

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

Graded Exercise: IN2064 / Retake

Date: Thursday 1st April, 2021

Examiner: Prof. Dr. Stephan Günnemann

Time: 16:30 – 18:30

Working instructions

- This graded exercise consists of **34 pages** with a total of **11 problems**.
Please make sure now that you received a complete copy of the graded exercise.
- The total amount of achievable credits in this graded exercise is 40.
- 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

Left room from _____ to _____ / Early submission at _____

Problem 1: Probabilistic Inference (Version A) (4 credits)

We have observed N coin flips of which T landed tails and H landed heads ($T + H = N$). We model each coin flip with a $\text{Bernoulli}(\theta)$ distribution with a shared unknown probability θ of coming up heads, i.e. $p(\text{heads} \mid \theta) = \theta$.

Furthermore, we assume that θ follows a $\text{Beta}(a, b)$ distribution with parameters $a > 0$ and $b > 0$. In this problem we are not interested in estimating θ , but rather estimating the parameters a and b .

- 0 ☐ a) Is it possible to compute a maximum likelihood estimate (MLE) $\arg \max_{a,b} p(\mathcal{D} \mid a, b)$ of a and b in this model?
 1 ☐ If yes, briefly describe a way to do it and how the variable θ is handled.
 2 ☐ If no, explain why it is not possible and the role of the variable θ .

- 0 ☐ b) Is it possible to compute a maximum a posteriori (MAP) estimate $\arg \max_{a,b} p(a, b \mid \mathcal{D})$ of a and b in this model?
 1 ☐ If yes, briefly describe how to compute the MAP estimate of a and b .
 2 ☐ If no, describe how and why the model would need to be changed/extended to allow MAP estimation of a and b .

$$p(x \mid \theta) = \text{Ber}(\theta) = \theta^H (1-\theta)^{N-H}$$

$$p(\theta \mid a, b) = \text{Beta}(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \theta^{a-1} (1-\theta)^{b-1}$$

$$p(\mathcal{D} \mid a, b) = \prod p(x_i \mid \theta) = \prod \int p(x_i \mid \theta) p(\theta \mid a, b) d\theta$$

$$= \int \prod p(x_i \mid \theta) p(\theta \mid a, b) d\theta$$

$$p(a, b \mid \mathcal{D}) = p(\mathcal{D} \mid a, b) \underbrace{p(a, b)}_{\text{prior}} \quad \text{no}$$

Problem 1: Probabilistic Inference (Version B) (4 credits)

We have observed N coin flips of which T landed tails and H landed heads ($T + H = N$). We model each coin flip with a Bernoulli(θ) distribution with a shared unknown probability θ of coming up heads, i.e. $p(\text{heads} \mid \theta) = \theta$.

Furthermore, we assume that θ follows a Beta(a, b) distribution with parameters $a > 0$ and $b > 0$. In this problem we are not interested in estimating θ , but rather estimating the parameters a and b .

a) Is it possible to compute a maximum likelihood estimate (MLE) $\arg \max_{a,b} p(\mathcal{D} \mid a, b)$ of a and b in this model?

If yes, briefly describe a way to do it and how the variable θ is handled.

If no, explain why it is not possible and the role of the variable θ .

<input type="checkbox"/>	0
<input type="checkbox"/>	
<input type="checkbox"/>	1
<input type="checkbox"/>	
<input type="checkbox"/>	2

b) Is it possible to compute a maximum a posteriori (MAP) estimate $\arg \max_{a,b} p(a, b \mid \mathcal{D})$ of a and b in this model?

If yes, briefly describe how to compute the MAP estimate of a and b .

If no, describe how and why the model would need to be changed/extended to allow MAP estimation of a and b .

<input type="checkbox"/>	0
<input type="checkbox"/>	
<input type="checkbox"/>	1
<input type="checkbox"/>	
<input type="checkbox"/>	2

Problem 2: Decision Trees (Version A) (5 credits)

0 ☐ a) Suppose we randomly sample subsets of features to learn separate trees which are then combined. What is this technique called?

1 ☐ ~~embedded~~ *bagging*
Ensemble

0 ☐ b) We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N instances with $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits.

1 ☐ Specify a small dataset \mathcal{D} so that the learned decision tree has no splits – the root node is a leaf. Write down all (\mathbf{x}_i, y_i) tuples in your \mathcal{D} and make sure it contains at least one instance from each class. Justify your answer.

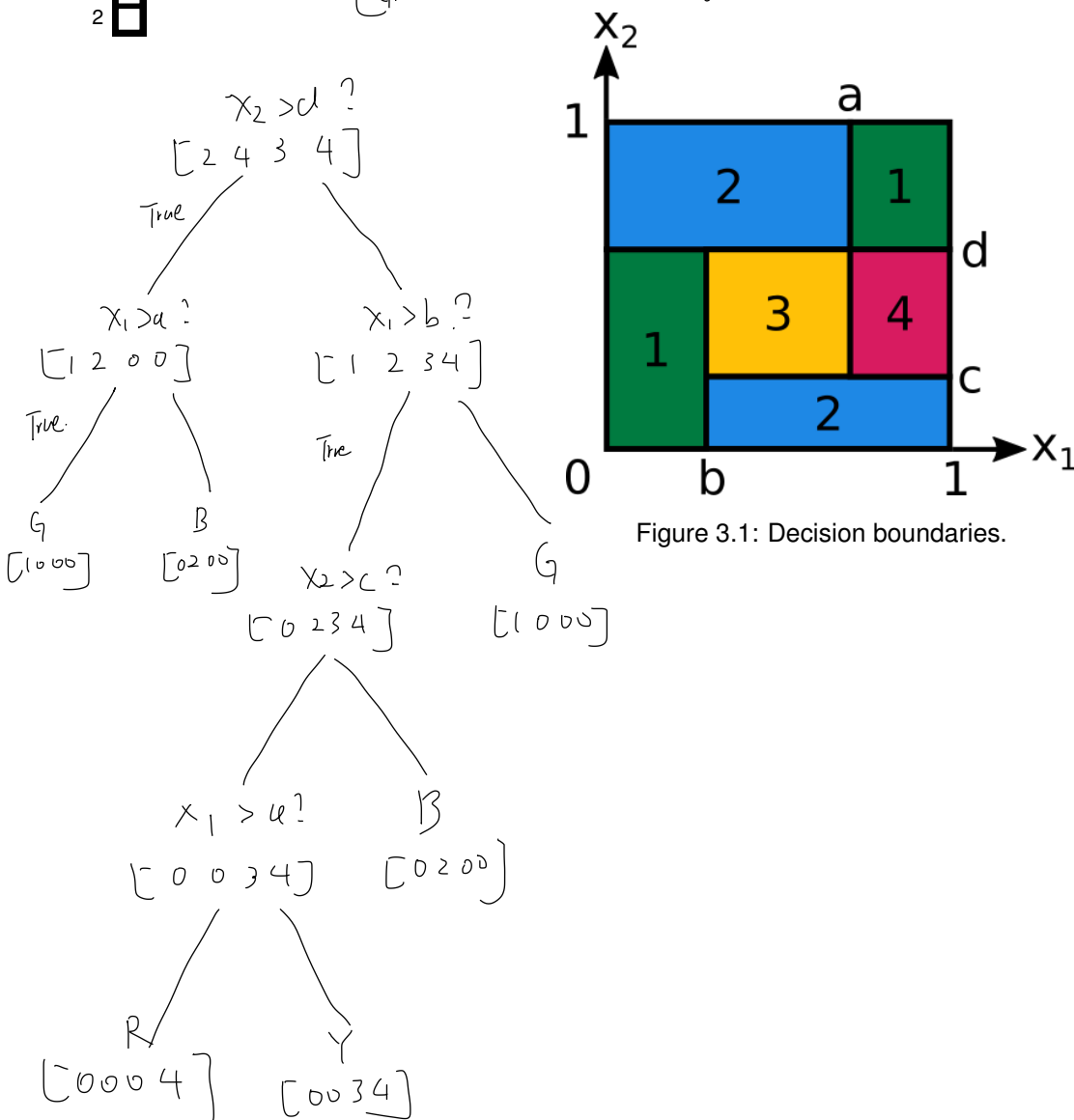
2 ☐ *Hint: you do not need more than a few instances.*

$$y_1 = y_2 \text{ or } y_1 \neq y_2 \text{ when } x_1 = x_1, x_2 = x_2$$

0 ☐ c) Draw the decision tree corresponding to the decision boundaries shown on Figure 3.1 where $0 \leq a, b, c, d \leq 1$ are some arbitrary constants and there are four classes marked with four numbers and different colors. Assume $x_1, x_2 \in [0, 1]$. *[Green Blue Yellow Red]*

