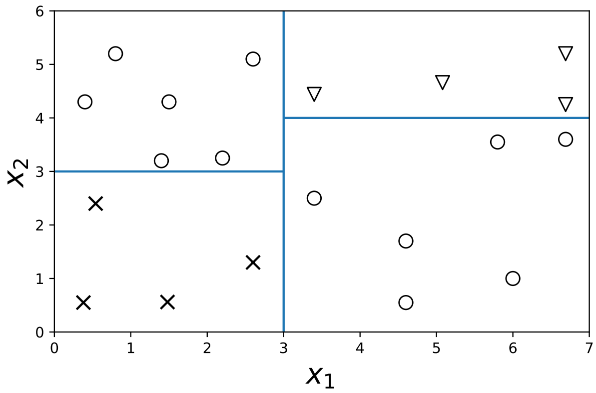
0
1

e) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the misclassification rate as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem c).

0
1

Problem 1: Decision Trees (Version C) (8 credits)

You are given a dataset with points from three different classes and want to classify them using a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols [$\times, \circ, \triangledown$]) and the decision boundaries of a decision tree of depth 2.



0 1 2 a) Draw the corresponding decision tree. Make sure that you include the feature (x_1 or x_2) and threshold of each inner node (i.e. decision node). Also, for each inner node and leaf node, indicate the number of samples of each class that pass the node.

0 1 2 b) Compute the Gini index of each node of your decision tree.

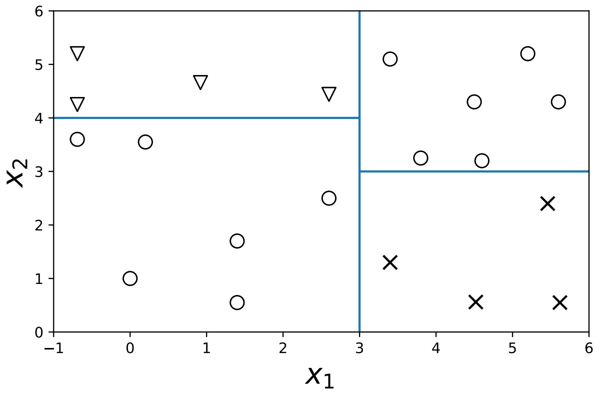
0 1 2 c) Compute the misclassification rate of each node of your decision tree.

0 1 d) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the Gini index as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem b).

0 1 e) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the misclassification rate as the impurity measure? Briefly justify your answer. *Note: You may refer to the solution of subproblem c).*

Problem 1: Decision Trees (Version D) **(8 credits)**

You are given a dataset with points from three different classes and want to classify them using a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols [$\times, \circ, \triangledown$]) and the decision boundaries of a decision tree of depth 2.



a) Draw the corresponding decision tree. Make sure that you include the feature (x_1 or x_2) and threshold of each inner node (i.e. decision node). Also, for each inner node and leaf node, indicate the number of samples of each class that pass the node.

b) Compute the Gini index of each node of your decision tree.

c) Compute the misclassification rate of each node of your decision tree.

d) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the Gini index as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem b).

e) Could this tree have been obtained from greedy optimization (as discussed in the lecture) using the misclassification rate as the impurity measure? Briefly justify your answer.
Note: You may refer to the solution of subproblem c).

Problem 2: Probabilistic Inference (Version A) (8 credits)

Consider the probabilistic model composed of the two following densities

$$\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} \theta_c^{\alpha_c - 1}$$

$$\mathbb{P}(x \mid \boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(x=c)} \text{ with } \sum_{c=1}^{C} \theta_c = 1$$

with probabilistic parameters $\boldsymbol{\theta} \in [0,1]^C$, a set of observations $\mathcal{D} = \{x_1, ..., x_N\}$ consisting of N samples $x_i \in \{1, ..., C\}$, and some given $\boldsymbol{\alpha} \in \mathbb{R}^C$. Here, Γ is the Gamma function.

0 1 2 3 4 5 6 7 8

Derive the maximum a posteriori estimate of the parameter $\boldsymbol{\theta}$ denoted $\boldsymbol{\theta}_{\text{MAP}}$.

Note: The maximum a posteriori estimate must also fulfill $\sum_{c=1}^{C} \theta_{\text{MAP},c} = 1$

posterior distribution $p(\theta \mid D) = \dfrac{p(D \mid \theta) \cdot p(\theta \mid \alpha)}{p(D)}$

$$\propto p(D \mid \theta)\, p(\theta \mid \alpha)$$

$$\propto \prod_{i=1}^{N} \prod_{c=1}^{C} \theta_c^{\mathbb{I}(x=c)} \cdot \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \cdot \prod_{c=1}^{C} \theta_c^{\alpha_c - 1}$$

$$\ln p(\theta \mid D) = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{I}(x=c) \ln \theta_c + \ln(\Gamma(\textstyle\sum_{c=1}^{C} \alpha_c)) - \sum_{c=1}^{C} \ln \Gamma(\alpha_c) + \sum_{c=1}^{C} \alpha_c - 1 \ln \theta_c$$

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{I}(x=c) \ln \theta_c + \sum_{c=1}^{C} \alpha_{c-1} \ln \theta_c + \lambda(1 - \textstyle\sum_{c=1}^{C} \theta_c) + C$$

$$\frac{\partial \mathcal{L}}{\partial \theta_c} = N_c \frac{1}{\theta_c} + \frac{\alpha_c - 1}{\theta_c} - \lambda \stackrel{!}{=} 0$$

$$\frac{N_c + \alpha_c - 1}{\theta_c} = \lambda$$

$$\theta_c = \frac{N_c + \alpha_c - 1}{\lambda}$$

$$\sum_{c=1}^{C} \theta_c = 1$$

$$\sum_{c=1}^{C} (N_c + \alpha_c - 1) = \lambda$$

$$\theta_{c\,\text{MAP}} = \frac{N_c + \alpha_c - 1}{\sum_{c=1}^{C} (N_c + \alpha_c - 1)}$$

Problem 2: Probabilistic Inference (Version B) (8 credits)

Consider the probabilistic model composed of the two following densities

$$\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} \theta_c^{\alpha_c - 1}$$

$$\mathbb{P}(x \mid \boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(x=c)} \text{ with } \sum_{c=1}^{C} \theta_c = 1$$

with probabilistic parameters $\boldsymbol{\theta} \in [0,1]^C$, a set of observations $\mathcal{D} = \{x_1, ..., x_N\}$ consisting of N samples $x_i \in \{1, ..., C\}$, and some given $\boldsymbol{\alpha} \in \mathbb{R}^C$. Here, Γ is the Gamma function.

Derive the maximum a posteriori estimate of the parameter $\boldsymbol{\theta}$ denoted $\boldsymbol{\theta}_{\text{MAP}}$.
Note: The maximum a posteriori estimate must also fulfill $\sum_{c=1}^{C} \theta_{\text{MAP},c} = 1$

0
1
2
3
4
5
6
7
8

Problem 2: Probabilistic Inference (Version C) **(8 credits)**

Consider the probabilistic model composed of the two following densities

$$\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} \theta_c^{\alpha_c - 1}$$

$$\mathbb{P}(x \mid \boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(x=c)} \text{ with } \sum_{c=1}^{C} \theta_c = 1$$

with probabilistic parameters $\boldsymbol{\theta} \in [0, 1]^C$, a set of observations $\mathcal{D} = \{x_1, ..., x_N\}$ consisting of N samples $x_i \in \{1, ..., C\}$, and some given $\boldsymbol{\alpha} \in \mathbb{R}^C$. Here, Γ is the Gamma function.

0 Derive the maximum a posteriori estimate of the parameter $\boldsymbol{\theta}$ denoted $\boldsymbol{\theta}_{\text{MAP}}$.
1 *Note*: The maximum a posteriori estimate must also fulfill $\sum_{c=1}^{C} \theta_{\text{MAP},c} = 1$
2
3
4
5
6
7
8

Problem 2: Probabilistic Inference (Version D) (8 credits)

Consider the probabilistic model composed of the two following densities

$$\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} \theta_c^{\alpha_c - 1}$$

$$\mathbb{P}(x \mid \boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(x=c)} \text{ with } \sum_{c=1}^{C} \theta_c = 1$$

with probabilistic parameters $\boldsymbol{\theta} \in [0,1]^C$, a set of observations $\mathcal{D} = \{x_1, ..., x_N\}$ consisting of N samples $x_i \in \{1, ..., C\}$, and some given $\boldsymbol{\alpha} \in \mathbb{R}^C$. Here, Γ is the Gamma function.

