Data Analytics and Machine Learning Group
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.

Machine Learning

Exam:	IN2064 / Endterm	**Date:**	Saturday 11^{th} July, 2020
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	10:45 – 12:45

	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	P 9	P 10	P 11	P 12
I												

Working instructions

- This exam consists of **16 pages** with a total of **12 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 55 credits.
- Allowed resources:
  - all materials that you will use on your own (lecture slides, calculator etc.)
  - **not allowed are any forms of collaboration between examinees and plagiarism**
- Only write on the provided sheets, **submitting your own additional sheets is not possible**.
- Last three pages can be used as scratch paper.
- All sheets (including scratch paper) have to be submitted to the upload queue. Missing pages will be cosidered empty.
- **Only use a black or blue color (no red or green)!**
- Write your answers only in the provided solution boxes or the scratch paper.
- **For problems that say "Justify your answer" you only get points if you provide a valid explanation.**
- **For problems that say "Prove" you only get points if you provide a valid mathematical proof.**
- If a problem does not say "Justify your answer" or "Prove" it's sufficient to only provide the correct answer.
- Exam duration - 120 minutes.

Left room from __________ to __________ / Early submission at __________

Problem 1 KNN-Classification (4 credits)

0
1
2

a) Assume you use a KNN-classifier on the following training data, that contains at least 100 samples of each class.

PS	acceleration	max. velocity [km/h]	cylinder capacity [cm^3]	weight [kg]	class
150	12.5	178	1968	2001	van
600	3.6	250	3996	2150	car
113	3.5	200	937	227	motorcycle
...	...	...	...	...	...

You observe that the obtained model performs bad on the test set. What might be the problem? Name at least two possible problems and explain how you would solve them.

0
1
2

b) Would a decision tree have the same problems? Justify your answer.

Problem 2 **Overfitting (3 credits)**

0
1
2
3

Explain overfitting. When does it occur? Why is overfitting unwanted? How can we spot overfitting? How can we avoid it?

Problem 3 **Probabilistic inference (7 credits)**

0
1
2
3
4
5
6
7

Consider the following probabilistic model

$$p(\lambda \mid a, b) = \text{Gamma}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$

$$p(x \mid \lambda) = \text{Poisson}(x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

where $a \in (1, \infty)$ and $b \in (0, \infty)$. We have observed a single data point $x \in \mathbb{N}$. Derive the maximum a posteriori (MAP) estimate of the parameter λ for the above probabilistic model. Show your work.

Problem 4 Regression (5 credits)

a) Assume you train a linear regression model on dataset $D = \{\boldsymbol{x}_i, y_i\}_i$, $\boldsymbol{x}_i \in \mathbb{R}^D$, $y_i \in \mathbb{R}$ with the mean-square-error as loss function. After training is finished, you compute the MSE on individual data-points of the training-set. You notice that for three points you obtain a high MSE (1000 times higher than for the other points).
Evaluation on the test-set shows that your regression model does not perform that well. What might be the reason for that? How would you improve performance of your model? Justify your answer.

0
1
2
3

b) You want to train another linear regression model and decide to use the log-cosh-loss:

0
1
2

$$E_{lc} = \sum_i \log\cosh(\boldsymbol{w}^T \boldsymbol{x}_i - y_i)$$

How do you learn the parameter $\boldsymbol{w}$ of your model? Describe in one or two sentences.
Hint: $\cosh(z) = 0.5(e^z + e^{-z})$

Problem 5 Classification (4 credits)

We would like to design a generative classification model for the following data.
Each data point is represented by a D-dimensional feature vector $\boldsymbol{x} = (x_1, ..., x_D)$, where each entry x_j is a real number between 0 and 1, that is $x_j \in [0, 1]$ for $j = 1, ..., D$. Each data point belongs to one of $K > 2$ possible classes, that is $y \in \{1, ..., K\}$.

a) Which of the following distributions is the most reasonable choice for the class prior $p(y)$?

☐ Bernoulli ☐ Normal ☐ Beta ☐ Exponential ☐ Categorical

b) We decide to model the class conditional distribution $p(\boldsymbol{x}|y)$ as $p(\boldsymbol{x}|y) = \prod_{j=1}^{D} p(x_j|y)$, that is, the features x_j are conditionally independent give the class label y.

Which of the following distributions is the most reasonable choice for $p(x_j|y)$?

☐ Categorical ☐ Beta ☐ Normal ☐ Exponential ☐ Bernoulli

0 ☐
1 ☐
2 ☐

c) What is the name of the posterior distribution $p(y|\boldsymbol{x})$ for the model that you specified in subtasks (a) and (b)? Justify your answer.

Note that you need to provide the name of the distribution, not its probability density / mass functions. If the posterior distribution doesn't have a name, you should write "unknown distribution".

Problem 6 Alternative characterization of vertices (4 credits)

Consider a non-empty convex set $\mathcal{X} \subset \mathbb{R}^D$ and $\boldsymbol{x} \in \mathcal{X}$. Prove that if $\boldsymbol{x}$ is a vertex of $\mathcal{X}$ then $\mathcal{X} \setminus \{\boldsymbol{x}\}$ is convex.

