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

Esolution
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 / Endterm	**Date:**	Tuesday 16th February, 2021
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	11:00 – 13:00

Working instructions

- This graded exercise consists of **? pages** with a total of **23 problems**.
  Please make sure now that you received a complete copy of the answer sheet.
- The total amount of achievable credits in this graded exercise is 107 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**
- You have to sign the code of conduct. (Typing your name is fine)
- You have to either print this document and scan your solutions or paste scans/pictures of your handwritten solutions into the solution boxes in this PDF. **Editing the PDF digitally is prohibited except for signing the code of conduct and answering multiple choice questions**
- Make sure that the **QR codes are visible** on every uploaded page. Otherwise, we cannot grade your submission.
- **You must solve the specified version of the problem**. Different problems may have different version: e.g. Problem 1 (Version A), Problem 5 (Version C), etc. If you solve the wrong version you get **zero** points.
- 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 considered empty.
- **Only use a black or blue color (no red or green)! Pencils are allowed.**
- 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.
- Instructor announcements and clarifications will be posted **on Piazza** with email notifications.
- Exercise duration - 120 minutes.

Left room from ___________ to ___________ / Early submission at ___________

Problem 1 (Version A) (4 credits)

0 1 2 3 4

We have to find the most likely value s^* of s after incorporating the observations of t, i.e. the maximum a posteriori estimate.

$$
\begin{aligned}
s^* &= \arg\max_s \mathrm{p}(\mathcal{D} \mid s)\ \mathrm{p}(s) \\
&= \arg\max_s \log \mathrm{p}(\mathcal{D} \mid s) + \log \mathrm{p}(s) \\
&= \arg\max_s \sum_{i=1}^{N} \log s^2 \exp(-s^2 t_i) + \log \exp(-s^2) \\
&= \arg\max_s N \log s^2 - s^2 \sum_{i=1}^{N} t_i - s^2 \\
&= \arg\max_s N \log s^2 - s^2(T+1) \text{ where } T = \sum_{i=1}^{N} t_i
\end{aligned}
$$

This expression is symmetric in the sign of s, so we can restrict ourselves to the case of $s \geq 0$. On this restricted domain, the expression is also concave in s, so we can find the maximum by differentiation.

$$
\frac{\partial}{\partial s} N \log s^2 - s^2(T+1) = \frac{2N}{s} - 2(T+1)s = 0 \Leftrightarrow s = \pm\sqrt{\frac{N}{T+1}}
$$

Summing the observations, we get $T = 19$ and so the positive most likely severity of the disease is $s^* = \sqrt{\frac{5}{19+1}} = \sqrt{\frac{1}{4}} = \frac{1}{2}$.

Note: The problem description had a small mistake and depending on if the students worked with $\exp(-s^2)$ or $\exp\left(-\frac{s^2}{2}\right)$, the students might also have arrived at

$$
s^* = \arg\max_s N \log s^2 - s^2(T + \frac{1}{2}).
$$

Then their end result would be $s^* = \sqrt{\frac{5}{19+\frac{1}{2}}} = \sqrt{\frac{10}{39}} \approx 0.506$.

Problem 1 (Version B) (4 credits)

0
1
2
3
4

We have to find the most likely value s^* of s after incorporating the observations of t, i.e. the maximum a posteriori estimate.

$$
\begin{aligned}
s^* &= \arg\max_s \mathrm{p}(\mathcal{D} \mid s)\ \mathrm{p}(s) \\
&= \arg\max_s \log \mathrm{p}(\mathcal{D} \mid s) + \log \mathrm{p}(s) \\
&= \arg\max_s \sum_{i=1}^{N} \log s^2 \exp(-s^2 t_i) + \log \exp(-s^2) \\
&= \arg\max_s N \log s^2 - s^2 \sum_{i=1}^{N} t_i - s^2 \\
&= \arg\max_s N \log s^2 - s^2(T+1) \text{ where } T = \sum_{i=1}^{N} t_i
\end{aligned}
$$

This expression is symmetric in the sign of s, so we can restrict ourselves to the case of $s \geq 0$. On this restricted domain, the expression is also concave in s, so we can find the maximum by differentiation.

$$
\frac{\partial}{\partial s} N \log s^2 - s^2(T+1) = \frac{2N}{s} - 2(T+1)s = 0 \Leftrightarrow s = \pm\sqrt{\frac{N}{T+1}}
$$

Summing the observations, we get $T = 26$ and so the positive most likely severity of the disease is $s^* = \sqrt{\frac{3}{26+1}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$.