Derive the maximum a posteriori estimate of the parameter $\boldsymbol{\theta}$ denoted $\boldsymbol{\theta}_{\text{MAP}}$.
Note: The maximum a posteriori estimate must also fulfill $\sum_{c=1}^{C} \theta_{\text{MAP},c} = 1$

0
1
2
3
4
5
6
7
8

Problem 3: Linear Regression (Version A) **(8 credits)**

We want to perform regression on a dataset consisting of N samples $\mathbf{x}_i \in \mathbb{R}^D$ ($D > 1$) with corresponding targets $y_i \in \mathbb{R}$ (represented compactly as $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\mathbf{y} \in \mathbb{R}^N$).
Assume that we have a deep neural network with L layers such that $f(\mathbf{x}, \mathbb{W}) = \mathbf{W}^{(L)}\sigma\left(\mathbf{W}^{(L-1)} \dots \left(\sigma\left(\mathbf{W}^{(0)}\mathbf{x}\right)\right)\right)$ where $\mathbb{W} = \{\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L)}\}$. Here, σ is the sigmoid activation function applied element-wise.

0 1 2 3 4

a) Suppose the weights $\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L-1)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *last* layer:

$$\arg\min_{\mathbf{W}^{(L)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(L)}$ of this optimization problem in closed-form ? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(L)}$. If no, explain how we can approximately solve this problem in practice.

0 1 2 3 4

b) Suppose the weights $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *first* layer:

$$\arg\min_{\mathbf{W}^{(0)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(0)}$ of this optimization problem in closed-form? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(0)}$. If no, explain how we can approximately solve this problem in practice.

set $\sigma(W^{L-1} \dots) = \phi(x)$

$f(x, W) = W^{L}\phi(x)$

$W^{*(L)} = (\Phi^T\Phi)^{-1}\Phi y$

No because chain rule each sigmoid contain W^0 cannot get close form

use GD and its variance

Problem 3: Linear Regression (Version B) (8 credits)

We want to perform regression on a dataset consisting of N samples $\mathbf{x}_i \in \mathbb{R}^D$ ($D > 1$) with corresponding targets $y_i \in \mathbb{R}$ (represented compactly as $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\mathbf{y} \in \mathbb{R}^N$).
Assume that we have a deep neural network with L layers such that $f(\mathbf{x}, \mathbb{W}) = \mathbf{W}^{(L)}\sigma\left(\mathbf{W}^{(L-1)} \dots \left(\sigma\left(\mathbf{W}^{(0)}\mathbf{x}\right)\right)\right)$ where $\mathbb{W} = \{\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L)}\}$. Here, σ is the sigmoid activation function applied element-wise.

a) Suppose the weights $\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L-1)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *last* layer:

0 1 2 3 4

$$\arg\min_{\mathbf{W}^{(L)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(L)}$ of this optimization problem in closed-form ? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(L)}$. If no, explain how we can approximately solve this problem in practice.

b) Suppose the weights $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *first* layer:

0 1 2 3 4

$$\arg\min_{\mathbf{W}^{(0)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(0)}$ of this optimization problem in closed-form? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(0)}$. If no, explain how we can approximately solve this problem in practice.

Problem 3: Linear Regression (Version C) **(8 credits)**

We want to perform regression on a dataset consisting of N samples $\mathbf{x}_i \in \mathbb{R}^D$ ($D > 1$) with corresponding targets $y_i \in \mathbb{R}$ (represented compactly as $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\mathbf{y} \in \mathbb{R}^N$).
Assume that we have a deep neural network with L layers such that $f(\mathbf{x}, \mathbb{W}) = \mathbf{W}^{(L)}\sigma\left(\mathbf{W}^{(L-1)} \dots \left(\sigma\left(\mathbf{W}^{(0)}\mathbf{x}\right)\right)\right)$ where $\mathbb{W} = \{\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L)}\}$. Here, σ is the sigmoid activation function applied element-wise.

0 1 2 3 4

a) Suppose the weights $\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L-1)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *last* layer:

$$\arg\min_{\mathbf{W}^{(L)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(L)}$ of this optimization problem in closed-form ? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(L)}$. If no, explain how we can approximately solve this problem in practice.

0 1 2 3 4

b) Suppose the weights $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *first* layer:

$$\arg\min_{\mathbf{W}^{(0)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(0)}$ of this optimization problem in closed-form? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(0)}$. If no, explain how we can approximately solve this problem in practice.

Problem 3: Linear Regression (Version D) **(8 credits)**

We want to perform regression on a dataset consisting of N samples $\mathbf{x}_i \in \mathbb{R}^D$ ($D > 1$) with corresponding targets $y_i \in \mathbb{R}$ (represented compactly as $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\mathbf{y} \in \mathbb{R}^N$).
Assume that we have a deep neural network with L layers such that $f(\mathbf{x}, \mathbb{W}) = \mathbf{W}^{(L)}\sigma\left(\mathbf{W}^{(L-1)} \dots \left(\sigma\left(\mathbf{W}^{(0)}\mathbf{x}\right)\right)\right)$ where $\mathbb{W} = \{\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L)}\}$. Here, σ is the sigmoid activation function applied element-wise.

a) Suppose the weights $\mathbf{W}^{(0)}, \dots, \mathbf{W}^{(L-1)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *last* layer:

0 1 2 3 4

$$\arg\min_{\mathbf{W}^{(L)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(L)}$ of this optimization problem in closed-form ? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(L)}$. If no, explain how we can approximately solve this problem in practice.

b) Suppose the weights $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)}$ are non-zero, known and fixed and we want to solve the following optimization problem over the weights of the *first* layer:

0 1 2 3 4

$$\arg\min_{\mathbf{W}^{(0)}} \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i, \mathbb{W}) - y_i)^2$$

Can you compute the optimal weights $\mathbf{W}^{*(0)}$ of this optimization problem in closed-form? Justify your response. If yes, provide the closed-form solution $\mathbf{W}^{*(0)}$. If no, explain how we can approximately solve this problem in practice.

Problem 4: Linear classification (Version A) (9 credits)

A friend of yours owns an ice cream truck. To improve his profitability, he wants to predict whether it is profitable for him to drive to the lake and sell ice cream. For this reason, he created a dataset about his days at the lake. Whenever your friend was not able to collect the data, he noted down a "-". The last row represents the data for today.

x_1 Rain probability	x_2 Temperature forecast	x_3 # Visitors at 9 am	x_4 Public holiday	x_5 Weekday	y Profitable day
10%	25.1	51	True	False	True
20%	-	2	False	True	True
90%	37.8	5	False	False	False
-	23.0	27	True	True	True
45%	-2.1	21	True	True	False
0%	-10.2	18	True	False	?

0 1 2

a)
Your friend is a big fan of logistic regression. Can you name a reason why logistic regression might not be such a good idea here? Justify your answer.

Missing data // too much feature, dimensionality curse // $w^T x$ $\begin{pmatrix} x_1 \\ x_2 \\ x \\ x_5 \end{pmatrix}$ $\in \mathbb{R}^5$

0 1 2 3 4

b)
You, finally, convinced your friend to use a Naïve Bayes classifier. However, he is unsure which distribution is a good choice for each feature. Which of the following closs-conditional distributions should he use for each of the features x_1, x_2, x_3, x_4 and x_5?: (a) Bernoulli, (b) Normal distribution, (c) Poisson distribution, (d) Beta distribution, and (e) exponential distribution. Justify your answer.

d b c a a

[0,1] prob cont. Int Rolle

0 1 2 3

c)
After you have fitted the model, your friend is still not happy with the predictive performance of the model. Name a reason why the dataset might be not ideal for a Naïve Bayes model (assuming the data follows perfectly the chosen distributions). Justify your answer.

feature correlated not i.i.d data

Problem 4: Linear classification (Version B) (9 credits)

A friend of yours owns an ice cream truck. To improve his profitability, he wants to predict whether it is profitable for him to drive to the lake and sell ice cream. For this reason, he created a dataset about his days at the lake. Whenever your friend was not able to collect the data, he noted down a "-". The last row represents the data for today.

x_1 Rain probability	x_2 Temperature forecast	x_3 # Visitors at 9 am	x_4 Public holiday	x_5 Weekday	y Profitable day
10%	25.1	51	True	False	True
20%	-	2	False	True	True
90%	37.8	5	False	False	False
-	23.0	27	True	True	True
45%	-2.1	21	True	True	False
0%	-10.2	18	True	False	?

a)
Your friend is a big fan of logistic regression. Can you name a reason why logistic regression might not be such a good idea here? Justify your answer.