1 ☐ *[Green Blue Yellow Red]*

2 ☐



Problem 2: Decision Trees (Version B) (5 credits)

a) Suppose we randomly sample subsets of features to learn separate trees which are then combined. What is this technique called?

☐ 0
☐ 1

b) We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N instances with $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits.

☐ 0
☐ 1
☐ 2

Specify a small dataset \mathcal{D} so that the learned decision tree has no splits – the root node is a leaf. Write down all (\mathbf{x}_i, y_i) tuples in your \mathcal{D} and make sure it contains at least one instance from each class. Justify your answer.

Hint: you do not need more than a few instances.

c) Draw the decision tree corresponding to the decision boundaries shown on Figure 4.1 where $0 \leq a, b, c, d \leq 1$ are some arbitrary constants and there are four classes marked with four numbers and different colors. Assume $x_1, x_2 \in [0, 1]$.

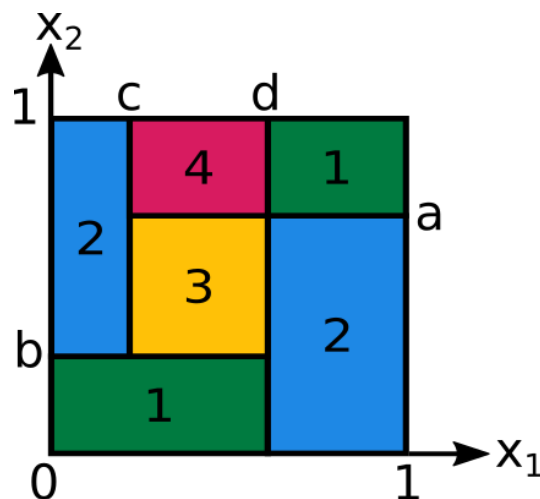

☐ 0
☐ 1
☐ 2


Figure 4.1: Decision boundaries.

Problem 2: Decision Trees (Version C) (5 credits)

0  a) Suppose we randomly sample subsets of features to learn separate trees which are then combined. What is this technique called?

0 ☐ b) We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N instances with $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits.

1 ☐

2 **E** Specify a small dataset \mathcal{D} so that the learned decision tree has no splits – the root node is a leaf. Write down all (\mathbf{x}_i, y_i) tuples in your \mathcal{D} and make sure it contains at least one instance from each class. Justify your answer.

Hint: you do not need more than a few instances.

0 ☐ c) Draw the decision tree corresponding to the decision boundaries shown on Figure 5.1 where $0 \leq a, b, c, d \leq 1$
 1 ☐ are some arbitrary constants and there are four classes marked with four numbers and different colors. Assume $x_1, x_2 \in [0, 1]$.

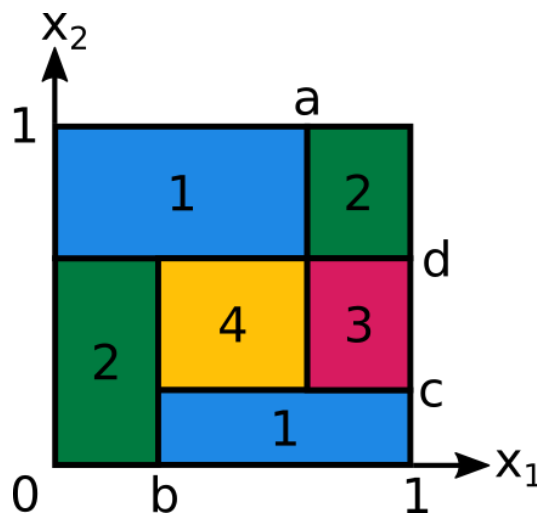


Figure 5.1: Decision boundaries.

Problem 2: Decision Trees (Version D) (5 credits)

a) Suppose we randomly sample subsets of features to learn separate trees which are then combined. What is this technique called?

☐ 0
☐ 1

b) We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N instances with $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits.

☐ 0
☐ 1
☐ 2

Specify a small dataset \mathcal{D} so that the learned decision tree has no splits – the root node is a leaf. Write down all (\mathbf{x}_i, y_i) tuples in your \mathcal{D} and make sure it contains at least one instance from each class. Justify your answer.

Hint: you do not need more than a few instances.

c) Draw the decision tree corresponding to the decision boundaries shown on Figure 6.1 where $0 \leq a, b, c, d \leq 1$ are some arbitrary constants and there are four classes marked with four numbers and different colors. Assume $x_1, x_2 \in [0, 1]$.

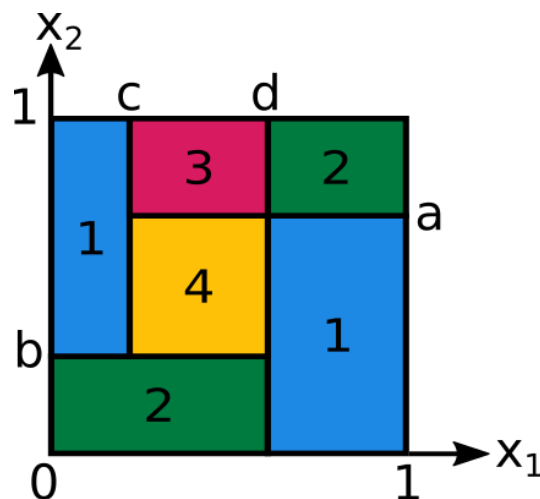
☐ 0
☐ 1
☐ 2


Figure 6.1: Decision boundaries.

Problem 3: Linear Regression (Version A) (2 credits)

0 ☐
1 ☐
2 ☐

We have a dataset $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^2, y_i \in \mathbb{R}\}_{i=1}^N$ and want to fit the following model with three parameters $\mathbf{w} = (a \ b \ c)^T \in \mathbb{R}^3$ to it.

$$f(\mathbf{x}, \mathbf{w}) = a \sin(x_2) + \frac{1}{2} b \|\mathbf{x}\|_1 - \mathbf{x}_1^2 \mathbf{x}_2 c$$

Give a closed form expression for the optimal \mathbf{w} minimizing the squared error between the predictions and targets

$$\sum_{i=1}^N (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2.$$

Justify your answer.

Note: You can use results from the lecture without deriving them again.

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \begin{pmatrix} \sin(x_2) \\ \frac{1}{2} \|\mathbf{x}\|_1 \\ -\mathbf{x}_1^2 \mathbf{x}_2 \end{pmatrix} = \mathbf{w}^T \phi(\mathbf{x})$$

$$\begin{aligned} \mathbf{w}^* &= \arg\min_{\mathbf{w}} \left[\mathbf{w}^T \Phi - \mathbf{y} \right]^2 \\ &= (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \end{aligned}$$

Problem 3: Linear Regression (Version B) (2 credits)

We have a dataset $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^2, y_i \in \mathbb{R}\}_{i=1}^N$ and want to fit the following model with three parameters $\mathbf{w} = (a \ b \ c)^\top \in \mathbb{R}^3$ to it.

$$f(\mathbf{x}, \mathbf{w}) = a\|\mathbf{x}\|_2 - \frac{1}{2}\mathbf{x}_1^2\mathbf{x}_2b + c\cos(\mathbf{x}_1)$$

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2

Give a closed form expression for the optimal \mathbf{w} minimizing the squared error between the predictions and targets

$$\sum_{i=1}^N (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2.$$

Justify your answer.