0
1
2
3
4

Hint: additionally to the definition from the lecture you can use that $\boldsymbol{x} \in \mathcal{X}$ is a vertex of $\mathcal{X}$ if and only if for all $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathcal{X}$ with $\boldsymbol{x}_0 \neq \boldsymbol{x}_1$ and all $\lambda \in (0, 1)$ there holds that $\boldsymbol{x} \neq \boldsymbol{x}_\lambda$, where $\boldsymbol{x}_\lambda = \lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_0$ (i.e. $\boldsymbol{x}$ does not lie between two different points from $\mathcal{X}$).

Problem 7 Classification with Hinge loss and L_∞ penalty (7 credits)

For $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ and $y_1, \dots, y_N \in \{-1, 1\}$ consider the following optimization problem with a fixed parameter $\lambda > 0$ and $\|\mathbf{w}\|_\infty = \max\left(|w_1|, \dots, |w_D|\right)$.

$$\text{minimize}_{\mathbf{w},b} \quad \sum_{i=1}^{N} \max\left(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)\right) + \lambda \|\mathbf{w}\|_\infty \,. \tag{1}$$

a) In this task you have to choose all correct options. Problem (1) as formulated above is

- ☐ concave.
- ☐ a quadratic problem.
- ☐ unconstrained.
- ☐ not a quadratic problem.
- ☐ convex.
- ☐ constrained.
- ☐ a linear problem.
- ☐ a minimization problem.
- ☐ non-convex.

0 ☐
1 ☐
2 ☐
3 ☐
4 ☐
5 ☐

b) Reformulate problem (1) as an optimization problem with a linear objective and linear constraints. Justify your answer.

Hint: you can introduce new variables to the problem.

Problem 8 Deep learning (4 credits)

0
1
2
3
4

We are using a fully-connected neural network with 2 hidden layers for binary classification of points in $\mathbb{R}^D$

$$f(\boldsymbol{x}, \boldsymbol{W}) = \sigma_2(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\sigma_0(\boldsymbol{W}_0\boldsymbol{x}))).$$

where $\boldsymbol{W} = \{\boldsymbol{W}_0, \boldsymbol{W}_1, \boldsymbol{W}_2\}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{D_1 \times D}$, $\boldsymbol{W}_1 \in \mathbb{R}^{D_2 \times D_1}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times D_2}$ are the weights of the neural network.
The neural network outputs probabilities of the positive class, i.e. $p(y = 1|\boldsymbol{x}, \boldsymbol{W}) = f(\boldsymbol{x}, \boldsymbol{W})$, and is trained by minimizing the binary cross-entropy loss. We use the following activation functions:

$$\sigma_0(t) = t\sqrt{69} \qquad \sigma_1(t) = -\frac{t}{54\pi} \qquad \sigma_2(t) = \frac{1}{1 + \exp(-67t)}$$

The neural network achieves 100% classification accuracy on a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. Which of the following statements is true? Justify your answer.

1. $\mathcal{D}$ is linearly separable.
2. $\mathcal{D}$ is NOT linearly separable.
3. There is not enough information to determine if $\mathcal{D}$ is linearly separable.

Problem 9 **Principal Component Analysis (4 credits)**

Consider the data

$$\boldsymbol{X} = \begin{pmatrix} 0.37 & 0.95 & 0.73 & 0.60 \\ 0.16 & 0.16 & 0.06 & 0.87 \\ 0.60 & 0.71 & 0.02 & 0.97 \\ 0.83 & 0.21 & 0.18 & 0.18 \\ 0.30 & 0.52 & 0.43 & 0.29 \\ 0.61 & 0.14 & 0.29 & 0.37 \\ 0.46 & 0.79 & 0.20 & 0.51 \\ 0.59 & 0.05 & 0.61 & 0.17 \end{pmatrix},$$

where each row of $\boldsymbol{X}$ represents a sample.

0 1 2 3 4

In each of the following PCA solutions the first row of $\boldsymbol{\Gamma}$ corresponds to the first principal component (associated with the first variance), the second row to the second, etc. Only one of these solutions is correct. Which one is it? For each wrong solution give a reason for why it is wrong!