0
1
2

b)
You, finally, convinced your friend to use a Naïve Bayes classifier. However, he is unsure which distribution is a good choice for each feature. Which of the following closs-conditional distributions should he use for each of the features x_1, x_2, x_3, x_4 and x_5?: (a) Bernoulli, (b) Normal distribution, (c) Poisson distribution, (d) Beta distribution, and (e) exponential distribution. Justify your answer.

0
1
2
3
4

c)
After you have fitted the model, your friend is still not happy with the predictive performance of the model. Name a reason why the dataset might be not ideal for a Naïve Bayes model (assuming the data follows perfectly the chosen distributions). Justify your answer.

0
1
2
3

Problem 4: Linear classification (Version C) **(9 credits)**

A friend of yours owns an ice cream truck. To improve his profitability, he wants to predict whether it is profitable for him to drive to the lake and sell ice cream. For this reason, he created a dataset about his days at the lake. Whenever your friend was not able to collect the data, he noted down a "-". The last row represents the data for today.

x_1 Rain probability	x_2 Temperature forecast	x_3 # Visitors at 9 am	x_4 Public holiday	x_5 Weekday	y Profitable day
10%	25.1	51	True	False	True
20%	-	2	False	True	True
90%	37.8	5	False	False	False
-	23.0	27	True	True	True
45%	-2.1	21	True	True	False
0%	-10.2	18	True	False	?

0 1 2

a)
Your friend is a big fan of logistic regression. Can you name a reason why logistic regression might not be such a good idea here? Justify your answer.

0 1 2 3 4

b)
You, finally, convinced your friend to use a Naïve Bayes classifier. However, he is unsure which distribution is a good choice for each feature. Which of the following closs-conditional distributions should he use for each of the features x_1, x_2, x_3, x_4 and x_5?: (a) Bernoulli, (b) Normal distribution, (c) Poisson distribution, (d) Beta distribution, and (e) exponential distribution. Justify your answer.

0 1 2 3

c)
After you have fitted the model, your friend is still not happy with the predictive performance of the model. Name a reason why the dataset might be not ideal for a Naïve Bayes model (assuming the data follows perfectly the chosen distributions). Justify your answer.

Problem 4: Linear classification (Version D) (9 credits)

A friend of yours owns an ice cream truck. To improve his profitability, he wants to predict whether it is profitable for him to drive to the lake and sell ice cream. For this reason, he created a dataset about his days at the lake. Whenever your friend was not able to collect the data, he noted down a "-". The last row represents the data for today.

x_1 Rain probability	x_2 Temperature forecast	x_3 # Visitors at 9 am	x_4 Public holiday	x_5 Weekday	y Profitable day
10%	25.1	51	True	False	True
20%	-	2	False	True	True
90%	37.8	5	False	False	False
-	23.0	27	True	True	True
45%	-2.1	21	True	True	False
0%	-10.2	18	True	False	?

a)
Your friend is a big fan of logistic regression. Can you name a reason why logistic regression might not be such a good idea here? Justify your answer.

0
1
2

b)
You, finally, convinced your friend to use a Naïve Bayes classifier. However, he is unsure which distribution is a good choice for each feature. Which of the following closs-conditional distributions should he use for each of the features x_1, x_2, x_3, x_4 and x_5?: (a) Bernoulli, (b) Normal distribution, (c) Poisson distribution, (d) Beta distribution, and (e) exponential distribution. Justify your answer.

0
1
2
3
4

c)
After you have fitted the model, your friend is still not happy with the predictive performance of the model. Name a reason why the dataset might be not ideal for a Naïve Bayes model (assuming the data follows perfectly the chosen distributions). Justify your answer.

0
1
2
3

Problem 5: Optimization – Convexity (Version A) (8 credits)

Consider the two functions

$$f_a(\mathbf{x}) = \max\left[0, (\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a})\right]$$

PSD

$$f_b(\mathbf{x}) = 2^{\max_i |\mathbf{x}_i - \mathbf{a}_i|}$$

with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and positive semidefinite $\mathbf{A} \in \mathbb{R}^{n\times n}$.
For your reference, here are the convexity-preserving rules from the lecture. Let $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex functions, and $g : \mathbb{R}^d \to \mathbb{R}$ be a concave function, then:

1. $h(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ is convex
2. $h(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$ is convex
3. $h(\boldsymbol{x}) = c \cdot f_1(\boldsymbol{x})$ is convex if $c \geq 0$
4. $h(\boldsymbol{x}) = c \cdot g(\boldsymbol{x})$ is convex if $c \leq 0$
5. $h(\boldsymbol{x}) = f_1(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is convex ($\boldsymbol{A}$ matrix, $\boldsymbol{b}$ vector)
6. $h(\boldsymbol{x}) = m(f_1(\boldsymbol{x}))$ is convex if $m : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing.

$(x^TA - a^TA)(x-a)$

$\overbrace{x^TAx}\ \underbrace{-2x^TAa + a^TAa.}_{\text{linear}}$

$(x^TAx)'' > 0$

convex .

0 1 2 3 4

a)
Prove or disprove that $f_a(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

$|x|$ convex can we denote?)

convex $\dfrac{|x-a|}{Ax+b.}$ ✓

0 1 2 3 4

b)
Prove or disprove that $f_b(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

Problem 5: Optimization – Convexity (Version B) (8 credits)

Consider the two functions

$$f_a(\mathbf{x}) = \max\left[0, (\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a})\right]$$

$$f_b(\mathbf{x}) = 2^{\max_i |\mathbf{x}_i - \mathbf{a}_i|}$$

with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and positive semidefinite $\mathbf{A} \in \mathbb{R}^{n\times n}$.
For your reference, here are the convexity-preserving rules from the lecture. Let $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex functions, and $g : \mathbb{R}^d \to \mathbb{R}$ be a concave function, then:

1. $h(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ is convex
2. $h(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$ is convex
3. $h(\boldsymbol{x}) = c \cdot f_1(\boldsymbol{x})$ is convex if $c \geq 0$
4. $h(\boldsymbol{x}) = c \cdot g(\boldsymbol{x})$ is convex if $c \leq 0$
5. $h(\boldsymbol{x}) = f_1(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is convex ($\boldsymbol{A}$ matrix, $\boldsymbol{b}$ vector)
6. $h(\boldsymbol{x}) = m(f_1(\boldsymbol{x}))$ is convex if $m : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing.

a)
Prove or disprove that $f_a(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

0
1
2
3
4

b)
Prove or disprove that $f_b(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

0
1
2
3
4

Problem 5: Optimization – Convexity (Version C) **(8 credits)**

Consider the two functions

$$f_a(\mathbf{x}) = \max\left[0, (\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a})\right]$$

$$f_b(\mathbf{x}) = 2^{\max_i |\mathbf{x}_i - \mathbf{a}_i|}$$

with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and positive semidefinite $\mathbf{A} \in \mathbb{R}^{n\times n}$.
For your reference, here are the convexity-preserving rules from the lecture. Let $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex functions, and $g : \mathbb{R}^d \to \mathbb{R}$ be a concave function, then:

1. $h(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ is convex
2. $h(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$ is convex
3. $h(\boldsymbol{x}) = c \cdot f_1(\boldsymbol{x})$ is convex if $c \geq 0$
4. $h(\boldsymbol{x}) = c \cdot g(\boldsymbol{x})$ is convex if $c \leq 0$
5. $h(\boldsymbol{x}) = f_1(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is convex ($\boldsymbol{A}$ matrix, $\boldsymbol{b}$ vector)
6. $h(\boldsymbol{x}) = m(f_1(\boldsymbol{x}))$ is convex if $m : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing.

0 1 2 3 4

a)
Prove or disprove that $f_a(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

0 1 2 3 4

b)
Prove or disprove that $f_b(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

Problem 5: Optimization – Convexity (Version D) (8 credits)

Consider the two functions

$$f_a(\mathbf{x}) = \max\left[0, (\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a})\right]$$

$$f_b(\mathbf{x}) = 2^{\max_i |\mathbf{x}_i - \mathbf{a}_i|}$$

with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and positive semidefinite $\mathbf{A} \in \mathbb{R}^{n\times n}$.

For your reference, here are the convexity-preserving rules from the lecture. Let $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex functions, and $g : \mathbb{R}^d \to \mathbb{R}$ be a concave function, then:

1. $h(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ is convex
2. $h(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$ is convex
3. $h(\boldsymbol{x}) = c \cdot f_1(\boldsymbol{x})$ is convex if $c \geq 0$
4. $h(\boldsymbol{x}) = c \cdot g(\boldsymbol{x})$ is convex if $c \leq 0$
5. $h(\boldsymbol{x}) = f_1(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is convex ($\boldsymbol{A}$ matrix, $\boldsymbol{b}$ vector)
6. $h(\boldsymbol{x}) = m(f_1(\boldsymbol{x}))$ is convex if $m : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing.

a)

Prove or disprove that $f_a(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

0 1 2 3 4

b)

Prove or disprove that $f_b(\mathbf{x})$ is convex in $\mathbf{x}$. If you use a convexity preserving operation (stated above), clearly refer to the used rule. Any intermediate steps that are not direct applications of the given convexity-preserving operations must be proven.

