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:**	Thursday 13th February, 2020
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	17:00 – 19:00

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

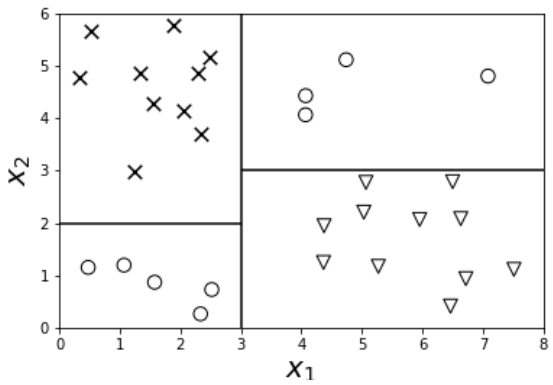
Working instructions

- This exam consists of **16 pages** with a total of **10 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 92 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - A4 sheet of handwritten notes (two sides)
  - **no other materials (e.g. books, cell phones, calculators) are allowed!**
- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Last two pages can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Only use a black or a blue pen (no pencils, red or green pens)!**
- 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 _____

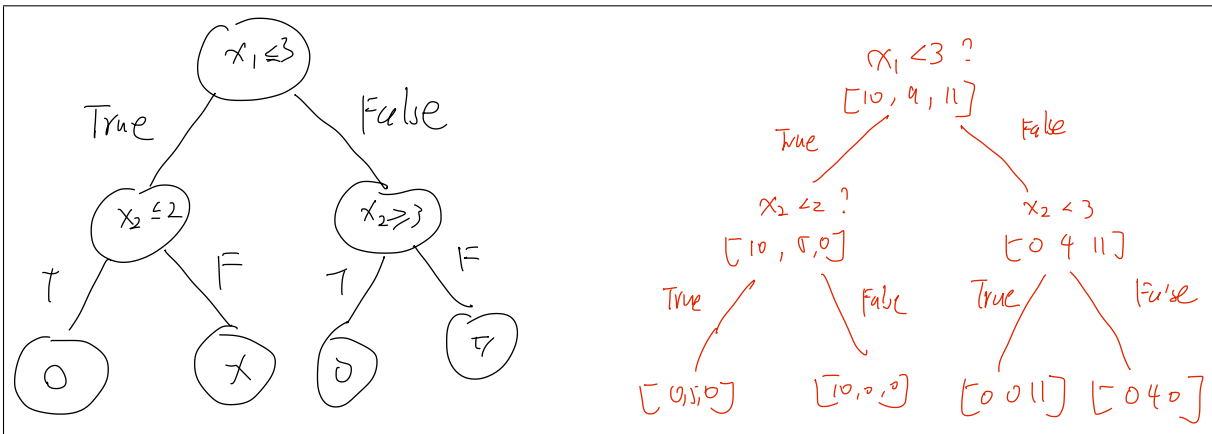
Problem 1 Decision Trees (12 credits)

You are given a dataset with points from three different classes and want to classify them based on 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.



0 1 2 3 4

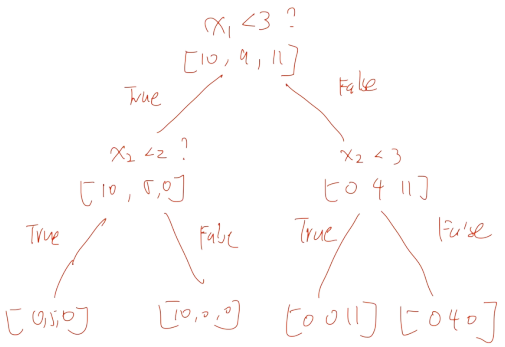
a) Draw the corresponding decision tree. Make sure that you include the feature (x_1 or x_2) and threshold of the split as well as the number of samples of each class that pass the corresponding inner node or leaf node.