Variances	Principal component matrix $\boldsymbol{\Gamma}$	Answer
$\begin{pmatrix} 0.16 \\ 0.10 \\ 0.05 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.72 & 0.14 \\ 0.01 & 0.54 & 0.71 \\ -0.81 & -0.37 & 0.4 \\ -0.52 & 0.23 & -0.56 \end{pmatrix}$	
$\begin{pmatrix} 0.16 \\ 0.10 \\ 0.05 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.72 & 0.14 & -0.63 \\ 0.01 & 0.54 & 0.71 & -0.46 \\ -0.81 & -0.37 & 0.41 & 0.18 \end{pmatrix}$	
$\begin{pmatrix} 0.16 \\ -0.10 \\ 0.05 \\ 0.01 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.72 & 0.14 & -0.63 \\ 0.01 & 0.54 & 0.71 & -0.46 \\ -0.81 & -0.37 & 0.41 & 0.18 \\ -0.52 & 0.23 & -0.56 & -0.60 \end{pmatrix}$	
$\begin{pmatrix} 0.16 \\ 0.05 \\ 0.10 \\ 0.01 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.72 & 0.14 & -0.63 \\ -0.81 & -0.37 & 0.41 & 0.18 \\ 0.01 & 0.54 & 0.71 & -0.46 \\ -0.52 & 0.23 & -0.56 & -0.60 \end{pmatrix}$	
$\begin{pmatrix} 0.16 \\ 0.10 \\ 0.05 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.72 & 0.14 & -0.63 \\ 0.01 & 0.54 & 0.71 & -0.46 \\ 0.50 & -0.72 & 0.14 & -0.63 \end{pmatrix}$	

Problem 10 **Mixture Models (1 credit)**

Let $z \sim \text{Cat}(\boldsymbol{\pi})$ be a random variable with categorical distribution on $\{1, \dots, K\}$ with probabilities $\mathrm{p}(z = k) = \pi_k$ for $k \in \{1, \dots, K\}$. Furthermore, let x be a random variable dependent on z with an arbitrary likelihood, i.e. $\mathrm{p}(x \mid z)$ can be any probability distribution. Which of the following is the general form of $\mathrm{p}(z = k \mid x)$?

- ☐ $\pi_k \prod_{i=1}^{K} \log\left(\mathrm{p}(x \mid z = i)\, \pi_i\right)$
- ☐ $\mathrm{p}(x \mid z = k) \cdot \frac{\pi_k}{\sum_{i=1}^{K} \pi_i}$
- ☐ $\mathrm{p}(z = k) \cdot \mathbb{E}_z\left[\mathrm{p}\left(x \mid z\right)\right]$
- ☐ $\mathrm{p}(x \mid z = k)\, \pi_k \left(\sum_{i=1}^{K} \mathrm{p}(x \mid z = i)\, \pi_i\right)^{-1}$

Problem 11 **EM Algorithm (10 credits)**

Consider a one-dimensional mixture of exponential distributions with K components and a uniform prior over components, i.e.

$$\mathrm{p}(z_i = k) = \frac{1}{K} \qquad \mathrm{p}(x \mid \lambda_k, z_i = k) = \lambda_k \exp\left(-\lambda_k x\right) \qquad \text{where } \lambda_k > 0.$$

We have observed N values $x_i \in \mathbb{R}_{\geq 0}$ ($i = 1 \dots N$) and want to fit this mixture model with the EM algorithm.

a) Derive the M-step, i.e. the responsibilites respectively the posterior $\gamma(z_i = k) = \mathrm{p}(z_i = k \mid x_i)$.

☐ 0
☐ 1
☐ 2
☐ 3

0
1
2
3
4
5

b) Derive the E-step, i.e. find $\arg\max_\lambda \mathbb{E}_{\boldsymbol{Z}\sim\gamma}\left[\log \mathrm{p}\left(\boldsymbol{Z}, \boldsymbol{X} \mid \boldsymbol{\lambda}\right)\right]$. Here $\boldsymbol{Z}$ represents all z_i and $\boldsymbol{X}$ all x_i ($i = 1 \dots N$).

c) Is the EM algorithm guranteed to converge to a global optimum in general? If yes, justify why. If no, how to avoid getting stuck in local optima or saddle points?

0
1
2

Problem 12 Differential Privacy (2 credits)

0
1
2

Let $\mathcal{M}_f : \mathbb{R}^D \to \mathbb{R}^D$ be an $\epsilon -$ DP mechanism with a privacy parameter ϵ applied to the function $f : \mathbb{R}^D \to \mathbb{R}^D$. Similarly, let $\mathcal{N}_g : \mathbb{R}^D \to \mathbb{R}^D$ be a $\sigma -$ DP mechanism with a privacy parameter σ applied to the function g.

Let $h_1 : \mathbb{R}^D \to \mathbb{R}^D$ and $h_2 : \mathbb{R}^D \to \mathbb{R}^D$ be arbitrary functions and $\boldsymbol{X} \in \mathbb{R}^D$. Can we provide differential privacy guarantees for the following mappings? If yes, what is their respective privacy parameter? If no, why not?

a) $\boldsymbol{X} \mapsto (\mathcal{M}_f(\boldsymbol{X}), \mathcal{N}_g(\boldsymbol{X}))$

b) $\boldsymbol{X} \mapsto h_1(\mathcal{N}_g(h_2(\boldsymbol{X})))$

c) $\boldsymbol{X} \mapsto h_2(\mathcal{M}_f(\boldsymbol{X}))$

d) $\boldsymbol{X} \mapsto (\mathcal{M}_f(h_1(\boldsymbol{X})), \mathcal{N}_g(\boldsymbol{X}))$

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