0 1 2 3 4

Problem 6: Deep learning (Version A) (8 credits)

0 1 2 3 4 5 6 7 8

Given the input $x \in \mathbb{R}$, output $y \in \mathbb{R}$, and learnable parameters $\alpha, \beta \in \mathbb{R}$ we define the following four models:

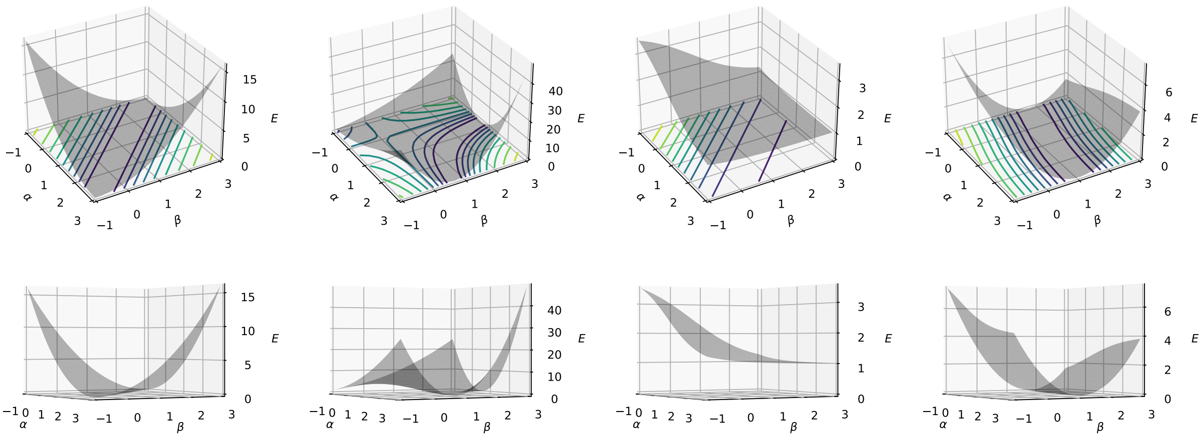
1. $f(x) = \sigma(\alpha x + \beta)$ $f(1) = \sigma(\alpha 1 + \beta)$ $\alpha = 1$ $\beta = 3$ 1
2. $f(x) = \alpha\beta x$ $f(1) = \alpha\beta$ 9
3. $f(x) = \alpha x + \beta$ $f(1) = \alpha + \beta$ 6
4. $f(x) = \sigma(\alpha x) + \beta$ $f(1) = \sigma(\alpha) + \beta$ 4

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Additionally, for the true output y and the model output $f(x)$, we define the error as $E = (y - f(x))^2$.

In the four plots below, we **fix the input and output** to $x = 1$ and $y = 2$ and plot the error curve of each model (1.-4.) as a function of α and β.

Connect the plots (a)-(d) with the models 1.-4. Justify your answer.

NOTE: The plots on the top and the bottom row are the same but shown at a different angle. For example, the two plots in (a) show the same error curve which corresponds to one of the models. The contour plot that spans the α-β plane shows the projected error contours. Darker colors denote lower values and the values along a single contour are constant.



(a) 3 (b) 2 (c) 1 (d) 4

Problem 6: Deep learning (Version B) (8 credits)

0 1 2 3 4 5 6 7 8

Given the input $x \in \mathbb{R}$, output $y \in \mathbb{R}$, and learnable parameters $\alpha, \beta \in \mathbb{R}$ we define the following four models:

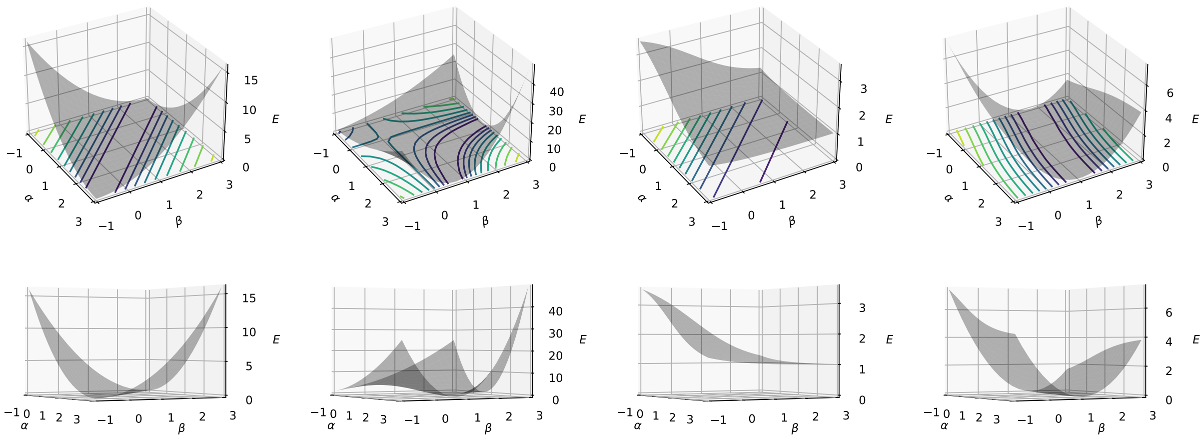
1. $f(x) = \sigma(\alpha x) + \beta$
2. $f(x) = \alpha\beta x$
3. $f(x) = \alpha x + \beta$
4. $f(x) = \sigma(\alpha x + \beta)$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Additionally, for the true output y and the model output $f(x)$, we define the error as $E = (y - f(x))^2$.

In the four plots below, we **fix the input and output** to $x = 1$ and $y = 2$ and plot the error curve of each model (1.-4.) as a function of α and β.

Connect the plots (a)-(d) with the models 1.-4. Justify your answer.

NOTE: The plots on the top and the bottom row are the same but shown at a different angle. For example, the two plots in (a) show the same error curve which corresponds to one of the models. The contour plot that spans the α-β plane shows the projected error contours. Darker colors denote lower values and the values along a single contour are constant.



(a) (b) (c) (d)

Problem 6: Deep learning (Version C) (8 credits)

0 1 2 3 4 5 6 7 8

Given the input $x \in \mathbb{R}$, output $y \in \mathbb{R}$, and learnable parameters $\alpha, \beta \in \mathbb{R}$ we define the following four models:

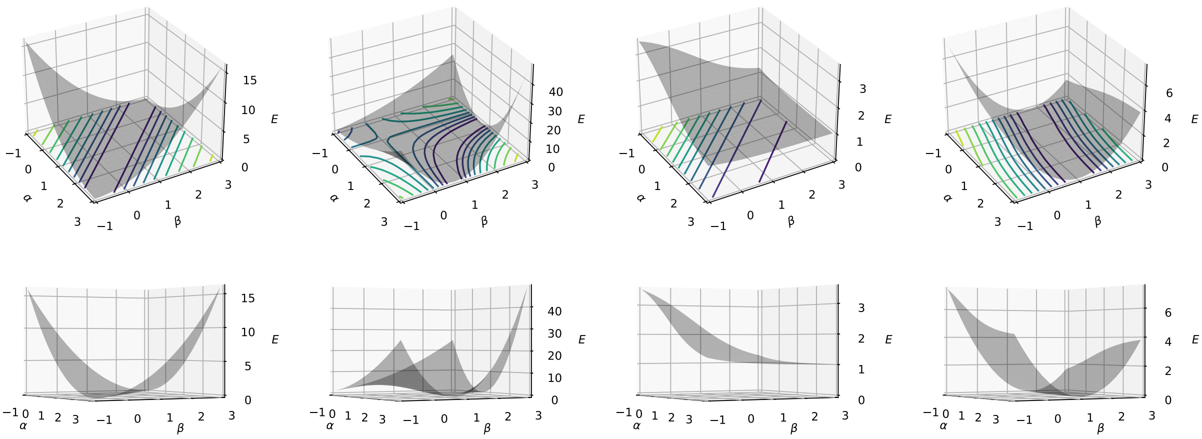
1. $f(x) = \sigma(\alpha x + \beta)$
2. $f(x) = \sigma(\alpha x) + \beta$
3. $f(x) = \alpha \beta x$
4. $f(x) = \alpha x + \beta$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Additionally, for the true output y and the model output $f(x)$, we define the error as $E = (y - f(x))^2$.