Note: You can use results from the lecture without deriving them again.

Problem 3: Linear Regression (Version C) (2 credits)

0 ☐
1 ☐
2 ☐

We have a dataset $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^2, y_i \in \mathbb{R}\}_{i=1}^N$ and want to fit the following model with three parameters $\mathbf{w} = (a \ b \ c)^\top \in \mathbb{R}^3$ to it.

$$f(\mathbf{x}, \mathbf{w}) = -\mathbf{x}_1 \mathbf{x}_2^2 a + b \tan(\mathbf{x}_2) + \frac{1}{2} c \|\mathbf{x}\|_\infty$$

Give a closed form expression for the optimal \mathbf{w} minimizing the squared error between the predictions and targets

$$\sum_{i=1}^N (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2.$$

Justify your answer.

Note: You can use results from the lecture without deriving them again.

Problem 4: Logistic Regression (Version A) (6 credits)

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}\}$ be a classification dataset that is not linearly separable. Furthermore, let $\mathcal{D}_\alpha = \{(\alpha \mathbf{x}_i, y_i) \mid (\mathbf{x}_i, y_i) \in \mathcal{D}\}$ be a scaled copy of \mathcal{D} with $\alpha > 1$. $f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 1]$ is a logistic regression model on \mathbb{R}^m for some m with parameters \mathbf{w} . $f(\mathbf{x}, \mathbf{w})$ outputs the predicted probabilities for class 1.

You train two logistic regression models on \mathcal{D} and \mathcal{D}_α *without regularization* and obtain the optimal parameters \mathbf{w}^* and $\mathbf{w}^{*,\alpha}$, respectively. Consider a test point $\mathbf{x}_{\text{test}} \in \mathbb{R}^d$ and the predicted probabilities by the two models, $s = f(\mathbf{x}_{\text{test}}, \mathbf{w}^*)$ and $t = f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,\alpha})$.

$$\alpha \mathbf{w}^{*,\alpha} = \mathbf{w}^*$$

a) Is $s > t$ possible? Is $s = t$ possible? Is $s < t$ possible? Justify your answers.

$$f(\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}_i) = y_i$$

σ monotonically increasing

$$\sigma(\mathbf{w}^\top \alpha \mathbf{x}_i) = y_i$$

$$\sigma(\mathbf{w}^\top \mathbf{x}) < \sigma(\mathbf{w}^\top \alpha \mathbf{x}) \text{ for } \alpha > 0$$

$$s = \sigma(\mathbf{w}^{*,\alpha} \cdot \mathbf{x}_{\text{test}})$$

$$t = \sigma\left(\frac{\mathbf{w}^*}{\alpha} \cdot \mathbf{x}_{\text{test}}\right)$$

for $\alpha > 0$
if $\mathbf{w}^{*,\alpha} \cdot \mathbf{x}_{\text{test}} > 0$

$s > t$ if $\mathbf{w}^{*,\alpha} \cdot \mathbf{x}_{\text{test}} < 0$

$s < t$

Now consider another dataset \mathcal{D}_3 and a logistic regression model $f(\mathbf{x}, \mathbf{w})$ on \mathcal{D}_3 .

$$\mathcal{D}_3 = \{(\mathbf{x}_i, 1) \mid \mathbf{x}_i \in \mathbb{R}^2, \mathbf{x}_{i,1} > 0, \mathbf{x}_{i,2} > 0\} \cup \{(\mathbf{x}_i, 0) \mid \mathbf{x}_i \in \mathbb{R}^2, \mathbf{x}_{i,1} < 0, \mathbf{x}_{i,2} < 0\}$$

In the following, treat ∞ as an actual value, i.e. $\infty \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ would be a vector in the same direction as $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ but with infinite length/norm.

b) Briefly explain why the maximum likelihood estimate of \mathbf{w} on \mathcal{D}_3 obtained by training *without regularization* is not unique in this setting.

$$E(\mathbf{w}) = - \sum y_i \ln \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i) \ln (1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$$

$$\min E(\mathbf{w}) \quad \mathbf{w} \uparrow \rightarrow \infty$$

c) Give two possible maximum likelihood estimates (without regularization) of the parameters \mathbf{w} on \mathcal{D}_3 , $\mathbf{w}^{*,a}$ and $\mathbf{w}^{*,b}$, and a test point \mathbf{x}_{test} such that either $f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,a}) < \frac{1}{2} < f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,b})$ or $f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,a}) > \frac{1}{2} > f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,b})$ holds. Justify your answer.

$$\uparrow$$

$$\mathbf{w}^{*,a} < 0 \quad \mathbf{w}^{*,b} > 0$$

$$\text{or } \mathbf{w}^{*,a} > 0, \mathbf{w}^{*,b} < 0$$

Problem 4: Logistic Regression (Version B) (6 credits)

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}\}$ be a classification dataset that is *not* linearly separable. Furthermore, let $\mathcal{D}_\alpha = \{(\alpha \mathbf{x}_i, y_i) \mid (\mathbf{x}_i, y_i) \in \mathcal{D}\}$ be a scaled copy of \mathcal{D} with $\alpha > 1$. $f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 1]$ is a logistic regression model on \mathbb{R}^m for some m with parameters \mathbf{w} . $f(\mathbf{x}, \mathbf{w})$ outputs the predicted probabilities for class 1.

You train two logistic regression models on \mathcal{D} and \mathcal{D}_α *without regularization* and obtain the optimal parameters \mathbf{w}^* and $\mathbf{w}^{*,\alpha}$, respectively. Consider a test point $\mathbf{x}_{\text{test}} \in \mathbb{R}^d$ and the predicted probabilities by the two models, $s = f(\mathbf{x}_{\text{test}}, \mathbf{w}^*)$ and $t = f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,\alpha})$.

0 ☐ a) Is $s > t$ possible? Is $s = t$ possible? Is $s < t$ possible? Justify your answers.

1 ☐

2 ☐

Now consider another dataset \mathcal{D}_3 and a logistic regression model $f(\mathbf{x}, \mathbf{w})$ on \mathcal{D}_3 .

$$\mathcal{D}_3 = \{(\mathbf{x}_i, 1) \mid \mathbf{x}_i \in \mathbb{R}^2, \mathbf{x}_{i,1} > 0, \mathbf{x}_{i,2} > 0\} \cup \{(\mathbf{x}_i, 0) \mid \mathbf{x}_i \in \mathbb{R}^2, \mathbf{x}_{i,1} < 0, \mathbf{x}_{i,2} < 0\}$$

In the following, treat ∞ as an actual value, i.e. $\infty \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ would be a vector in the same direction as $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ but with infinite length/norm.

0 ☐ b) Briefly explain why the maximum likelihood estimate of \mathbf{w} on \mathcal{D}_3 obtained by training *without* regularization is not unique in this setting.

1 ☐

2 ☐

0 ☐ c) Give two possible maximum likelihood estimates (without regularization) of the parameters \mathbf{w} on \mathcal{D}_3 , $\mathbf{w}^{*,a}$ and $\mathbf{w}^{*,b}$, and a test point \mathbf{x}_{test} such that either $f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,a}) < \frac{1}{2} < f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,b})$ or $f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,a}) > \frac{1}{2} > f(\mathbf{x}_{\text{test}}, \mathbf{w}^{*,b})$ holds. Justify your answer.