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

Note: Your answer may contain improper fractions (e.g. $\frac{33}{117}$)

0 1 2 3 4



$$iG(1) = 1 - \left[\left(\frac{10}{30}\right)^2 + \left(\frac{9}{30}\right)^2 + \left(\frac{11}{30}\right)^2\right]$$

$$= 1 - \frac{100+81+121}{900}$$

$$= \frac{598}{900}$$

$$iG(2_L) = 1 - \left[\left(\frac{10}{15}\right)^2 + \left(\frac{5}{15}\right)^2\right]$$

$$= 1 - \frac{125}{225}$$

$$= \frac{100}{225}$$

$$iG(2_R) = 1 - \left[\left(\frac{4}{15}\right)^2 + \left(\frac{11}{15}\right)^2\right]$$

$$= 1 - \frac{137}{225}$$

$$= \frac{88}{225}$$

$$iG(3_{LL}) = 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$$

$$iG(3_{LR}) = 1 - \left(\frac{5}{5}\right)^2 = 0$$

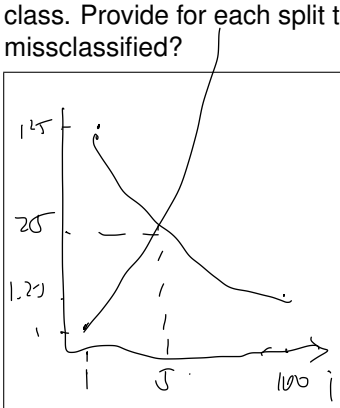
$$iG(3_{RL}) = 1 - \left(\frac{11}{11}\right)^2 = 0$$

$$iG(3_{RR}) = 1 - \left(\frac{4}{4}\right)^2 = 0$$

c) Assume you have a dataset with two-dimensional points from two different classes C_1 and C_2. The points from class C_1 are given by $A = \{(i, i^2) \mid i \in \{1...100\}\} \subseteq \mathbb{R}^2$, while the points from class C_2 are $B = \{(i, \frac{125}{i}) \mid i \in \{1...100\}\} \subseteq \mathbb{R}^2$.

0 1 2 3 4

Construct a decision tree of minimal depth that assigns as many data points as possible to the correct class. Provide for each split the feature and corresponding thresholds. How many and which datapoints are missclassified?



$$i^2 = \frac{125}{i}$$

$$i = 5$$

$\{5, 25\}$ point will be missclassified.

Problem 2 **Probabilistic inference (4 credits)**

0 1 2 3 4

We are interested in estimating a discrete parameter z that can take values in $\{1, 2, 3, 4\}$.

- We place a categorical prior on z, that is $p(z \mid \pi) = \text{Categorical}(z \mid \pi)$ with $\pi = (0.1, 0.05, 0.85, 0.0)$.
- We choose the following likelihood function: $p(x \mid z) = \text{Exponential}(x \mid 2^z) = 2^z \exp(-x2^z)$.
- We have observed one sample $x = 32$.

What is the posterior probability that z is equal to 4, i.e. what is $p(z = 4 \mid x, \pi)$? Justify your answer.

$$p(z=4 \mid x, \pi) = \frac{p(x \mid z=4, \pi)\, p(z=4 \mid \pi)}{p(x \mid z=1)\, p(z=1) + p(\cdots}$$

$$= \frac{2^4 \exp(-32 \cdot 2^4) \cdot 0)}{\cdot \quad -} \quad \approx 0$$

Problem 3 **Probabilistic inference (8 credits)**

0 1 2 3 4 5 6 7 8

We are interested in estimating the parameter $\theta \in \mathbb{R}$ of the following probabilistic model:

$$p(x \mid \theta) = \exp(\theta - x - \exp(\theta - x)).$$

We have observed a single sample $x \in \mathbb{R}$ drawn from the above model. Derive the maximum likelihood estimate (MLE) of the parameter θ. Justify your answer.