In the four plots below, we **fix the input and output** to $x = 1$ and $y = 2$ and plot the error curve of each model (1.-4.) as a function of α and β.

Connect the plots (a)-(d) with the models 1.-4. Justify your answer.

NOTE: The plots on the top and the bottom row are the same but shown at a different angle. For example, the two plots in (a) show the same error curve which corresponds to one of the models. The contour plot that spans the α-β plane shows the projected error contours. Darker colors denote lower values and the values along a single contour are constant.



(a) (b) (c) (d)

Problem 6: Deep learning (Version D) (8 credits)

0 1 2 3 4 5 6 7 8

Given the input $x \in \mathbb{R}$, output $y \in \mathbb{R}$, and learnable parameters $\alpha, \beta \in \mathbb{R}$ we define the following four models:

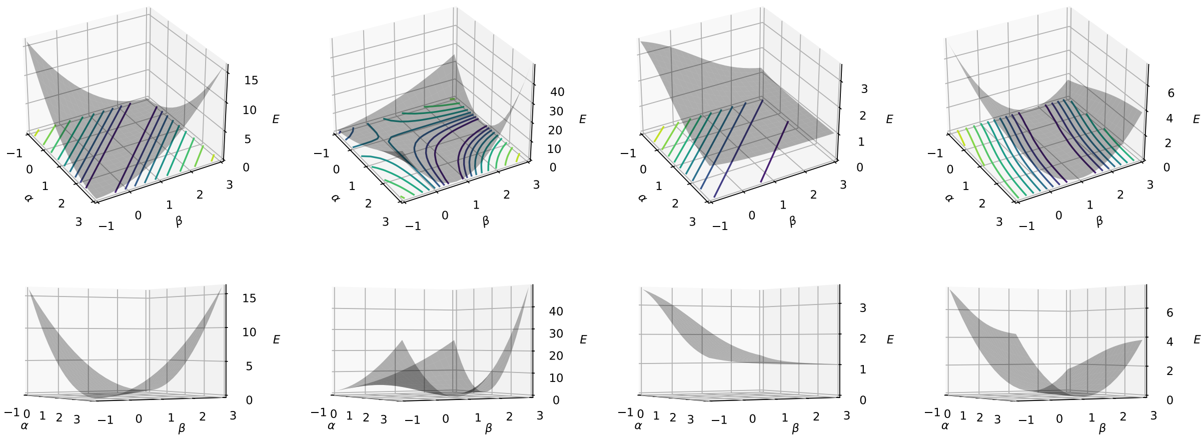
1. $f(x) = \sigma(\alpha x + \beta)$
2. $f(x) = \alpha x + \beta$
3. $f(x) = \sigma(\alpha x) + \beta$
4. $f(x) = \alpha\beta x$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Additionally, for the true output y and the model output $f(x)$, we define the error as $E = (y - f(x))^2$.

In the four plots below, we **fix the input and output** to $x = 1$ and $y = 2$ and plot the error curve of each model (1.-4.) as a function of α and β.

Connect the plots (a)-(d) with the models 1.-4. Justify your answer.

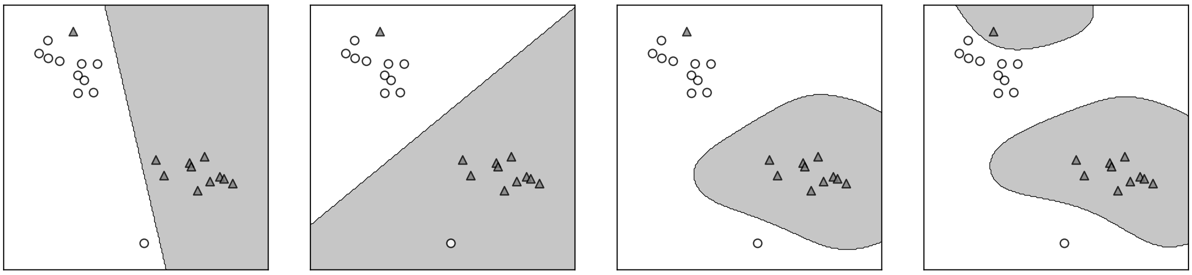
NOTE: The plots on the top and the bottom row are the same but shown at a different angle. For example, the two plots in (a) show the same error curve which corresponds to one of the models. The contour plot that spans the α-β plane shows the projected error contours. Darker colors denote lower values and the values along a single contour are constant.



(a) (b) (c) (d)

Problem 7: Support Vector Machines (Version A) **(4 credits)**

Below are shown four different decision boundaries plotted for four different models trained on the same dataset. Connect the decision boundaries (a)-(d) with the models 1-4. Justify your answer.



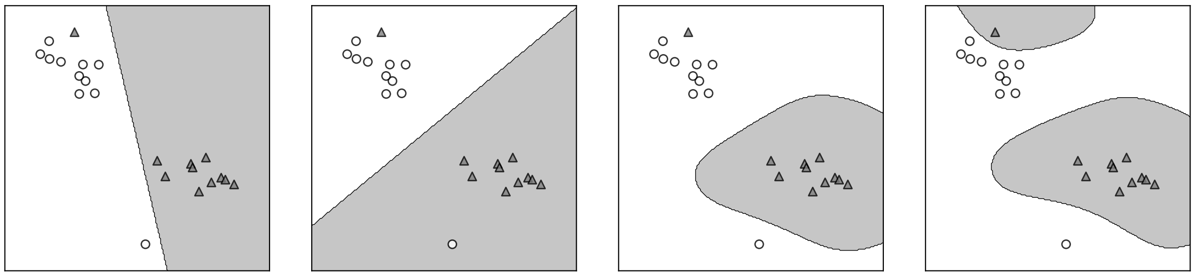
(a) 1 (b) 3 (c) 2 (d) 4.

1. `LogisticRegression(lambda=1.0, penalty='l2')`
2. `SVM(C=0.01, kernel='rbf')` → non-linear
3. `SVM(C=0.01, kernel='linear')` → maximize margin
4. `SVM(C=1000, kernel='rbf')` → hard SVM

Note: `rbf` *denotes radial basis (Gaussian) kernel.* `kernel='linear'` *corresponds to the plain SVM discussed in Section* 2 *of lecture 9. Parameters* `lambda` *and* `C` *appear in the loss for the respective models, as defined in the lecture.*

Problem 7: Support Vector Machines (Version B) **(4 credits)**

Below are shown four different decision boundaries plotted for four different models trained on the same dataset. Connect the decision boundaries (a)-(d) with the models 1-4. Justify your answer.



(a) (b) (c) (d)

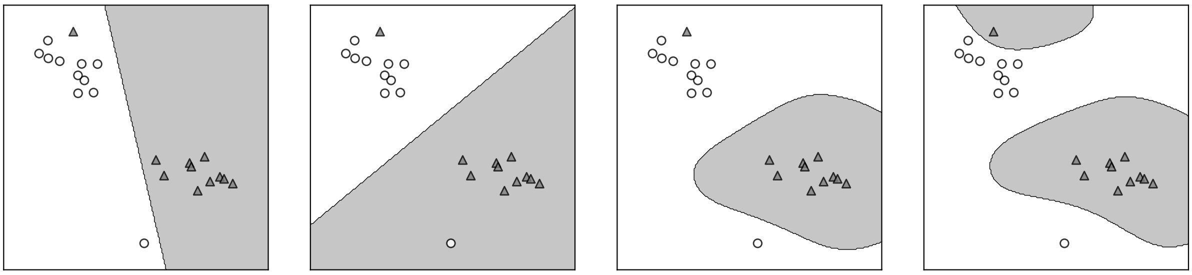
1. `SVM(C=0.01, kernel='rbf')`
2. `LogisticRegression(lambda=1.0, penalty='l2')`
3. `SVM(C=1000, kernel='rbf')`
4. `SVM(C=0.01, kernel='linear')`

Note: `rbf` *denotes radial basis (Gaussian) kernel.* `kernel='linear'` *corresponds to the plain SVM discussed in Section* 2 *of lecture 9. Parameters* `lambda` *and* `C` *appear in the loss for the respective models, as defined in the lecture.*

Problem 7: Support Vector Machines (Version C) **(4 credits)**

0 1 2 3 4

Below are shown four different decision boundaries plotted for four different models trained on the same dataset. Connect the decision boundaries (a)-(d) with the models 1-4. Justify your answer.