1 ☐

2 ☐

Problem 5: Optimization (Version A) (3 credits)

Suppose we're minimizing some differentiable convex function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ using one of the following variants of gradient descent. Let θ_t be the value of the parameter at iteration t , and θ_{t+1} be the value of the parameter at the iteration $t+1$ for some integer t .

a) Suppose we are using gradient descent with line search.

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

0
1

Yes $t^* = \arg\min_t f(\theta + t\Delta\theta)$

guarante to find a t^* make $f(\theta_{t+1}) \leq f(\theta_t)$

b) Suppose we are using gradient descent with fixed step size.

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

0
1

No $\theta_{t+1} \leftarrow \theta_t - \tau \nabla_{\theta} L(\theta)$

If τ too big, and θ_t is close to θ^* , may overshoot

c) Suppose we are using gradient descent with adaptive learning rate (Adam - Adaptive moment estimation).

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

0
1

No because in time t , these optimization doesn't know the next value $f(\theta_{t+1}) \leq f(\theta_t)$ by spatial τ

so it can still overshoot.

Problem 5: Optimization (Version B) (3 credits)

Suppose we're minimizing some differentiable convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ using one of the following variants of gradient descent. Let θ_t be the value of the parameter at iteration t , and θ_{t+1} be the value of the parameter at the iteration $t + 1$ for some integer t .

- 0 ☐ a) Suppose we are using gradient descent with fixed step size.
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

- 0 ☐ b) Suppose we are using gradient descent with adaptive learning rate (Adam - Adaptive moment estimation).
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

- 0 ☐ c) Suppose we are using gradient descent with line search.
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

Problem 5: Optimization (Version C) (3 credits)

Suppose we're minimizing some differentiable convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ using one of the following variants of gradient descent. Let θ_t be the value of the parameter at iteration t , and θ_{t+1} be the value of the parameter at the iteration $t + 1$ for some integer t .

a) Suppose we are using gradient descent with line search.

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

<input type="checkbox"/>	0
<input type="checkbox"/>	1

b) Suppose we are using gradient descent with adaptive learning rate (Adam - Adaptive moment estimation).

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

<input type="checkbox"/>	0
<input type="checkbox"/>	1

c) Suppose we are using gradient descent with fixed step size.

Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

<input type="checkbox"/>	0
<input type="checkbox"/>	1

Problem 5: Optimization (Version D) (3 credits)

Suppose we're minimizing some differentiable convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ using one of the following variants of gradient descent. Let θ_t be the value of the parameter at iteration t , and θ_{t+1} be the value of the parameter at the iteration $t + 1$ for some integer t .

- 0 ☐ a) Suppose we are using gradient descent with adaptive learning rate (Adam - Adaptive moment estimation).
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

- 0 ☐ b) Suppose we are using gradient descent with fixed step size.
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

- 0 ☐ c) Suppose we are using gradient descent with line search.
1 ☐ Is the inequality $f(\theta_{t+1}) \leq f(\theta_t)$ guaranteed to always hold? Justify your answer.

Problem 6: Deep Learning (Version A) (3 credits)

Suppose $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ are two vectors. We define the function $f: \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}, \mathbf{y}) = \log(1 + \exp(\mathbf{x}^T \mathbf{y})).$$

The code below implements the computation of $f(\mathbf{x}, \mathbf{y})$ as well as its gradients w.r.t. \mathbf{x} and \mathbf{y} using backpropagation (similarly to how we did in Exercise sheet 7). However, some code fragments are missing. Your task is to complete the missing code fragments.

$$\frac{1}{1 + \exp(\mathbf{x}^T \mathbf{y})} \cdot \exp(\mathbf{x}^T \mathbf{y}) \cdot \mathbf{y}$$

```
import numpy as np
```

```
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
```

```
        #####
```

```
        # MISSING CODE FRAGMENT #2
```

```
        #####
```

```
        return d_x, d_y
```

$$out = np.log(1 + np.exp(np.dot(x.T, y)))$$

$$dx = (1 / (1 + np.exp(np.dot(x.T, y)))) \times np.exp(x.T @ y) \times y$$

```
# Example usage
```

```
f = F()
```

```
x = np.array([1., 2., 3])
```

```
y = np.array([-2., 3., -1.])
```

```
z = f.forward(x, y)
```

```
d_z = 1.0
```

```
d_x, d_y = f.backward(d_z)
```

a) Complete the MISSING CODE FRAGMENT #1

<input type="checkbox"/>	0
<input type="checkbox"/>	1

b) Complete the MISSING CODE FRAGMENT #2

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2

Problem 6: Deep Learning (Version B) (3 credits)

Suppose $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ are two vectors. We define the function $f: \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}, \mathbf{y}) = \log(\exp(\mathbf{x}^T \mathbf{y}) - 1).$$

The code below implements the computation of $f(\mathbf{x}, \mathbf{y})$ as well as its gradients w.r.t. \mathbf{x} and \mathbf{y} using backpropagation (similarly to how we did in Exercise sheet 7). However, some code fragments are missing. Your task is to complete the missing code fragments.

```
import numpy as np

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
        #####
        # MISSING CODE FRAGMENT #2
        #####
        return d_x, d_y

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

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

0 ☐ a) Complete the MISSING CODE FRAGMENT #1

1 ☐

0 ☐ b) Complete the MISSING CODE FRAGMENT #2

1 ☐

2 ☐

Problem 6: Deep Learning (Version C) (3 credits)

Suppose $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ are two vectors. We define the function $f: \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}, \mathbf{y}) = \log(1 + \exp(\mathbf{x}^T \mathbf{y})).$$

The code below implements the computation of $f(\mathbf{x}, \mathbf{y})$ as well as its gradients w.r.t. \mathbf{x} and \mathbf{y} using backpropagation (similarly to how we did in Exercise sheet 7). However, some code fragments are missing. Your task is to complete the missing code fragments.

```
import numpy as np

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
        #####
        # MISSING CODE FRAGMENT #2
        #####
        return d_x, d_y

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

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

a) Complete the MISSING CODE FRAGMENT #1

<input type="checkbox"/>	0
<input type="checkbox"/>	1

b) Complete the MISSING CODE FRAGMENT #2

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2

Problem 6: Deep Learning (Version D) (3 credits)

Suppose $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ are two vectors. We define the function $f: \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}, \mathbf{y}) = \log(\exp(\mathbf{x}^T \mathbf{y}) - 1).$$