$$\theta_{MLE} = \arg\max_{\theta} f(\theta) = \arg\max_{\theta} p(x \mid \theta) = \arg\max_{\theta} \ln p(x \mid \theta)$$

$$\ln p(x \mid \theta) = \theta - x - \exp(\theta - x)$$

$$\frac{\partial \ln p(x \mid \theta)}{\partial \theta} = 1 - \exp^{(\theta - x)} \stackrel{!}{=} 0.$$

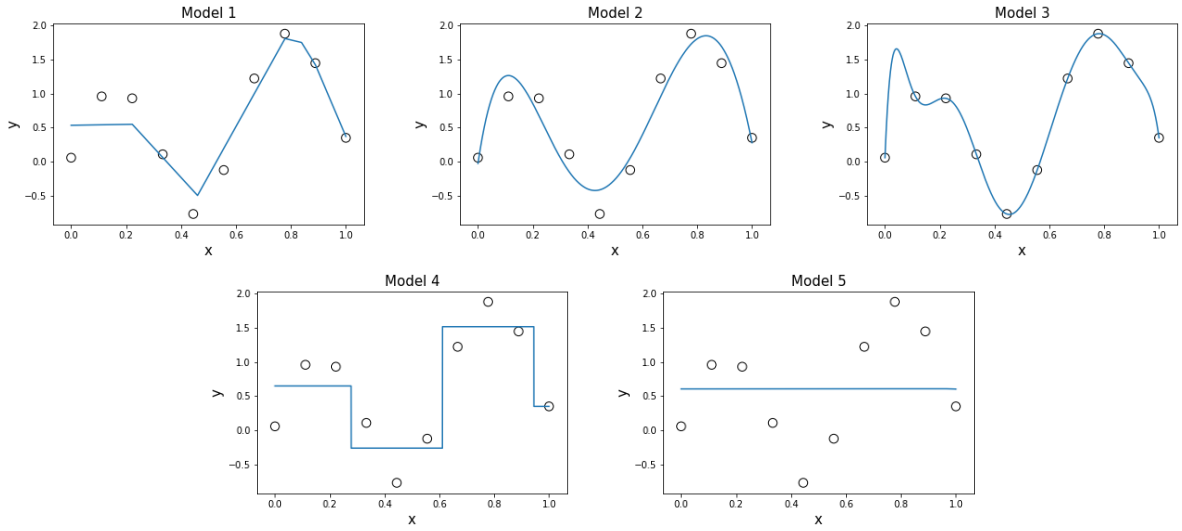
$$\exp^{(\theta - x)} = \exp^{(0)}$$

$$\theta^*_{MLE} = x.$$

Mark correct answers with a cross ☒
To undo a cross, completely fill out the answer option ■
To re-mark an option, use a human-readable marking ✕■

Problem 4 Regression (10 credits)

The following five plots show five different regression models fitted to the same dataset. Your task is to assign each of the plots to the corresponding model.



a) Polynomial regression (degree = 5), no regularization

☒ Model 1 ☒ Model 2 ☐ Model 3 ☐ Model 4 ☐ Model 5

b) Polynomial regression (degree = 10), no regularization

☐ Model 1 ☒ Model 2 ☒ Model 3 ☐ Model 4 ☐ Model 5

c) Polynomial regression (degree = 50), L_2 regularization with $\lambda = 10^3$

☐ Model 1 ☐ Model 2 ☒ Model 3 ☐ Model 4 ☒ Model 5

d) Feed-forward neural network with ReLU activation functions, no regularization

☒ Model 1 ☐ Model 2 ☐ Model 3 ☒ Model 4 ☐ Model 5

e) Decision tree of depth 2

☐ Model 1 ☐ Model 2 ☐ Model 3 ☒ Model 4 ☒ Model 5

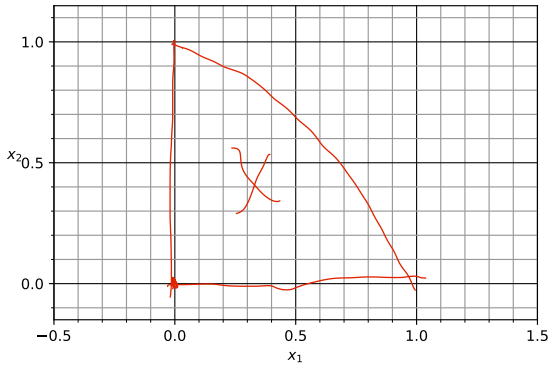
$\sqrt{x_1^2 + x_2^2}$

Problem 5 Convex optimization (18 credits)