Note: The problem description had a small mistake and depending on if the students worked with $\exp(-s^2)$ or $\exp\left(-\frac{s^2}{2}\right)$, the students might also have arrived at

$$
s^* = \arg\max_s N \log s^2 - s^2(T + \frac{1}{2})
$$

Then their end result would be $s^* = \sqrt{\frac{3}{26+\frac{1}{2}}} = \sqrt{\frac{6}{53}} \approx 0.336$.

Problem 1 (Version C) (4 credits)

0 1 2 3 4

We have to find the most likely value s^* of s after incorporating the observations of t, i.e. the maximum a posteriori estimate.

$$
\begin{aligned}
s^* &= \arg\max_s \mathrm{p}(\mathcal{D} \mid s)\ \mathrm{p}(s) \\
&= \arg\max_s \log \mathrm{p}(\mathcal{D} \mid s) + \log \mathrm{p}(s) \\
&= \arg\max_s \sum_{i=1}^{N} \log s^2 \exp(-s^2 t_i) + \log \exp(-s^2) \\
&= \arg\max_s N \log s^2 - s^2 \sum_{i=1}^{N} t_i - s^2 \\
&= \arg\max_s N \log s^2 - s^2 (T+1) \text{ where } T = \sum_{i=1}^{N} t_i
\end{aligned}
$$

This expression is symmetric in the sign of s, so we can restrict ourselves to the case of $s \geq 0$. On this restricted domain, the expression is also concave in s, so we can find the maximum by differentiation.

$$
\frac{\partial}{\partial s} N \log s^2 - s^2(T+1) = \frac{2N}{s} - 2(T+1)s = 0 \Leftrightarrow s = \pm\sqrt{\frac{N}{T+1}}
$$

Summing the observations, we get $T = 35$ and so the positive most likely severity of the disease is $s^* = \sqrt{\frac{4}{35+1}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$.

Note: The problem description had a small mistake and depending on if the students worked with $\exp(-s^2)$ or $\exp\left(-\frac{s^2}{2}\right)$, the students might also have arrived at

$$
s^* = \arg\max_s N \log s^2 - s^2 (T + \frac{1}{2}).
$$

Then their end result would be $s^* = \sqrt{\frac{4}{35+\frac{1}{2}}} = \sqrt{\frac{8}{71}} \approx 0.336$.

Problem 2 (Version A) (4 credits)

a)

0 1 2

The value of $k = 5$ minimizes the LOOCV error.The error is $4/13$.

b)

0 1

Yes. One counter example is that (1, 3) was previously labeled with **-** but is now labeled with **+** since the new point (1, 2) is closest.

c)

0 1

We need to move it to the closest data point with the same label to keep the decision boundary the same. In this case this is (5, 1). The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

Problem 2 (Version B) (4 credits)

0 1 2

a)

The value of $k = 5$ minimizes the LOOCV error.The error is $4/13$.

0 1

b)

Yes. One counter example is that (2, 2) was previously labeled with **-** but is now labeled with **+** since the new point (1, 2) is closest.

0 1

c)

We need to move it to the closest data point with the same label to keep the decision boundary the same. In this case this is (2, 6). The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

Problem 2 (Version C) (4 credits)

a)

The value of $k = 5$ minimizes the LOOCV error.The error is $4/13$.

0 1 2

b)

Yes. One counter example is that (2, 2) was previously labeled with **+** but is now labeled with **-** since the new point (1, 2) is closest.

0 1

c)

We need to move it to the closest data point with the same label to keep the decision boundary the same. In this case this is (2, 6). The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

0 1

Problem 2 (Version D) (4 credits)

a)

The value of $k = 5$ minimizes the LOOCV error.The error is $4/13$.

b)

Yes. One counter example is that $(1, 3)$ was previously labeled with **+** but is now labeled with **-** since the new point $(1, 2)$ is closest.

c)

We need to move it to the closest data point with the same label to keep the decision boundary the same. In this case this is $(5, 1)$. The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

Problem 3 (Version A) (6 credits)

a)

0 1 2 3

Because of convexity, we can find the optimal w_{D+1} by finding the zero of the derivative.

$$\frac{\partial}{\partial w_{D+1}} J(\boldsymbol{w}) = \sum_{i=1}^{N} \left(\boldsymbol{w}^{\mathrm{T}} \tilde{\boldsymbol{x}}^{(i)} - y^{(i)} \right) + \lambda w_{D+1}$$

$$= \boldsymbol{w}_{1:D}^{\mathrm{T}} \sum_{i=1}^{N} \boldsymbol{x}^{(i)} + w_{D+1} \sum_{i=1}^{N} 1 - \sum_{i=1}^{N} y^{(i)} + \lambda w_{D+1}$$

$\sum_{i=1}^{N} \boldsymbol{x}^{(i)}$ is zero because we have assumed that the $\boldsymbol{x}^{i}$ are centered.