The code below implements the computation of $f(\mathbf{x}, \mathbf{y})$ as well as its gradients w.r.t. \mathbf{x} and \mathbf{y} using backpropagation (similarly to how we did in Exercise sheet 7). However, some code fragments are missing. Your task is to complete the missing code fragments.

```
import numpy as np

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
        #####
        # MISSING CODE FRAGMENT #2
        #####
        return d_x, d_y

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

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

0 ☐ a) Complete the MISSING CODE FRAGMENT #1

1 ☐

0 ☐ b) Complete the MISSING CODE FRAGMENT #2

1 ☐

2 ☐

Problem 7: Kernels (Version A) (3 credits)

Let $\Sigma \in \mathbb{R}^{D \times D}$ be a given invertible, positive semi-definite matrix and $c \in \mathbb{R}$ be a given constant. Consider the kernel:

$$k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}, \quad k(\mathbf{x}_1, \mathbf{x}_2) = c^2 \exp \left(-\frac{1}{2} (\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2) \right).$$

Prove or disprove that k is a valid kernel.

Hint: If $k_1(\mathbf{x}_1, \mathbf{x}_2)$ then $\exp(k_1(\mathbf{x}_1, \mathbf{x}_2))$ is also a kernel.

$$= c^2 \exp \left(-\frac{1}{2} (\mathbf{x}_1^T \Sigma^{-1} - \mathbf{x}_2^T \Sigma^{-1}) (\mathbf{x}_1 - \mathbf{x}_2) \right)$$

$$= c^2 \exp \left(-\frac{1}{2} (\mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_1 - 2 \mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_2 + \mathbf{x}_2^T \Sigma^{-1} \mathbf{x}_2) \right)$$

$$= c^2 \exp \left(-\frac{1}{2} \mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_1 + \mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_2 - \frac{1}{2} \mathbf{x}_2^T \Sigma^{-1} \mathbf{x}_2 \right)$$

$$= c^2 \exp \left(-\frac{1}{2} \mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_1 \right) \exp \left(-\frac{1}{2} \mathbf{x}_2^T \Sigma^{-1} \mathbf{x}_2 \right) \exp \left(\mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_2 \right)$$

$$\underbrace{\phi(\mathbf{x}) = \exp \left(-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right)}_{\text{kernel}} \quad \underbrace{\exp \left(\mathbf{x}_1^T \Sigma^{-1} \mathbf{x}_2 \right)}_{\text{kernel}}$$

kernel

<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2
<input type="checkbox"/>	3

Problem 7: Kernels (Version B) (3 credits)

Let $\Sigma \in \mathbb{R}^{D \times D}$ be a given invertible, positive semi-definite matrix and $a \in \mathbb{R}$ be a given constant. Consider the kernel:

$$k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}, \quad k(\mathbf{x}_1, \mathbf{x}_2) = a^2 \exp \left(-\frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)^\top \Sigma^{-1}(\mathbf{x}_1 - \mathbf{x}_2) \right).$$

0	<input type="checkbox"/>
1	<input type="checkbox"/>
2	<input type="checkbox"/>
3	<input type="checkbox"/>

Prove or disprove that k is a valid kernel.

Hint: If $k_1(\mathbf{x}_1, \mathbf{x}_2)$ then $\exp(k_1(\mathbf{x}_1, \mathbf{x}_2))$ is also a kernel.

Problem 8: Probabilistic inference & SVD (Version A) (4 credits)

Consider a generative model where $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the observed variable and $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^D$ are the model parameters. We assume the following generative process:

$$p(\mathbf{X}|\mathbf{a}, \mathbf{b}) = \prod_{i=1}^N \prod_{j=1}^D p(X_{ij}|\mathbf{a}, \mathbf{b}) = \prod_{i=1}^N \prod_{j=1}^D \mathcal{N}(X_{ij}|\mathbf{a}_i \cdot \mathbf{b}_j, 1).$$

Here $\mathcal{N}(x|\mu, \sigma)$ denotes the density of the normal distribution

$$\mathcal{N}(x|\mu, \sigma) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

a) Suppose you observed \mathbf{X} . Derive a maximum likelihood estimate (MLE) \mathbf{a}^* , \mathbf{b}^* of the parameters \mathbf{a} , \mathbf{b} .

Hint: SVD can be helpful here. No need to take derivatives.

$$\begin{aligned} \ln p(\mathbf{X}|\mathbf{a}, \mathbf{b}) &= \sum_{i=1}^N \sum_{j=1}^D \ln \exp\left(-\frac{(X_{ij} - \mathbf{a}_i \cdot \mathbf{b}_j)^2}{2}\right) \\ &= -\frac{1}{2} \sum_{i,j} (X_{ij} - \mathbf{a}_i \cdot \mathbf{b}_j)^2 \\ &= -\frac{1}{2} (\mathbf{X} - \mathbf{a} \cdot \mathbf{b}^T)^T (\mathbf{X} - \mathbf{a} \cdot \mathbf{b}^T) \end{aligned}$$

$$\begin{aligned} &= -\frac{1}{2} \|\mathbf{X} - \mathbf{a} \mathbf{b}^T\|_F^2 \\ &\mathbf{B} = \mathbf{U} \Sigma \mathbf{V}^T \\ &= \begin{matrix} \uparrow & \uparrow \\ \sqrt{\sigma_1} \mathbf{u}_1 & \sqrt{\sigma_1} \mathbf{v}_1 \end{matrix} \end{aligned}$$