Consider the set $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_2 \leq 1 \text{ and } x_i \geq 0 \text{ for all } i = 1, \ldots, D\}$ with $1 < D \in \mathbb{N}$.

0 1

a) Draw $\mathcal{X}$ on the provided axes below for $D = 2$.



0 1 2 3 4

b) Write down the function $\pi_{\mathcal{X}}$ projecting an arbitrary point $\boldsymbol{p} \in \mathbb{R}^2$ on $\mathcal{X}$ for the case $D = 2$.

Note: if you decide to split $\mathbb{R}^2$ into regions and consider them separately then you have to describe the regions analytically (just a reference to your plot from a) will not be sufficient).

let $R^2 = X \cup X_0 \cup X_1$

$X_0 = \{ x \in R^D : x_1 < 0 \text{ or } x_2 < 0 \text{ for } i = 1,2 \}$

$\pi_X(x) = \begin{pmatrix} \min(1, \max(0, x_1)) \\ \min(1, \max(0, x_2)) \end{pmatrix}$ for $x \in R^2$

$X_1 = \{ x \in R^D : \|x\| > 1, x_1 \geq 0 \text{ and } x_2 \geq 0 \}$

$\pi_X(x) = \frac{x}{\|x\|_2}$

$X = \{ x \in R^D : \|x\|_2 < 1 \text{ and } x_i \geq 0 \}$

$\pi_X(x) = x$

From now on we consider the setting with an arbitrary $1 < D \in \mathbb{N}$.

c) Prove that $\mathcal{X}$ is convex.

0 1 2 3

$X = \{x \in \mathbb{R}^D : \|x\|_2 \leq 1\} \cap \{x \in \mathbb{R}^D : 0 \leq x_i \leq 1\}$.

unit ball cube $[0,1]^D$

d) Fill in the space in the box below using mathematical notation with a description of the vertices of $\mathcal{X}$. Note that just writing down the definition of vert($\mathcal{X}$) will not be sufficient.

0 1 2

vert($\mathcal{X}$) =

$X = \{x \in \mathbb{R}^D : \|x\| = 1\} \cup \{0\}$

0 1 2 3 4 5 6 7 8

e) Find the maximum of the following constrained optimization problem. Justify your answer, all properties of the objective function and $\mathcal{X}$ that you use should be clearly stated and derived from the previous tasks or results considered in the course.

Hint: results from c) and d) might help you.
Hint: for arbitrary $\boldsymbol{c} \in \mathbb{R}^D$ the maximum of the constrained problem $\max_{\boldsymbol{x}} \boldsymbol{c}^T \boldsymbol{x}$ subject to $\|\boldsymbol{x}\|_2 = 1$ is $\|\boldsymbol{c}\|_2$.

$$\begin{aligned} \text{maximize}_{\boldsymbol{x}} \quad & \sum_{i=1}^{D} x_i + e^{\|\boldsymbol{x}\|_2^2} \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

解答

Maximum of convex function over convex domain are its vertices

Problem 6 **Kernels (10 credits)** PSD

Let $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ be a positive semi-definite matrix and consider for $p \in \mathbb{N}$ the following function

$$k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}, \quad k(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1^T \boldsymbol{A} \boldsymbol{x}_2 + 1)^p .$$

1×n n×1

a) Prove that k is a valid kernel using kernel preserving operations known from the course.