$$= N w_{D+1} - \sum_{i=1}^{N} y^{(i)} + \lambda w_{D+1} = (N + \lambda) w_{D+1} - \sum_{i=1}^{N} y^{(i)}$$

Solving for w_{D+1} we get

$$w_{D+1} = \frac{1}{N+\lambda} \sum_{i=1}^{N} y^{(i)}.$$

b)

0 1 2 3

We propose a biased centering of the regression targets, i.e.

$$\widehat{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i)} \quad \text{and} \quad \widehat{y}^{(i)} = y^{(i)} - \frac{1}{N+\lambda} \sum_{j=1}^{N} y^{(j)}.$$

The ridge regression loss evaluated on $\widetilde{\mathcal{D}}$ is

$$\mathcal{L}(\widetilde{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widetilde{\boldsymbol{w}}_{1:D}^{\mathrm{T}} \boldsymbol{x}^{(i)} + \widetilde{\boldsymbol{w}}_{D+1} - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{w}}\|_2^2 + \frac{\lambda}{2} \widetilde{\boldsymbol{w}}_{D+1}^2.$$

The gradient and therefore the optimal value of $\widetilde{\boldsymbol{w}}_{D+1}$ is independent of $\widetilde{\boldsymbol{w}}_{1:D}$, so for the optimal values of $\widetilde{\boldsymbol{w}}_{1:D}$ it is equivalent to minimize $\mathcal{L}$ with $\widetilde{\boldsymbol{w}}_{D+1}^{*}$ plugged in.

$$\mathcal{L}(\widetilde{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widetilde{\boldsymbol{w}}_{1:D}^{\mathrm{T}} \boldsymbol{x}^{(i)} + \left(\frac{1}{N+\lambda} \sum_{j=1}^{N} y^{(j)} \right) - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{w}}_{1:D}\|_2^2 + \frac{\lambda}{2} \left(\frac{1}{N+\lambda} \sum_{i=1}^{N} y^{(i)} \right)^2.$$

The last part has zero gradient with respect to $\widetilde{\boldsymbol{w}}_{1:D}$, so it does not influence the optimal $\widetilde{\boldsymbol{w}}_{1:D}^{*}$ and we can drop it since $\widetilde{\boldsymbol{w}}_{D+1}$ has been eliminated. If we then absorb the $\frac{1}{N+\lambda} \sum_{j=1}^{N} y^{(j)}$ term in the least squares regression sum into $y^{(i)}$, we get the ridge regression loss evaluated on $\widehat{\mathcal{D}}$

$$\mathcal{L}(\widehat{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widehat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{x}^{(i)} - \widehat{y}^{(i)} \right)^2 + \frac{\lambda}{2} \|\widehat{\boldsymbol{w}}\|_2^2.$$

showing that ridge regression on $\widehat{\mathcal{D}}$ is equivalent to ridge regression on $\widetilde{\mathcal{D}}$.

Problem 3 (Version B) (6 credits)

0 1 2 3

a)

Because of convexity, we can find the optimal w_{D+1} by finding the zero of the derivative.

$$\frac{\partial}{\partial w_{D+1}} J(\boldsymbol{w}) = \sum_{i=1}^{N} \left(\boldsymbol{w}^{\mathsf{T}} \tilde{\boldsymbol{x}}^{(i)} - y^{(i)} \right) + \lambda w_{D+1}$$

$$= \boldsymbol{w}_{1:D}^{\mathsf{T}} \sum_{i=1}^{N} \boldsymbol{x}^{(i)} + w_{D+1} \sum_{i=1}^{N} 1 - \sum_{i=1}^{N} y^{(i)} + \lambda w_{D+1}$$

$\sum_{i=1}^{N} \boldsymbol{x}^{(i)}$ is zero because we have assumed that the $\boldsymbol{x}^{i}$ are centered.

$$= N w_{D+1} - \sum_{i=1}^{N} y^{(i)} + \lambda w_{D+1} = (N + \lambda) w_{D+1} - \sum_{i=1}^{N} y^{(i)}$$

Solving for w_{D+1} we get

$$w_{D+1} = \frac{1}{N + \lambda} \sum_{i=1}^{N} y^{(i)}.$$

0 1 2 3

b)

We propose a biased centering of the regression targets, i.e.

$$\widehat{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i)} \quad \text{and} \quad \widehat{y}^{(i)} = y^{(i)} - \frac{1}{N + \lambda} \sum_{j=1}^{N} y^{(j)}.$$

The ridge regression loss evaluated on $\widetilde{\mathcal{D}}$ is

$$\mathcal{L}(\widetilde{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widetilde{\boldsymbol{w}}_{1:D}^{\mathsf{T}} \boldsymbol{x}^{(i)} + \widetilde{w}_{D+1} - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{w}}\|_2^2 + \frac{\lambda}{2} \widetilde{w}_{D+1}^2.$$