0
1
2

b) Suppose all singular values of the observed matrix \mathbf{X} are distinct. Is the MLE of the parameters \mathbf{a} , \mathbf{b} unique in this case? Justify your answer.

$$\frac{1}{c} \mathbf{a} \quad c \mathbf{b}$$

0
1
2

Problem 8: Probabilistic inference & SVD (Version B) (4 credits)

Consider a generative model where $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the observed variable and $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^D$ are the model parameters. We assume the following generative process:

$$p(\mathbf{X}|\mathbf{a}, \mathbf{b}) = \prod_{i=1}^N \prod_{j=1}^D p(X_{ij}|\mathbf{a}, \mathbf{b}) = \prod_{i=1}^N \prod_{j=1}^D \mathcal{N}(X_{ij}|\mathbf{a}_i \cdot \mathbf{b}_j, 1).$$

Here $\mathcal{N}(x|\mu, \sigma)$ denotes the density of the normal distribution

$$\mathcal{N}(x|\mu, \sigma) \propto \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

- 0 ☐
- 1 ☐
- 2 ☐
- a) Suppose you observed \mathbf{X} . Derive a maximum likelihood estimate (MLE) \mathbf{a}^* , \mathbf{b}^* of the parameters \mathbf{a} , \mathbf{b} .
Hint: SVD can be helpful here. No need to take derivatives.
- 0 ☐
- 1 ☐
- 2 ☐
- b) Suppose all singular values of the observed matrix \mathbf{X} are distinct. Is the MLE of the parameters \mathbf{a} , \mathbf{b} unique in this case? Justify your answer.

Problem 9: Dimensionality Reduction (Version A) (2 credits)

Figure 24.1 shows a scatter plot of your two-dimensional data ($N = 13$ instances). You want to apply a non-linear dimensionality reduction technique based on neighbor graphs (e.g. T-SNE or UMAP). As a first step you compute the $N \times N$, weighted adjacency matrix representing the neighbor graph. Assume that the weights are computed with

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma^2\right)}$$

where $\mathbf{x}_i \in \mathbb{R}^2$ and you set $p_{i|i} = 0$. Finally, you obtain the similarity between instances i and j with $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$.

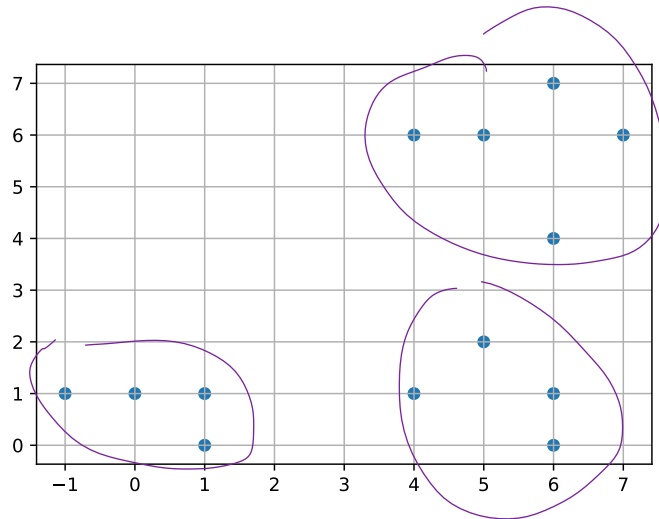


Figure 24.1: Scatter plot of the data

Which of the following neighbor graph plots (pixel in position i, j shows the value of p_{ij}) corresponds to the given dataset and the stated formula for $\sigma = 2$? What is your answer for $\sigma = 5$? Justify your answers!

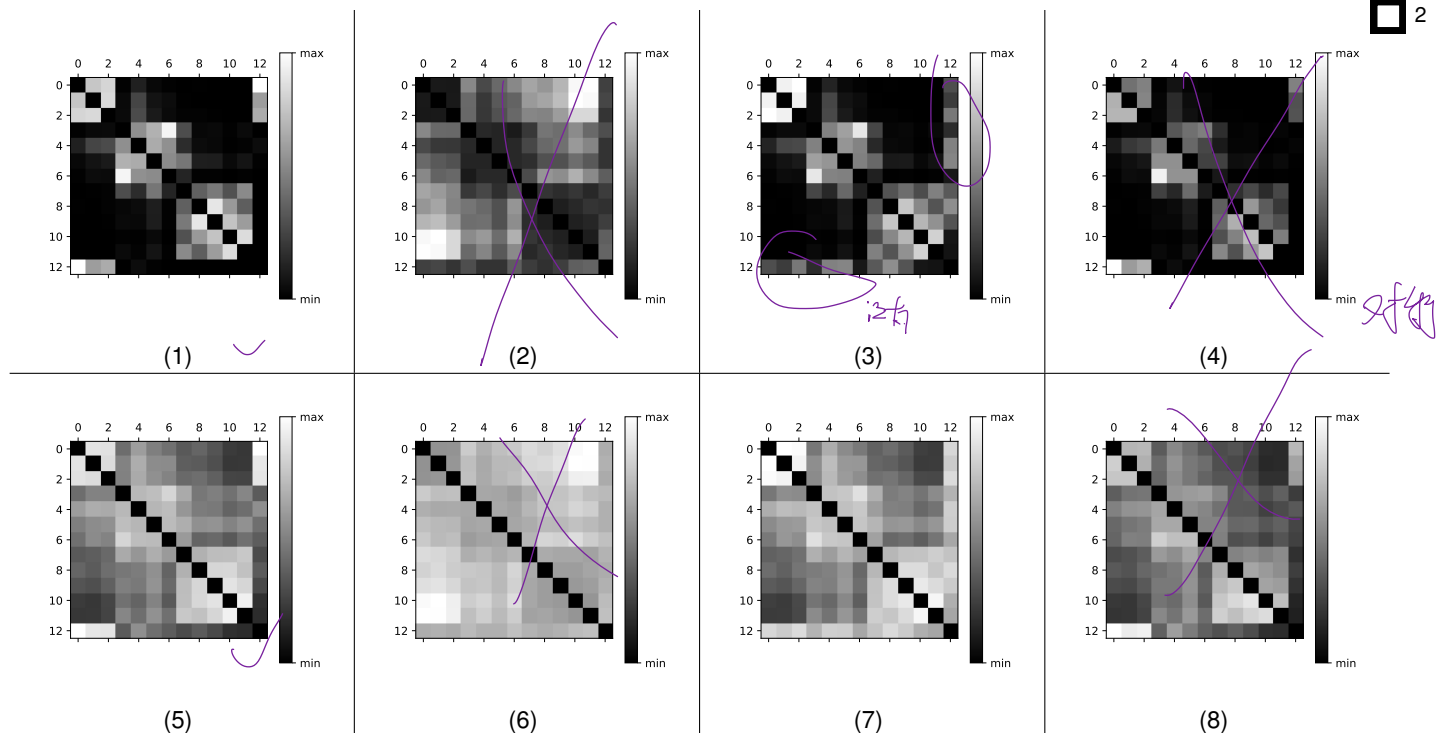
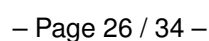


Figure 25.1 shows a scatter plot of your two-dimensional data ($N = 13$ instances). You want to apply a non-linear dimensionality reduction technique based on neighbor graphs (e.g. T-SNE or UMAP). As a first step you compute the $N \times N$, weighted adjacency matrix representing the neighbor graph. Assume that the weights are computed with

where $\mathbf{x}_i \in \mathbb{R}^2$ and you set $p_{iji} = 0$. Finally, you obtain the similarity between instances i and j with $p_{ij} = \frac{p_{ilj} + p_{jli}}{2}$.



0	
1	
2	



Problem 9: Dimensionality Reduction (Version C) (2 credits)

Figure 26.1 shows a scatter plot of your two-dimensional data ($N = 13$ instances). You want to apply a non-linear dimensionality reduction technique based on neighbor graphs (e.g. T-SNE or UMAP). As a first step you compute the $N \times N$, weighted adjacency matrix representing the neighbor graph. Assume that the weights are computed with

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma^2\right)}$$

where $\mathbf{x}_i \in \mathbb{R}^2$ and you set $p_{i|i} = 0$. Finally, you obtain the similarity between instances i and j with $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$.

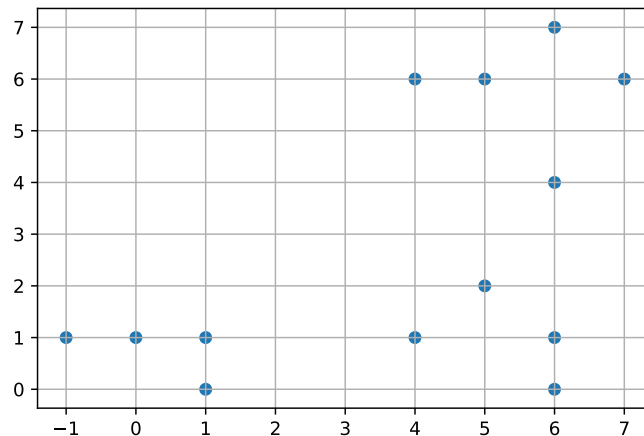
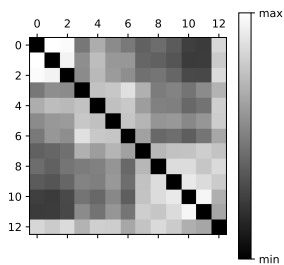
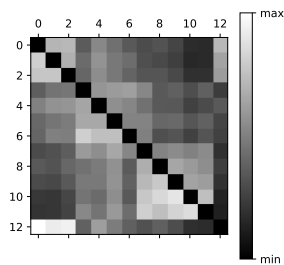


Figure 26.1: Scatter plot of the data

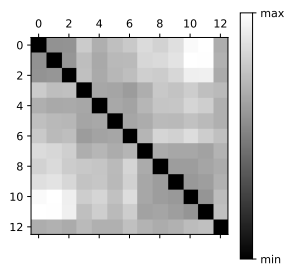
Which of the following neighbor graph plots (pixel in position i, j shows the value of p_{ij}) corresponds to the given dataset and the stated formula for $\sigma = 2$? What is your answer for $\sigma = 5$? *Justify your answers!*



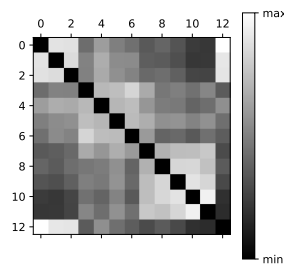
(1)



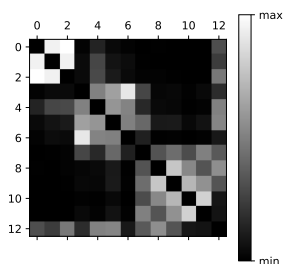
(2)



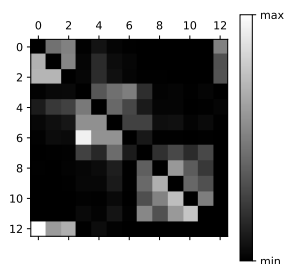
(3)



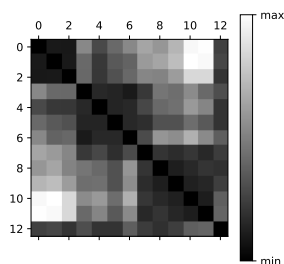
(4)



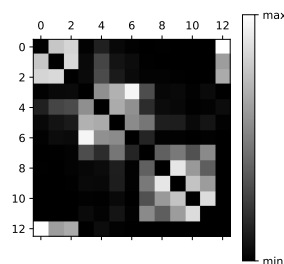
(5)



(6)



(7)



(8)

Problem 9: Dimensionality Reduction (Version D) (2 credits)