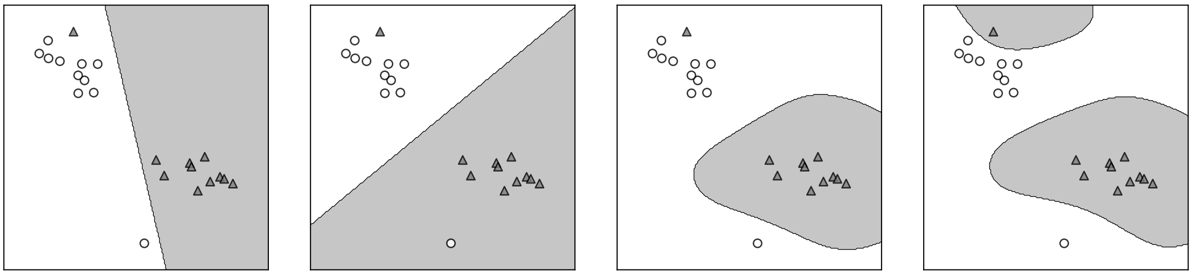
(a) (b) (c) (d)

1. `SVM(C=1000, kernel='rbf')`
2. `SVM(C=0.01, kernel='linear')`
3. `LogisticRegression(lambda=1.0, penalty='l2')`
4. `SVM(C=0.01, kernel='rbf')`

Note: `rbf` *denotes radial basis (Gaussian) kernel.* `kernel='linear'` *corresponds to the plain SVM discussed in Section* 2 *of lecture 9. Parameters* `lambda` *and* `C` *appear in the loss for the respective models, as defined in the lecture.*

Problem 7: Support Vector Machines (Version D) **(4 credits)**

Below are shown four different decision boundaries plotted for four different models trained on the same dataset. Connect the decision boundaries (a)-(d) with the models 1-4. Justify your answer.



(a) (b) (c) (d)

1. `LogisticRegression(lambda=1.0, penalty='l2')`
2. `SVM(C=0.01, kernel='rbf')`
3. `SVM(C=1000, kernel='rbf')`
4. `SVM(C=0.01, kernel='linear')`

Note: `rbf` *denotes radial basis (Gaussian) kernel.* `kernel='linear'` *corresponds to the plain SVM discussed in Section* 2 *of lecture 9. Parameters* `lambda` *and* `C` *appear in the loss for the respective models, as defined in the lecture.*

Problem 8: Kernels (Version A) (8 credits)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a neural network with two hidden layers and ReLU activations between them. Let $p(\mathbf{x})$ denote the probability density function on $\mathbb{R}^N$. Prove or disprove that the following functions are valid kernels.

0 1 2 3 4

a) $k(\mathbf{x}, \mathbf{y}) = \underbrace{\sum_i^M f(\mathbf{x})_i}_{\phi(x_i) \atop X} + \underbrace{\sum_i^M f(\mathbf{y})_i}_{\phi(y_i) \atop Y}$

$$K = \begin{pmatrix} X+X & X+Y \\ Y+X & Y+Y \end{pmatrix} < 0$$

0 1 2 3 4

b) $k(\mathbf{x}, \mathbf{y}) = \log \underbrace{p(\mathbf{x})}_{\phi(x)} \log \underbrace{p(\mathbf{y})}_{\phi(y)}$

Problem 8: Kernels (Version B) (8 credits)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and let $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ be a neural network with two hidden layers and ReLU activations between them. Let $p(\mathbf{x})$ denote the probability density function on $\mathbb{R}^N$. Prove or disprove that the following functions are valid kernels.

a) $k(\mathbf{x}, \mathbf{y}) = \left(\sum_i^M f(\mathbf{x})_i\right) \left(\sum_i^M f(\mathbf{y})_i\right)$

0
1
2
3
4

b) $k(\mathbf{x}, \mathbf{y}) = \sum_z p(\mathbf{x}|z)p(\mathbf{y}|z)p(z)$, where $z \in \{1, \ldots, Z\}$

0
1
2
3
4

Problem 8: Kernels (Version C) (8 credits)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a neural network with two hidden layers and ReLU activations between them. Let $p(\mathbf{x})$ denote the probability density function on $\mathbb{R}^N$. Prove or disprove that the following functions are valid kernels.

0 1 2 3 4

a) $k(\mathbf{x}, \mathbf{y}) = \sum_i^M f(\mathbf{x})_i + \sum_i^M f(\mathbf{y})_i$

0 1 2 3 4

b) $k(\mathbf{x}, \mathbf{y}) = \sum_z p(\mathbf{x}|z)p(\mathbf{y}|z)p(z)$, where $z \in \{1, \ldots, Z\}$

Problem 8: Kernels (Version D) (8 credits)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a neural network with two hidden layers and ReLU activations between them. Let $p(\mathbf{x})$ denote the probability density function on $\mathbb{R}^N$. Prove or disprove that the following functions are valid kernels.

a) $k(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}) \log p(\mathbf{y})$

0 1 2 3 4

b) $k(\mathbf{x}, \mathbf{y}) = \sum_i^M f(\mathbf{x})_i + \sum_i^M f(\mathbf{y})_i$

0 1 2 3 4

Problem 9: Dimensionality Reduction (Version A) **(6 credits)**

Consider the following centered data matrix, where the data points are stored as rows

$$X = \begin{pmatrix} 0 & -5 & 0 & -2 \\ 0 & 0 & 2 & 6 \\ 2 & 1 & 0 & -2 \\ 0 & 0 & -2 & 6 \\ 1 & 1 & -1 & -2 \\ -2 & 1 & 0 & -2 \\ -2 & 1 & 0 & -2 \\ 0 & -1 & 0 & 0 \\ 1 & 1 & 1 & -2 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

We performed eigenvalue decomposition of the respective covariance matirix $\mathbf{\Sigma_X} = \mathbf{\Gamma}^T \mathbf{\Lambda} \mathbf{\Gamma}$. We obtained the following eigenvalue matrix

$$\Lambda = \begin{pmatrix} 9.6 & 0 & 0 & 0 \\ 0 & 3.2 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and the following eigenvector matrix

$$\Gamma = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

where the eigenvectors are stored as columns and sorted according to the eigenvalues in decreasing order.

0
1
2
3

a) Compute the coordinates of all data points when projected onto the subspace spanned by the top-2 principal components (i.e. the eigenvectors with the two largest eigenvalues). Justify your answer.

$$Y = \hat{X} \cdot \Gamma = \left(\begin{array}{cc} -2 & -8 \\ 6 & 0 \\ -2 & 1 \\ & 1 \\ 1 & 1 \\ 1 & \end{array} \right.$$

0
1
2
3

b) Now, suppose that we projected the original data points on the subspace defined by the THIRD principal component (i.e. the eigenvector with the third-largest eigenvalues). Compute the variance of the projected data. Justify your answer.

$$\Sigma_{\hat{X}} = \Gamma \Lambda \Gamma^T = \sim$$

$$\Sigma_{\hat{X}}' = \Sigma_X \cdot \Gamma$$

Problem 9: Dimensionality Reduction (Version B) **(6 credits)**

Consider the following centered data matrix, where the data points are stored as rows

$$\begin{pmatrix} 0 & 2 & 0.5 & -1 \\ 2 & 1 & -1 & 1 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & 3.5 & 1 \\ -2 & 0 & 0 & -1 \\ 1 & -1 & 0 & -1 \\ -2 & 1 & -2 & 1 \\ -1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

We performed eigenvalue decomposition of the respective covariance matirix $\mathbf{\Sigma_X} = \mathbf{\Gamma}^T \mathbf{\Lambda} \mathbf{\Gamma}$. We obtained the following eigenvalue matrix

$$\begin{pmatrix} 2.05 & 0 & 0 & 0 \\ 0 & 1.6 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 0.8 \end{pmatrix}$$

and the following eigenvector matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where the eigenvectors are stored as columns and sorted according to the eigenvalues in decreasing order.

a) Compute the coordinates of all data points when projected onto the subspace spanned by the top-2 principal components (i.e. the eigenvectors with the two largest eigenvalues). Justify your answer.

0 1 2 3

b) Now, suppose that we projected the original data points on the subspace defined by the THIRD principal component (i.e. the eigenvector with the third-largest eigenvalues). Compute the variance of the projected data. Justify your answer.

0 1 2 3

Problem 9: Dimensionality Reduction (Version C) (6 credits)

Consider the following centered data matrix, where the data points are stored as rows

$$\begin{pmatrix} 1 & 1 & 0 & -4 \\ 2 & 2 & 1 & 2 \\ 0 & -4.5 & 0 & 1 \\ 0 & 1.5 & -2 & 1 \\ 1 & -2 & 0 & -1 \\ 1 & 0 & 0 & 1 \\ -2 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & -2 \\ -1 & 1 & 0 & 1 \end{pmatrix}$$

We performed eigenvalue decomposition of the respective covariance matirix $\mathbf{\Sigma_X} = \mathbf{\Gamma}^T \mathbf{\Lambda} \mathbf{\Gamma}$. We obtained the following eigenvalue matrix

$$\begin{pmatrix} 3.35 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 0.6 \end{pmatrix}$$

and the following eigenvector matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

where the eigenvectors are stored as columns and sorted according to the eigenvalues in decreasing order.