The gradient and therefore the optimal value of $\widetilde{w}_{D+1}$ is independent of $\widetilde{\boldsymbol{w}}_{1:D}$, so for the optimal values of $\widetilde{\boldsymbol{w}}_{1:D}$ it is equivalent to minimize $\mathcal{L}$ with $\widetilde{w}_{D+1}^{*}$ plugged in.

$$\mathcal{L}(\widetilde{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widetilde{\boldsymbol{w}}_{1:D}^{\mathsf{T}} \boldsymbol{x}^{(i)} + \left(\frac{1}{N + \lambda} \sum_{j=1}^{N} y^{(j)} \right) - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{w}}_{1:D}\|_2^2 + \frac{\lambda}{2} \left(\frac{1}{N + \lambda} \sum_{i=1}^{N} y^{(i)} \right)^2.$$

The last part has zero gradient with respect to $\widetilde{\boldsymbol{w}}_{1:D}$, so it does not influence the optimal $\widetilde{\boldsymbol{w}}_{1:D}^{*}$ and we can drop it since $\widetilde{w}_{D+1}$ has been eliminated. If we then absorb the $\frac{1}{N+\lambda} \sum_{j=1}^{N} y^{(j)}$ term in the least squares regression sum into $y^{(i)}$, we get the ridge regression loss evaluated on $\widehat{\mathcal{D}}$

$$\mathcal{L}(\widehat{\boldsymbol{w}}) = \frac{1}{2} \sum_{i=1}^{N} \left(\widehat{\boldsymbol{w}}^{\mathsf{T}} \boldsymbol{x}^{(i)} - \widehat{y}^{(i)} \right)^2 + \frac{\lambda}{2} \|\widehat{\boldsymbol{w}}\|_2^2.$$

showing that ridge regression on $\widehat{\mathcal{D}}$ is equivalent to ridge regression on $\widetilde{\mathcal{D}}$.

Problem 4 (Version A) (6 credits)

a)

In a naive Bayes classifier, the features are independent, so we can choose a different probability distribution for each of them. We choose to model the continuous feature as a normal distribution, $x_1 \mid y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 \mid y = c \sim \text{Bernoulli}(\alpha_c)$ where yes is 1, no is 0 and α_c gives the success probability.
The distribution of the classes y is a categorical distribution with parameter π, $y \sim \text{Categorical}(\pi)$.
The maximum likelihood estimates of the parameters are

$$\pi = \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{pmatrix}^T$$

$$\mu_1 = 1 \quad \mu_2 = 0 \quad \mu_3 = 5$$

$$\alpha_1 = \frac{1}{2} \quad \alpha_2 = \frac{1}{3} \quad \alpha_3 = 1$$

b)

The unnormalized posterior is $p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_1^{(b)} \mid y^{(b)})\, p(\boldsymbol{x}_2^{(b)} \mid y^{(b)})\, p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

$$p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto \begin{pmatrix} e^0 \frac{1}{2}\frac{2}{7} & e^{-\frac{1}{2}}\frac{1}{3}\frac{3}{7} & e^{-8}\, 1\, \frac{2}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7} & \frac{1}{7\sqrt{e}} & \frac{2}{7e^8} \end{pmatrix}^T$$

c)

We do not know anything about this data point, so the posterior distribution is just the prior distribution.

$$p(y^{(c)} \mid \boldsymbol{x}^{(c)}) = p(y) = \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{pmatrix}^T$$

d)

Since we only know the feature $x_2^{(d)}$, we only condition on that and get $p(y^{(d)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_2^{(b)} \mid y^{(d)})\, p(y^{(d)})$.

$$p(y^{(d)} \mid \boldsymbol{x}^{(d)}) = \begin{pmatrix} \frac{1}{2}\frac{2}{7} & \frac{2}{3}\frac{3}{7} & 0\frac{2}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} & 0 \end{pmatrix}^T$$

Problem 4 (Version B) (6 credits)

0 1 2 3

a)

In a naive Bayes classifier, the features are independent, so we can choose a different probability distribution for each of them. We choose to model the continuous feature as a normal distribution, $x_1 \mid y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 \mid y = c \sim \text{Bernoulli}(\alpha_c)$ where yes is 1, no is 0 and α_c gives the success probability.
The distribution of the classes y is a categorical distribution with parameter π, $y \sim \text{Categorical}(\pi)$.
The maximum likelihood estimates of the parameters are

$$\pi = \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{pmatrix}^T$$

$$\mu_1 = 1 \quad \mu_2 = 0 \quad \mu_3 = 5$$

$$\alpha_1 = \frac{1}{2} \quad \alpha_2 = \frac{1}{3} \quad \alpha_3 = 1$$

0 1

b)