Figure 27.1 shows a scatter plot of your two-dimensional data ($N = 13$ instances). You want to apply a non-linear dimensionality reduction technique based on neighbor graphs (e.g. T-SNE or UMAP). As a first step you compute the $N \times N$, weighted adjacency matrix representing the neighbor graph. Assume that the weights are computed with

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma^2\right)}$$

where $\mathbf{x}_i \in \mathbb{R}^2$ and you set $p_{i|i} = 0$. Finally, you obtain the similarity between instances i and j with $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$.

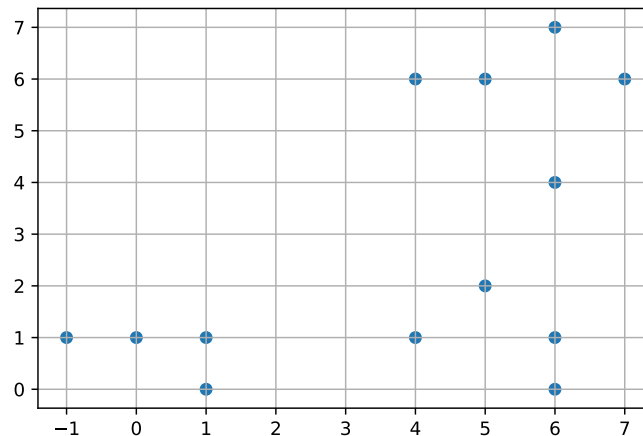
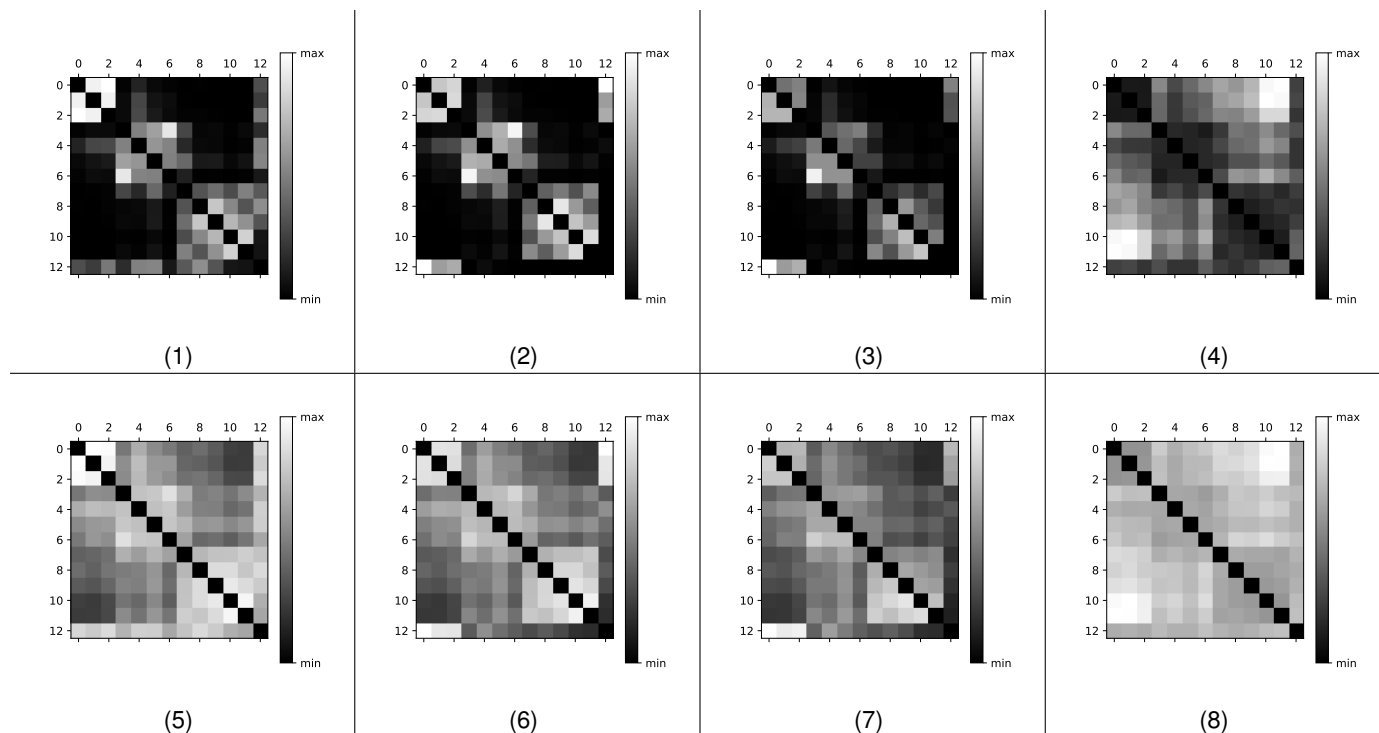


Figure 27.1: Scatter plot of the data

0 ☐
1 ☐
2 ☐

Which of the following neighbor graph plots (pixel in position i, j shows the value of p_{ij}) corresponds to the given dataset and the stated formula for $\sigma = 2$? What is your answer for $\sigma = 5$? *Justify your answers!*



Problem 10: Clustering (Version A) (4 credits)

a) Consider the K -means algorithm with two clusters and fixed centroids μ_1, μ_2 . Prove that the decision boundary is a hyperplane.

Hint: Consider the equation that defines the decision boundary.

b) Now consider the Gaussian mixture model (GMM) with two clusters. In this case the assigned cluster label for instance \mathbf{x}_i is given by

$$\arg \max_{k \in \{1,2\}} \gamma(\mathbf{z}_{ik}). \quad (1)$$

where $\gamma(\mathbf{z}_{ik})$ are the responsibilities. Is the decision boundary in a GMM with two clusters and fixed parameters π_k, μ_k, Σ_k linear in general? If yes, give the associated hyperplane. If no, specify the conditions on the parameters such that the decision boundary is linear *if and only if* the conditions hold. Justify your answer.

a) For \mathbf{x}_i is on the boundary

$$\cancel{2} \cancel{1} \cancel{1} \|\mathbf{x}_i - \mu_1\|_2^2 = \cancel{2} \cancel{1} \cancel{2} \|\mathbf{x}_i - \mu_2\|_2^2$$

$$\cancel{x_1^2} - 2x_1\mu_1 + \mu_1^2 = \cancel{x_1^2} - 2x_1\mu_2 + \mu_2^2$$

$$\mu_1^2 - \mu_2^2 = 2x_1(\mu_1 - \mu_2)$$

$$\mu_1 \neq \mu_2$$

$$2x_1 = \mu_1 + \mu_2$$

$$x_1 = \frac{\mu_1 + \mu_2}{2} \Rightarrow \text{linear boundary, hyper plane}$$

$$\|\mathbf{x}_1 - \mu_1\|_2^2 = \|\mathbf{x}_1 - \mu_2\|_2^2$$

$$(\mathbf{x}_1 - \mu_1)^T (\mathbf{x}_1 - \mu_1) = (\mathbf{x}_1 - \mu_2)^T (\mathbf{x}_1 - \mu_2)$$

$$\cancel{x_1^T x_1} - 2x_1^T \mu_1 + \mu_1^T \mu_1 = \cancel{x_1^T x_1} - 2x_1^T \mu_2 + \mu_2^T \mu_2$$

$$\mu_1^T \mu_1 - \mu_2^T \mu_2 = 2x_1^T (\mu_1 - \mu_2)$$

$$2x_1^T (\mu_1 - \mu_2) - \mu_1^T \mu_1 + \mu_2^T \mu_2 = 0$$

$$x^T w + b = 0 \Rightarrow \text{hyperplane}$$

$$w = 2(\mu_1 - \mu_2) \quad b = \mu_2^T \mu_2 - \mu_1^T \mu_1$$

boundary $\gamma(\mathbf{z}_{i1}) = \gamma(\mathbf{z}_{i2})$

:

$$x^T \Sigma_1^{-1} x = x^T \Sigma_2^{-1} x$$

Problem 10: Clustering (Version B) (4 credits)

0 ☐ a) Consider the K -means algorithm with two clusters and fixed centroids μ_1, μ_2 . Prove that the decision boundary is a hyperplane.

1 ☐ *Hint:* Consider the equation that defines the decision boundary.

2 ☐

0 ☐ b) Now consider the Gaussian mixture model (GMM) with two clusters. In this case the assigned cluster label for instance \mathbf{x}_i is given by

$$\arg \max_{k \in \{1,2\}} \gamma(\mathbf{z}_{ik}). \quad (2)$$

1 ☐ where $\gamma(\mathbf{z}_{ik})$ are the responsibilities. Is the decision boundary in a GMM with two clusters and fixed parameters π_k, μ_k, Σ_k linear in general? If yes, give the associated hyperplane. If no, specify the conditions on the parameters such that the decision boundary is linear *if and only if* the conditions hold. Justify your answer.

2 ☐

Problem 11: Fairness (Version A) (4 credits)

You are given data as shown on Table 30.1 where $X \in \mathbb{R}$ denotes the non-sensitive feature, $A \in \{a, b\}$ denotes the sensitive feature, and $Y \in \{0, 1\}$ denotes the ground-truth label.

Table 30.1: Fairness Data (each column is one data point)

ID	1	2	3	4	5	6	7
X	0.5	-1.0	-0.5	2.0	0.5	1.5	0.1
A	a	b	b	a	b	a	b
Y	1	1	0	0	0	0	1

a) Let the prediction $R = r(X)$ be some arbitrary function r that only depends on X . The sensitive attribute A is ignored. Can we conclude that the *Sufficiency* fairness criterion is satisfied for the data shown on Table 30.1? Justify your answer.

☐ 0
☐ 1

b) Let the prediction $R \in \{0, 1\}$ be

$$R = \begin{cases} 0 & \text{if } 2 \cdot X > 2 \text{ and } A = a \\ 0 & \text{if } 4 \cdot X > 1 \text{ and } A = b \\ 1 & \text{otherwise} \end{cases}$$