0 1 2 3

a) Compute the coordinates of all data points when projected onto the subspace spanned by the top-2 principal components (i.e. the eigenvectors with the two largest eigenvalues). Justify your answer.

0 1 2 3

b) Now, suppose that we projected the original data points on the subspace defined by the THIRD principal component (i.e. the eigenvector with the third-largest eigenvalues). Compute the variance of the projected data. Justify your answer.

Problem 9: Dimensionality Reduction (Version D) (6 credits)

Consider the following centered data matrix, where the data points are stored as rows

$$\begin{pmatrix} 0 & 3 & 0 & 3 \\ -1 & 1 & 1 & 1 \\ 1 & -1 & 4 & 3 \\ 1 & 1 & 1 & -7 \\ -2 & -1 & 1 & -2 \\ 1 & -1 & -2 & 1 \\ 0 & -2 & -1 & 1 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 \end{pmatrix}$$

We performed eigenvalue decomposition of the respective covariance matirix $\mathbf{\Sigma_X} = \mathbf{\Gamma}^T \mathbf{\Lambda} \mathbf{\Gamma}$. We obtained the following eigenvalue matrix

$$\begin{pmatrix} 7.6 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0.8 \end{pmatrix}$$

and the following eigenvector matrix

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

where the eigenvectors are stored as columns and sorted according to the eigenvalues in decreasing order.

a) Compute the coordinates of all data points when projected onto the subspace spanned by the top-2 principal components (i.e. the eigenvectors with the two largest eigenvalues). Justify your answer.

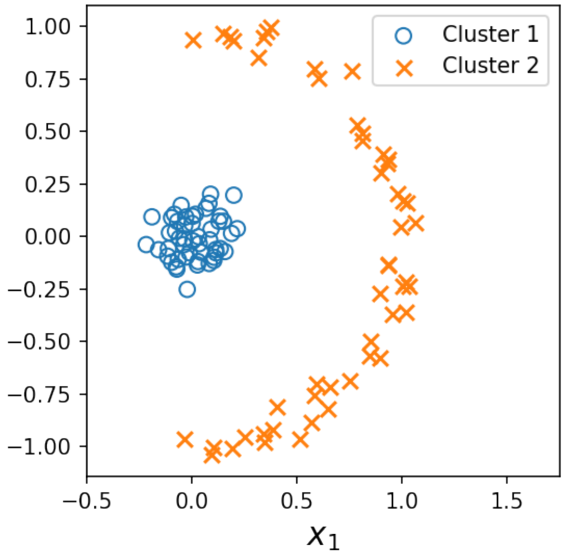
☐ 0
☐ 1
☐ 2
☐ 3

b) Now, suppose that we projected the original data points on the subspace defined by the THIRD principal component (i.e. the eigenvector with the third-largest eigenvalues). Compute the variance of the projected data. Justify your answer.

☐ 0
☐ 1
☐ 2
☐ 3

Problem 10: Clustering (Version A) (8 credits)

You would like to perform clustering on the following dataset.



0 1 2

a) Is K-means, in theory, able to recover the true clusters on the above dataset? Justify your answer.

No kmeans 2 clustre − linear boundary

0 1 2 3

b) Is GMM with FULL covariance matrix SHARED across all clusters, in theory, able to recover the true clusters on the above dataset? Justify your answer.

GMM with shared covarin matrix create linear boundary.

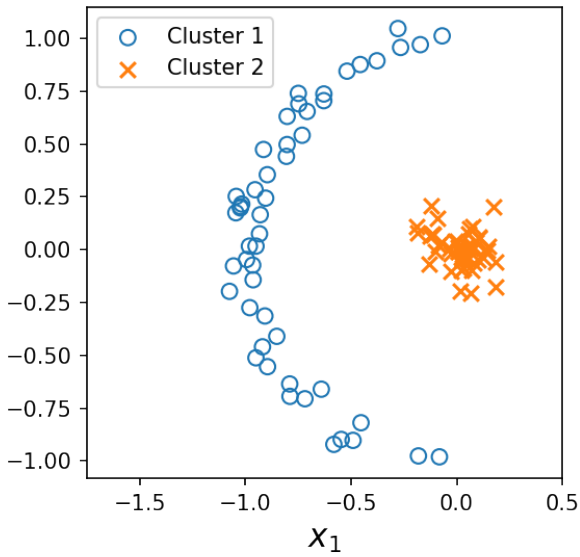
0 1 2 3

c) Is GMM with SEPARATE DIAGONAL covariance matrix for each cluster, in theory, able to recover the true clusters on the above dataset? Justify your answer.

different covariance

Problem 10: Clustering (Version B) (8 credits)

You would like to perform clustering on the following dataset.



a) Is K-means, in theory, able to recover the true clusters on the above dataset? Justify your answer.

0
1
2

b) Is GMM with FULL covariance matrix SHARED across all clusters, in theory, able to recover the true clusters on the above dataset? Justify your answer.

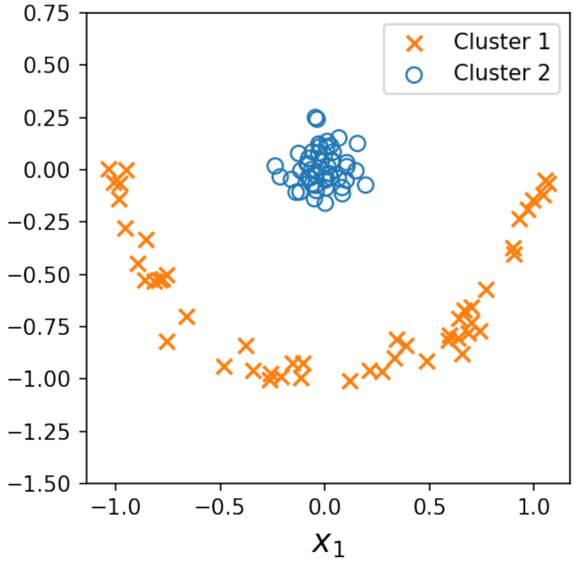
0
1
2
3

c) Is GMM with SEPARATE DIAGONAL covariance matrix for each cluster, in theory, able to recover the true clusters on the above dataset? Justify your answer.

0
1
2
3

Problem 10: Clustering (Version C) (8 credits)

You would like to perform clustering on the following dataset.



0 1 2

a) Is K-means, in theory, able to recover the true clusters on the above dataset? Justify your answer.

0 1 2 3

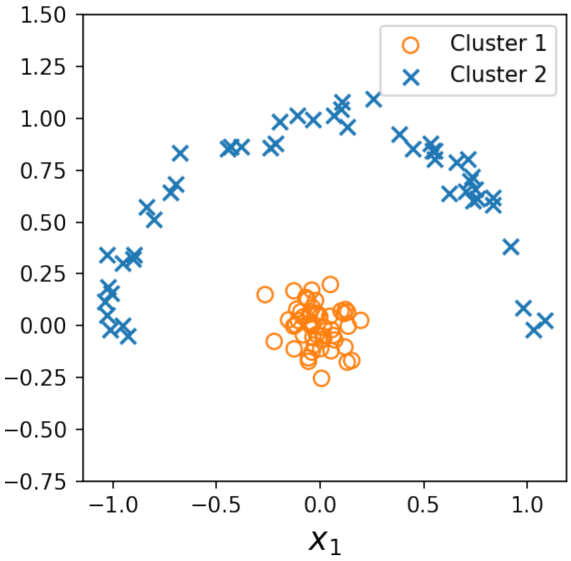
b) Is GMM with FULL covariance matrix SHARED across all clusters, in theory, able to recover the true clusters on the above dataset? Justify your answer.

0 1 2 3

c) Is GMM with SEPARATE DIAGONAL covariance matrix for each cluster, in theory, able to recover the true clusters on the above dataset? Justify your answer.

Problem 10: Clustering (Version D) **(8 credits)**

You would like to perform clustering on the following dataset.



a) Is K-means, in theory, able to recover the true clusters on the above dataset? Justify your answer.

☐ 0
☐ 1
☐ 2

b) Is GMM with FULL covariance matrix SHARED across all clusters, in theory, able to recover the true clusters on the above dataset? Justify your answer.

☐ 0
☐ 1
☐ 2
☐ 3

c) Is GMM with SEPARATE DIAGONAL covariance matrix for each cluster, in theory, able to recover the true clusters on the above dataset? Justify your answer.