The unnormalized posterior is $p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_1^{(b)} \mid y^{(b)})\, p(\boldsymbol{x}_2^{(b)} \mid y^{(b)})\, p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

$$p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto \begin{pmatrix} e^{-\frac{1}{2}}\frac{1}{2}\frac{2}{7} & e^{-2}\frac{1}{3}\frac{3}{7} & e^{-\frac{9}{2}}1\frac{2}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7\sqrt{e}} & \frac{1}{7e^2} & \frac{2}{7e^{\frac{9}{2}}} \end{pmatrix}^T$$

0 1

c)

We do not know anything about this data point, so the posterior distribution is just the prior distribution.

$$p(y^{(c)} \mid \boldsymbol{x}^{(c)}) = p(y) = \begin{pmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{pmatrix}^T$$

0 1

d)

Since we only know the feature $x_2^{(d)}$, we only condition on that and get $p(y^{(d)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_2^{(b)} \mid y^{(d)})\, p(y^{(d)})$.

$$p(y^{(d)} \mid \boldsymbol{x}^{(d)}) = \begin{pmatrix} \frac{1}{2}\frac{2}{7} & \frac{2}{3}\frac{3}{7} & 0\frac{2}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} & 0 \end{pmatrix}^T$$

Problem 4 (Version C) (6 credits)

a)

In a naive Bayes classifier, the features are independent, so we can choose a different probability distribution for each of them. We choose to model the continuous feature as a normal distribution, $x_1 \mid y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 \mid y = c \sim \text{Bernoulli}(\alpha_c)$ where yes is 1, no is 0 and α_c gives the success probability.
The distribution of the classes y is a categorical distribution with parameter π, $y \sim \text{Categorical}(\pi)$.
The maximum likelihood estimates of the parameters are

$$\pi = \begin{pmatrix} \frac{2}{7} & \frac{2}{7} & \frac{3}{7} \end{pmatrix}^T$$

$$\mu_1 = -2 \quad \mu_2 = 2 \quad \mu_3 = 4$$

$$\alpha_1 = \frac{1}{2} \quad \alpha_2 = 0 \quad \alpha_3 = \frac{2}{3}$$

b)

The unnormalized posterior is $p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_1^{(b)} \mid y^{(b)})\, p(\boldsymbol{x}_2^{(b)} \mid y^{(b)})\, p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

$$p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto \begin{pmatrix} e^{-\frac{9}{2}}\frac{1}{2}\frac{2}{7} & e^{-\frac{1}{2}}0\frac{2}{7} & e^{-\frac{9}{2}}\frac{2}{3}\frac{3}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7e^{\frac{9}{2}}} & 0 & \frac{2}{7e^{\frac{9}{2}}} \end{pmatrix}^T$$

c)

We do not know anything about this data point, so the posterior distribution is just the prior distribution.

$$p(y^{(c)} \mid \boldsymbol{x}^{(c)}) = p(y) = \begin{pmatrix} \frac{2}{7} & \frac{2}{7} & \frac{3}{7} \end{pmatrix}^T$$

d)

Since we only know the feature $x_2^{(d)}$, we only condition on that and get $p(y^{(d)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_2^{(b)} \mid y^{(d)})\, p(y^{(d)})$.

$$p(y^{(d)} \mid \boldsymbol{x}^{(d)}) = \begin{pmatrix} \frac{1}{2}\frac{2}{7} & 1\frac{2}{7} & \frac{1}{3}\frac{3}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} & \frac{1}{7} \end{pmatrix}^T$$

Problem 4 (Version D) (6 credits)

0 1 2 3

a)

In a naive Bayes classifier, the features are independent, so we can choose a different probability distribution for each of them. We choose to model the continuous feature as a normal distribution, $x_1 \mid y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 \mid y = c \sim \text{Bernoulli}(\alpha_c)$ where yes is 1, no is 0 and α_c gives the success probability.
The distribution of the classes y is a categorical distribution with parameter π, $y \sim \text{Categorical}(\pi)$.
The maximum likelihood estimates of the parameters are

$$\pi = \begin{pmatrix} \frac{2}{7} & \frac{2}{7} & \frac{3}{7} \end{pmatrix}^T$$

$$\mu_1 = -2 \quad \mu_2 = 2 \quad \mu_3 = 4$$

$$\alpha_1 = \frac{1}{2} \quad \alpha_2 = 0 \quad \alpha_3 = \frac{2}{3}$$

0 1

b)