0 1 2 3 4

rule 5 $k(x_1, x_2) = x_1^T A x_2$ A is PSD.

rule 1 $k(x_1, x_2) = x_1^T A x_2 + 1$ is kernel.

rule . 3 $k(x_1, x_2) = k_1(x_1, x_2)\, k_2(x_1, x_2) = k(x_1, x_2)^p$. is kern

$x_1 = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \\ x_D \end{pmatrix}$ $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$(x_{11}, x_{12}, x_{13} \cdots x_{1n}) \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}$

$x_1^T x_1 + x_2^T x_2 + \cdots + 1$

$x_1^T x_1 + x_2^T x_2$

$(x_1^T x_1 + 1)(x_2^T x_2 + 1)$

b) For the special case of $p = 2$ and $\boldsymbol{A} = \boldsymbol{I}$ (identity matrix) write down the corresponding feature map $\phi : \mathbb{R}^D \to \mathbb{R}^M$ such that

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \phi(\boldsymbol{x}_1)^T \phi(\boldsymbol{x}_2) .$$

What is the dimension M of the feature space in this case?

0 1 2 3 4 5 6

$k(x_1, x_2) = (x_1^T x_2 + 1)^2$

$\phi(x_1)^T \phi(x_2) = (x_1^T x_2 + 1)^T (x_1^T x_2 + 1)$

Problem 7 **Deep learning (8 credits)**

The code snippet below shows an implementation of two functions $f : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$.

Given two input vectors $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^N$, we perform the following computations:

$$z = f(\boldsymbol{x}, \boldsymbol{y})$$
$$t = g(z)$$

The code below uses backpropagation to compute $\frac{\partial t}{\partial \boldsymbol{x}}$ and $\frac{\partial t}{\partial \boldsymbol{y}}$ (similarly to how we did it in Tutorial 9: Deep Learning I). However, some code fragments are missing. Your task is to complete the missing code fragments.

Note: It's also fine to write your answer using pseudocode (we won't deduct points for small Python syntax errors, etc.).

```
class F:
    def forward(self, x, y):
        self.cache = (x, y)
        ###################################################
        # MISSING CODE FRAGMENT #1
        ###################################################
        return out

    def backward(self, d_out):
        # x, y are np.arrays of shape [N]
        x, y = self.cache
        N = len(x)
        # np.ones(N) returns a vector of ones of shape [N]
        d_y = (x + np.ones(N)) * d_out
        d_x = (y - np.ones(N)) * d_out
        return d_x, d_y

def sigmoid(a):
    return 1 / (1 + np.exp(-a))

class G:
    def forward(self, z):
        self.cache = z
        return sigmoid(z)

    def backward(self, d_out):
        z = self.cache
        ###################################################
        # MISSING CODE FRAGMENT #2
        ###################################################
        return d_z

# Example usage
f = F()
g = G()
x = np.array([1., 2., 3])
y = np.array([-2., 3., -1.])

z = f.forward(x, y)
t = g.forward(z)
d_z = g.backward(1.0)
d_x, d_y = f.backward(d_z)
```

a) Complete the MISSING CODE FRAGMENT #1

```
out = np.dot(x,y) + np.sum(y) - np.sum(x)
```

0 1 2 3 4

b) Complete the MISSING CODE FRAGMENT #2

```
dz = sigmoid(z) * (1 - sigmoid(z)) * d_out
```

0 1 2 3 4

Problem 8 SVD and linear regression (8 credits)

0 1 2 3 4 5 6 7 8

You want to perform linear regression on a data set with features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and targets $\boldsymbol{y} \in \mathbb{R}^{N}$. Assume that you have already computed the SVD of the feature matrix $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{T}$. Additionally, assume that $\boldsymbol{X}$ has full rank.

Show how we can compute the optimal linear regression weights $\boldsymbol{w}^{\star}$ in $\mathcal{O}(ND^2)$ operations by using the result of the SVD.