☐ 0
☐ 1
☐ 2
☐ 3

Problem 11: Fairness (Version A) **(7 credits)**

Consider the following probabilistic model for generating a dataset consisting of non-sensitive features $\mathbf{x} \in \mathbb{R}^2$, sensitive features $A \in \{a, b\}$ and class labels $Y \in \{0, 1\}$:

$$\Pr[A = a] = \frac{1}{2}$$
$$\Pr\left[Y = 0 \mid A = a\right] = \Pr\left[Y = 0 \mid A = b\right] = \frac{1}{2}$$
$$p(\mathbf{x}|A = s, Y = c) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_s^c, \mathbf{1}_2)$$

with $\boldsymbol{\mu}_a^0, \boldsymbol{\mu}_a^1, \boldsymbol{\mu}_b^0, \boldsymbol{\mu}_b^1 \in \mathbb{R}^2$, where $\mathbf{1}_2$ is the 2-dimensional identitiy matrix.
We want to classify data generated by this model via a linear classifier $r(\mathbf{x}, A) = \mathbb{I}\left[\mathbf{w}^T\mathbf{x} + d > 0\right]$ with $\mathbf{w} \in \mathbb{R}^2$, $||\mathbf{w}||_2 > 0$ and $d \in \mathbb{R}$. Note that r only operates on the non-sensitive features.
In the following, let

$$\boldsymbol{\mu}_a^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_a^1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_b^0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_b^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

For the following subtasks, you may use the fact that $\Pr_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{1}_2)}\left[\mathbf{w}^T\mathbf{x} + d > 0\right] = \Phi\left(\frac{w^T\mu + d}{||w||_2}\right)$ for any $\mathbf{w} \in \mathbb{R}^2$ with $||w|| > 0$, $d \in \mathbb{R}$, $\boldsymbol{\mu} \in \mathbb{R}^2$, where Φ is the CDF of a standard normal distribution.

0 1 2

a) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *equality of opportunity* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure equality of opportunity and justify your response. If no, justify your response.

0 1 2

b) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *separation* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure separation and justify your response. If no, justify your response.

0 1 2 3

c) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *independence* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure independence and justify your response. If no, justify your response.

Problem 11: Fairness (Version B) (7 credits)

Consider the following probabilistic model for generating a dataset consisting of non-sensitive features $\mathbf{x} \in \mathbb{R}^2$, sensitive features $A \in \{a, b\}$ and class labels $Y \in \{0, 1\}$:

$$\Pr[A = a] = \frac{1}{2}$$
$$\Pr\left[Y = 0 \mid A = a\right] = \Pr\left[Y = 0 \mid A = b\right] = \frac{1}{2}$$
$$p(\mathbf{x}|A = s, Y = c) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_s^c, \mathbf{1}_2)$$

with $\boldsymbol{\mu}_a^0, \boldsymbol{\mu}_a^1, \boldsymbol{\mu}_b^0, \boldsymbol{\mu}_b^1 \in \mathbb{R}^2$, where $\mathbf{1}_2$ is the 2-dimensional identitiy matrix.
We want to classify data generated by this model via a linear classifier $r(\mathbf{x}, A) = \mathbb{I}\left[\mathbf{w}^T\mathbf{x} + d > 0\right]$ with $\mathbf{w} \in \mathbb{R}^2$, $||\mathbf{w}||_2 > 0$ and $d \in \mathbb{R}$. Note that r only operates on the non-sensitive features.
In the following, let

$$\boldsymbol{\mu}_a^0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_a^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_b^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_b^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For the following subtasks, you may use the fact that $\Pr_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{1}_2)}\left[\mathbf{w}^T\mathbf{x} + d > 0\right] = \Phi\left(\frac{w^T\boldsymbol{\mu}+d}{||w||_2}\right)$ for any $\mathbf{w} \in \mathbb{R}^2$ with $||w|| > 0$, $d \in \mathbb{R}$, $\boldsymbol{\mu} \in \mathbb{R}^2$, where Φ is the CDF of a standard normal distribution.

a) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *equality of opportunity* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure equality of opportunity and justify your response. If no, justify your response.

0
1
2

b) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *separation* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure separation and justify your response. If no, justify your response.

0
1
2

c) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *independence* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure independence and justify your response. If no, justify your response.

0
1
2
3

Problem 11: Fairness (Version C) **(7 credits)**

Consider the following probabilistic model for generating a dataset consisting of non-sensitive features $\mathbf{x} \in \mathbb{R}^2$, sensitive features $A \in \{a, b\}$ and class labels $Y \in \{0, 1\}$:

$$\Pr[A = a] = \frac{1}{2}$$
$$\Pr\left[Y = 0 \mid A = a\right] = \Pr\left[Y = 0 \mid A = b\right] = \frac{1}{2}$$
$$p(\mathbf{x}|A = s, Y = c) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_s^c, \mathbf{1}_2)$$

with $\boldsymbol{\mu}_a^0, \boldsymbol{\mu}_a^1, \boldsymbol{\mu}_b^0, \boldsymbol{\mu}_b^1 \in \mathbb{R}^2$, where $\mathbf{1}_2$ is the 2-dimensional identitiy matrix.
We want to classify data generated by this model via a linear classifier $r(\mathbf{x}, A) = \mathbb{I}\left[\mathbf{w}^T\mathbf{x} + d > 0\right]$ with $\mathbf{w} \in \mathbb{R}^2$, $||\mathbf{w}||_2 > 0$ and $d \in \mathbb{R}$. Note that r only operates on the non-sensitive features.
In the following, let

$$\boldsymbol{\mu}_a^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_a^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_b^0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_b^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For the following subtasks, you may use the fact that $\Pr_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{1}_2)}\left[\mathbf{w}^T\mathbf{x} + d > 0\right] = \Phi\left(\frac{w^T\mu+d}{||w||_2}\right)$ for any $\mathbf{w} \in \mathbb{R}^2$ with $||w|| > 0$, $d \in \mathbb{R}$, $\boldsymbol{\mu} \in \mathbb{R}^2$, where Φ is the CDF of a standard normal distribution.

0
1
2

a) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *equality of opportunity* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure equality of opportunity and justify your response. If no, justify your response.

0
1
2

b) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *separation* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure separation and justify your response. If no, justify your response.

0
1
2
3

c) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *independence* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure independence and justify your response. If no, justify your response.

Problem 11: Fairness (Version D) (7 credits)

Consider the following probabilistic model for generating a dataset consisting of non-sensitive features $\mathbf{x} \in \mathbb{R}^2$, sensitive features $A \in \{a, b\}$ and class labels $Y \in \{0, 1\}$:

$$\Pr[A = a] = \frac{1}{2}$$

$$\Pr\left[Y = 0 \mid A = a\right] = \Pr\left[Y = 0 \mid A = b\right] = \frac{1}{2}$$

$$p(\mathbf{x}|A = s, Y = c) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_s^c, \mathbf{1}_2)$$

with $\boldsymbol{\mu}_a^0, \boldsymbol{\mu}_a^1, \boldsymbol{\mu}_b^0, \boldsymbol{\mu}_b^1 \in \mathbb{R}^2$, where $\mathbf{1}_2$ is the 2-dimensional identitiy matrix.
We want to classify data generated by this model via a linear classifier $r(\mathbf{x}, A) = \mathbb{I}\left[\mathbf{w}^T\mathbf{x} + d > 0\right]$ with $\mathbf{w} \in \mathbb{R}^2$, $||\mathbf{w}||_2 > 0$ and $d \in \mathbb{R}$. Note that r only operates on the non-sensitive features.
In the following, let

$$\boldsymbol{\mu}_a^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_a^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_b^0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_b^1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

For the following subtasks, you may use the fact that $\Pr_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{1}_2)}\left[\mathbf{w}^T\mathbf{x} + d > 0\right] = \Phi\left(\frac{w^T\boldsymbol{\mu} + d}{||w||_2}\right)$ for any $\mathbf{w} \in \mathbb{R}^2$ with $||w|| > 0$, $d \in \mathbb{R}$, $\boldsymbol{\mu} \in \mathbb{R}^2$, where Φ is the CDF of a standard normal distribution.

a) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *equality of opportunity* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure equality of opportunity and justify your response. If no, justify your response.

0
1
2

b) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *separation* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure separation and justify your response. If no, justify your response.

0
1
2

c) For the above generative model, are there values for $\mathbf{w}$ and d with $||\mathbf{w}||_2 > 0$ such that the linear classifier $r(\mathbf{x}, A)$ fulfills the *independence* fairness criterion? If yes, specify values of $\mathbf{w}$ and d that ensure independence and justify your response. If no, justify your response.

0
1
2
3

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.