The unnormalized posterior is $p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_1^{(b)} \mid y^{(b)})\, p(\boldsymbol{x}_2^{(b)} \mid y^{(b)})\, p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

$$p(y^{(b)} \mid \boldsymbol{x}^{(b)}) \propto \begin{pmatrix} e^{-8}\frac{1}{2}\frac{2}{7} & e^{0}\,0\,\frac{2}{7} & e^{-2}\frac{2}{3}\frac{3}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7e^8} & 0 & \frac{2}{7e^2} \end{pmatrix}^T$$

0 1

c)

We do not know anything about this data point, so the posterior distribution is just the prior distribution.

$$p(y^{(c)} \mid \boldsymbol{x}^{(c)}) = p(y) = \begin{pmatrix} \frac{2}{7} & \frac{2}{7} & \frac{3}{7} \end{pmatrix}^T$$

0 1

d)

Since we only know the feature $x_2^{(d)}$, we only condition on that and get $p(y^{(d)} \mid \boldsymbol{x}^{(b)}) \propto p(\boldsymbol{x}_2^{(b)} \mid y^{(d)})\, p(y^{(d)})$.

$$p(y^{(d)} \mid \boldsymbol{x}^{(d)}) = \begin{pmatrix} \frac{1}{2}\frac{2}{7} & 1\frac{2}{7} & \frac{1}{3}\frac{3}{7} \end{pmatrix}^T = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} & \frac{1}{7} \end{pmatrix}^T$$

Problem 5 (Version A) (2 credits)

0

1

2

We will prove that $f(\boldsymbol{x})$ is convex using convexity-preserving operations.

$\boldsymbol{a}^T\boldsymbol{x}$ is convex in $\boldsymbol{x}$ and e^z is an increasing convex function. Therefore, their composition $e^{\boldsymbol{a}^T\boldsymbol{x}}$ is convex in $\boldsymbol{x}$.

Similarly, $-\boldsymbol{a}^T\boldsymbol{x}$ is convex in $\boldsymbol{x}$, so $e^{-\boldsymbol{a}^T\boldsymbol{x}}$ is convex in $\boldsymbol{x}$ as well.

$e^{\boldsymbol{a}^T\boldsymbol{x}} + e^{-\boldsymbol{a}^T\boldsymbol{x}}$ is a sum of convex functions, so it's also convex in $\boldsymbol{x}$.

Finally, $\exp(e^{\boldsymbol{a}^T\boldsymbol{x}} + e^{-\boldsymbol{a}^T\boldsymbol{x}})$ is a composition of e^z (an increasing convex function) with another convex function.

Therefore $f(\boldsymbol{x})$ is convex in $\boldsymbol{x}$.

Problem 6 (Version A) (3 credits)

0
1
2
3

The output of conv1 will have shape [32, 16, 8]. Therefore,

- $C_{\text{in}} = 32$ since the output of conv1 has 8 channels.
- $C_{\text{out}} = 16$ we know that the output of the NN has 16 channels.
- $P = 1$ and $S = 1$ since no other combination of P and S will produce an output image with height 16 and width 8, since we don't need to perform downsampling in this layer.

Problem 6 (Version B) (3 credits)

0
1
2
3

The output of `conv1` will have shape [32, 64, 32]. Therefore,

- $C_{\text{in}} = 32$ since the output of `conv1` has 8 channels.
- $C_{\text{out}} = 16$ we know that the output of the NN has 16 channels.
- $P = 1$ (or $P = 0$) and $S = 4$ since no other combination of P and S will produce an output image with height 16 and width 8, i.e., where both dimensions are reduced by a factor of 4.

Problem 6 (Version C) (3 credits)

0
1
2
3

The output of conv1 will have shape [8, 16, 8]. Therefore,

- $C_{\text{in}} = 8$ since the output of conv1 has 8 channels.
- $C_{\text{out}} = 16$ we know that the output of the NN has 16 channels.
- $P = 1$ and $S = 1$ since no other combination of P and S will produce an output image with height 16 and width 8, since we don't need to perform downsampling in this layer.

Problem 6 (Version D) (3 credits)

0
1
2
3

The output of `conv1` will have shape [8, 64, 32]. Therefore,

- $C_{\text{in}} = 8$ since the output of `conv1` has 8 channels.
- $C_{\text{out}} = 16$ we know that the output of the NN has 16 channels.
- $P = 1$ (or $P = 0$) and $S = 4$ since no other combination of P and S will produce an output image with height 16 and width 8, i.e., where both dimensions are reduced by a factor of 4.

Problem 7 (All Versions) (5 credits)

0 1 2 3

a)

Since $\xi_q > 2$ the instance q is misclassified and lies on the wrong side of the decision boundary and it is *outside* of the margin.