Hint: Matrix operations have the following asymptotic complexity

- *Matrix multiplication* $\boldsymbol{AB}$ *for arbitrary* $\boldsymbol{A} \in \mathbb{R}^{P \times Q}$ *and* $\boldsymbol{B} \in \mathbb{R}^{Q \times R}$ *takes* $\mathcal{O}(PQR)$
- *Matrix multiplication* $\boldsymbol{AD}$ *for an arbitrary* $\boldsymbol{A} \in \mathbb{R}^{P \times Q}$ *and a diagonal* $\boldsymbol{D} \in \mathbb{R}^{Q \times Q}$ *takes* $\mathcal{O}(PQ)$
- *Matrix inversion* $\boldsymbol{C}^{-1}$ *for an arbitrary matrix* $\boldsymbol{C} \in \mathbb{R}^{M \times M}$ *takes* $\mathcal{O}(M^3)$
- *Matrix inversion* $\boldsymbol{D}^{-1}$ *for a diagonal matrix* $\boldsymbol{D} \in \mathbb{R}^{M \times M}$ *takes* $\mathcal{O}(M)$

$$
\begin{aligned}
w^{*} &= (X^T X)^{-1} X^T y \\
&= \left((U\Sigma V^T)^T (U\Sigma V^T) \right)^{-1} (U\Sigma V^T)^T y \\
&= (V\Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma U^T y \\
&= (V \Sigma^2 V^T)^{-1} V \Sigma U^T y \\
&= V \Sigma^{-2} V^{-1} V \Sigma U^T y \qquad (V^T = V^{-1},\ V^T V = I) \\
&= V \Sigma^{-1} U^T y
\end{aligned}
$$

$V: D \times D$, $\Sigma^{-1}: D \times N$, $U^T y: N \times 1$, result: $D \times 1$

$O(D^2) \cdot O(DN)$ $\quad O(DN)$ $\quad (N > 1)$

Problem 9 K-Means (10 credits)

Let $\gamma_i \in \mathbb{R}^D$ for $i = 1, \dots, K$ be a set of K points more than 4 apart, i.e. $\|\gamma_i - \gamma_j\|_2 > 4$ for all $i \neq j$. Consider K non-empty datasets $\mathcal{X}_i$ each contained within a unit ball around γ_i, i.e. $\|\boldsymbol{x} - \gamma_i\|_2 \leq 1$ for all $\boldsymbol{x} \in \mathcal{X}_i$. Let $\mathcal{X} = \bigcup_{i=1}^{K} \mathcal{X}_i$ be the combined dataset.

Now consider a centroid initialization procedure similar to k-means++, though it deterministically chooses the data point farthest away from all previous centroids. That means it initializes the cluster centers $\boldsymbol{\mu}_i$ as

$$\boldsymbol{\mu}_i = \begin{cases} \text{random sample from } \mathcal{X} & \text{if } i = 1 \\ \arg\max_{\boldsymbol{x} \in \mathcal{X}} \min_{j \in \{1,\dots,i-1\}} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|_2 & \text{if } i \in \{2, \dots, K\} \end{cases}$$

a) Explain why this deterministic k-means++ initialization of K clusters assigns each $\boldsymbol{\mu}_i$ to a different ball, i.e. for $i \neq j$ such that $\boldsymbol{\mu}_i \in \mathcal{X}_{i'}$ and $\boldsymbol{\mu}_j \in \mathcal{X}_{j'}$ it holds that $i' \neq j'$.

0 1 2 3 4 5

b) Assuming a), explain why k-means clustering of $\mathcal{X}$ with K clusters and our deterministic k-means++ initialization recovers the underlying structure of the data, i.e. all data points $\boldsymbol{x} \in \mathcal{X}_i$ will be assigned to the same centroid for all i.

0 1 2 3 4 5

Problem 10 **Differential Privacy & Fairness (4 credits)**

0 1 2

a) What is the *robustness to post-processing* property of Differential Privacy?

0 1 2

b) Suppose that we require that the same *percentage* of applicants from two different groups must receive a loan. What fairness criterion are we implementing? What is the biggest downside/con of this criterion?

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