☐ 0
☐ 1
☐ 2

Which ones of the following three fairness criteria *Independence*, *Separation*, and *Equality of Opportunity* are satisfied for the data shown on Table 30.1? Justify your answer.

c) Modify the *least* number of instances such that none of the above criteria are satisfied. You can only modify the non-sensitive features X . Write down the ID(s) of the modified instance(s) and their modified X value. Justify your answer!

☐ 0
☐ 1

Problem 11: Fairness (Version B) (4 credits)

You are given data as shown on Table 31.1 where $X \in \mathbb{R}$ denotes the non-sensitive feature, $A \in \{a, b\}$ denotes the sensitive feature, and $Y \in \{0, 1\}$ denotes the ground-truth label.

Table 31.1: Fairness Data (each column is one data point)

ID	1	2	3	4	5	6	7
X	0.5	-1.0	-0.5	2.0	0.5	1.5	0.1
A	b	a	a	b	a	b	a
Y	1	1	0	0	0	0	1

- 0 ☐ a) Let the prediction $R = r(X)$ be some arbitrary function r that only depends on X . The sensitive attribute A is ignored. Can we conclude that the *Sufficiency* fairness criterion is satisfied for the data shown on Table 31.1? Justify your answer.
- 1 ☐

- 0 ☐ b) Let the prediction $R \in \{0, 1\}$ be

$$R = \begin{cases} 0 & \text{if } 2 \cdot X > 0.5 \text{ and } A = a \\ 0 & \text{if } 3 \cdot X > 3.0 \text{ and } A = b \\ 1 & \text{otherwise} \end{cases}$$

- 1 ☐
- 2 ☐ Which ones of the following three fairness criteria *Independence*, *Separation*, and *Equality of Opportunity* are satisfied for the data shown on Table 31.1? Justify your answer.

- 0 ☐ c) Modify the *least* number of instances such that none of the above criteria are satisfied. You can only modify the non-sensitive features X . Write down the ID(s) of the modified instance(s) and their modified X value. Justify your answer!
- 1 ☐

Problem 11: Fairness (Version C) (4 credits)

You are given data as shown on Table 32.1 where $X \in \mathbb{R}$ denotes the non-sensitive feature, $A \in \{a, b\}$ denotes the sensitive feature, and $Y \in \{0, 1\}$ denotes the ground-truth label.

Table 32.1: Fairness Data (each column is one data point)

ID	1	2	3	4	5	6	7
X	0.5	-1.0	-0.5	2.0	0.5	1.5	0.1
A	a	b	b	a	b	a	b
Y	1	1	0	0	0	0	1

a) Let the prediction $R = r(X)$ be some arbitrary function r that only depends on X . The sensitive attribute A is ignored. Can we conclude that the *Sufficiency* fairness criterion is satisfied for the data shown on Table 32.1? Justify your answer.

☐

☐

0

1

b) Let the prediction $R \in \{0, 1\}$ be

$$R = \begin{cases} 0 & \text{if } 2 \cdot X > 2 \text{ and } A = a \\ 0 & \text{if } 4 \cdot X > 1 \text{ and } A = b \\ 1 & \text{otherwise} \end{cases}$$

☐

☐

☐

☐

0

1

2

Which ones of the following three fairness criteria *Independence*, *Separation*, and *Equality of Opportunity* are satisfied for the data shown on Table 32.1? Justify your answer.

c) Modify the *least* number of instances such that none of the above criteria are satisfied. You can only modify the non-sensitive features X . Write down the ID(s) of the modified instance(s) and their modified X value. Justify your answer!

☐

☐

☐

0

1

Problem 11: Fairness (Version D) (4 credits)

You are given data as shown on Table 33.1 where $X \in \mathbb{R}$ denotes the non-sensitive feature, $A \in \{a, b\}$ denotes the sensitive feature, and $Y \in \{0, 1\}$ denotes the ground-truth label.

Table 33.1: Fairness Data (each column is one data point)

ID	1	2	3	4	5	6	7
X	0.5	-1.0	-0.5	2.0	0.5	1.5	0.1
A	b	a	a	b	a	b	a
Y	1	1	0	0	0	0	1

- 0 ☐ a) Let the prediction $R = r(X)$ be some arbitrary function r that only depends on X . The sensitive attribute A is ignored. Can we conclude that the *Sufficiency* fairness criterion is satisfied for the data shown on Table 33.1? Justify your answer.
- 1 ☐

- 0 ☐ b) Let the prediction $R \in \{0, 1\}$ be

$$R = \begin{cases} 0 & \text{if } 2 \cdot X > 0.5 \text{ and } A = a \\ 0 & \text{if } 3 \cdot X > 3.0 \text{ and } A = b \\ 1 & \text{otherwise} \end{cases}$$

- 1 ☐
- 2 ☐ Which ones of the following three fairness criteria *Independence*, *Separation*, and *Equality of Opportunity* are satisfied for the data shown on Table 33.1? Justify your answer.

- 0 ☐ c) Modify the *least* number of instances such that none of the above criteria are satisfied. You can only modify the non-sensitive features X . Write down the ID(s) of the modified instance(s) and their modified X value. Justify your answer!
- 1 ☐