The vector $\boldsymbol{w}_{\text{soft}}$ is a *feasible* solution for the new hard-margin SVM, i.e. it satisfies all of the constraints because:

- By removing instance q we remove the corresponding constraint
- All other instances $i \neq q$ satisfy $y_i(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_i + b) \geq 1$ since $\xi_i = 0$

Since we already found one feasible solution, namely $\boldsymbol{w}_{\text{soft}}$ with the corresponding margin $m_{\text{soft}} = \frac{2}{||\boldsymbol{w}_{\text{soft}}||}$, the solution found by the hard-margin SVM with q removed can only be larger.Therefore, $m_{\text{hard}} \geq m_{\text{soft}}$.

0 1 2

b)

Since $\xi_q > 2$ the instance q is misclassified and lies on the wrong side of the decision boundary and it is *outside* of the margin.

As before, the vector $\boldsymbol{w}_{\text{soft}}$ is a *feasible* solution for the new hard-margin SVM, i.e. it satisfies all of the constraints. The constraint for instance q before was $y_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b) \geq 1 - \xi_q$. The optimal solution for

$$\xi_q \begin{cases} 1 - y_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b), & \text{if } y_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b) < 1 \\ 0, & \text{otherwise} \end{cases} .$$

$\xi_q > 2$ implies $y_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b) < -1$. If we now flip the sign of $-y_q = \tilde{y}_q$, we get $\tilde{y}_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b) > 1$. Hence, $\tilde{\xi}_q = 0$ (instances q is now correctly classified and outside the margin). As before, all other instances $i \neq q$ satisfy $y_i(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_i + b) \geq 1$ since $\xi_i = 0$.

Substituting $\xi_q > 2$ we have $y_q(\boldsymbol{w}_{\text{soft}}^T \boldsymbol{x}_q + b) \geq -1$. By relabeling instance q, i.e. multiplying y_q by -1 the hard-margin constraint is satisfied.
Since we already found one feasible solution, namely $\boldsymbol{w}_{\text{soft}}$ with the corresponding margin $m_{\text{soft}} = \frac{2}{||\boldsymbol{w}_{\text{soft}}||}$, the solution found by the hard-margin SVM with q relabeled can only be larger or be as large. Therefore, $m_{\text{hard}} \geq m_{\text{soft}}$ (we also accept $m_{\text{hard}} = m_{\text{soft}}$).

Problem 8 (All Versions) (6 credits)

a)

0 1 2

The training error is 0.Since $\boldsymbol{M}'$ is a rank 1 matrix $\boldsymbol{X}'$ and $\boldsymbol{y}'$ are linearly dependent which means we can perfectly reconstruct $\boldsymbol{y}'$ from $\boldsymbol{X}'$.

b)

0 1 2 3

Since the training error is 0 as we reasoned above we have: $\boldsymbol{w}^*\boldsymbol{X}' + b^* = \boldsymbol{y}'$.

Since $\boldsymbol{M}'$ is the *best* rank 1 approximation of $\boldsymbol{M}$ we have: $\boldsymbol{M}' = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$ where σ_1 is the largest singular value, and $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the corresponding singular vectors.

From here we can conclude that $\boldsymbol{X}' = \sigma_1 \boldsymbol{u}_1 v_{11}$ and $\boldsymbol{y}' = \sigma_1 \boldsymbol{u}_1 v_{12}$ where v_{11} and v_{12} are the first and second element of $\boldsymbol{v}_1$ respectively.Plugging $\boldsymbol{X}'$ and $\boldsymbol{y}'$ in we have:

$$\boldsymbol{w}^*\boldsymbol{X}' + b^* = \boldsymbol{y}'$$
$$\boldsymbol{w}^*\sigma_1 \boldsymbol{u}_1 v_{11} + b^* = \sigma_1 \boldsymbol{u}_1 v_{12}$$
$$\boldsymbol{w}^* v_{11} + b^* = v_{12}$$

From here we have: $b^* = 0$ and $\boldsymbol{w}^* = \frac{v_{12}}{v_{11}}$

c)

0 1

Since we assume that $\boldsymbol{X}'$ is full rank there are only two valid options: $K = D$ or $K = D + 1$. If $K = D$ then $\boldsymbol{y}'$ can be expressed as a liner combination of $\boldsymbol{X}'$ and we again achieve an error of 0. If $K = D + 1$ then the training error depends on the dataset and is in general ≥ 0.
Above, we made the simplifying assumption that $D \geq N$. However, the argument holds also for $D < N$ by substituting D with N.

Problem 9 (Version A) (6 credits)

0 1 2

a)

The objective for the (squared) Mahalanobis distance is $J(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \boldsymbol{z}_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$. By considering the optimization $\min_{\boldsymbol{Z}} J(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\mu})$ we can directly see the cluster assignment update from this:

$$\boldsymbol{z}_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Using the objective we can also derive the centroid update as

$$\frac{\partial J}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{i=1}^{N} \boldsymbol{z}_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = -\sum_{i=1}^{N} \boldsymbol{z}_{ik} 2 \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \tag{2}$$

$$\Leftrightarrow \quad \sum_{i=1}^{N} \boldsymbol{z}_{ik} \boldsymbol{\mu}_k = \sum_{i=1}^{N} \boldsymbol{z}_{ik} \boldsymbol{x}_i \quad \Leftrightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \boldsymbol{z}_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \boldsymbol{z}_{ik}} \tag{3}$$

Interestingly, the Mahanalobis distance does not have an influence on the centroid update.

0 1 2

b)

[Version A: This solution is much more thorough than necessary.]
Denote $\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$. The boundary between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is $\boldsymbol{x} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 + c(0, 1)^T$ for $c \geq 0$. For any boundary we have $d(\boldsymbol{x}, \boldsymbol{\mu}_1) = d(\boldsymbol{x}, \boldsymbol{\mu}_2)$. We thus have

$$(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) = (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \quad \Leftrightarrow \quad \begin{pmatrix} 1 & c \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} 1 \\ c \end{pmatrix} = \begin{pmatrix} -1 & c \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} -1 \\ c \end{pmatrix} \tag{4}$$

$$\Leftrightarrow \quad \sigma_{11} + 2c\sigma_{12} + c^2\sigma_{22} = \sigma_{11} - 2c\sigma_{12} + c^2\sigma_{22}$$

and therefore $\sigma_{12} = 0$. The boundary between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_3$ is $\boldsymbol{x} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3)/2 + c(1, 1)^T$ for a certain range of c. Considering $\sigma_{12} = 0$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3 = (1, -1)^T$ we have

$$(c + 0.5)^2 \sigma_{11} + (c - 0.5)^2 \sigma_{22} = (c - 0.5)^2 \sigma_{11} + (c + 0.5)^2 \sigma_{22} \tag{5}$$

and thus $\sigma_{11} = \sigma_{22}$. Since Σ is PSD and invertible, $\boldsymbol{\Sigma}^{-1}$ must be PD. We therefore have (for any $a > 0$)

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}. \tag{6}$$

c)

[Version A. This solution is much more thorough than necessary.]
Since there is a vertical/horizontal boundary in the center we have $\sigma_{12} = 0$ (see previous subproblem). The boundary between $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ is $\boldsymbol{x} = (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3)/2 + c(2,1)^T$ for a certain range of c. With $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = (1,-1)^T$ we therefore have

$$\begin{aligned} &\begin{pmatrix} 2c+0.5 & c-0.5 \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} 2c+0.5 \\ c-0.5 \end{pmatrix} = \begin{pmatrix} 2c-0.5 & c+0.5 \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} 2c-0.5 \\ c+0.5 \end{pmatrix} \\ \Leftrightarrow \quad &(4c^2 + 2c + 0.25)\sigma_{11} + (c^2 - c + 0.25)\sigma_{22} = (4c^2 - 2c + 0.25)\sigma_{11} + (c^2 + c + 0.25)\sigma_{22}. \end{aligned} \tag{7}$$

Considering only terms with c^1 we have $2\sigma_{11} - \sigma_{22} = -2\sigma_{11} + \sigma_{22} \;\Leftrightarrow\; 4\sigma_{11} = 2\sigma_{22}$.
Since a covariance matrix is PSD and $\boldsymbol{\Sigma}$ is invertible, $\boldsymbol{\Sigma}^{-1}$ must be positive definite. the solution for each version is (for any $a > 0$)

$$\boldsymbol{\Sigma}_A^{-1} = \begin{pmatrix} a & 0 \\ 0 & 2a \end{pmatrix}, \quad \boldsymbol{\Sigma}_B^{-1} = \begin{pmatrix} 2a & 0 \\ 0 & a \end{pmatrix}, \quad \boldsymbol{\Sigma}_C^{-1} = \begin{pmatrix} a & 0 \\ 0 & 2a \end{pmatrix}, \quad \boldsymbol{\Sigma}_D^{-1} = \begin{pmatrix} 2a & 0 \\ 0 & a \end{pmatrix}. \tag{8}$$

Problem 10 (Version A) (6 credits)

0 1 2 3

a)

Changing any single instance only modifies one of the groups G_i so it is sufficient to reason *only* about the sensitivity of the aggregation function operating on the groups.

Since f is bounded, the aggregation function takes as input m numbers, $g_1, \ldots, g_m$ in the interval $[a, b]$. Changing one instance can change at most one g_i, and in the worst case the change can be anywhere in the interval $[a, b]$.

In the worst-case the output of one g_i changes from b to a, and the global Δ_1 sensitivity of f' is $\frac{b-a}{m}$.

0 1 2

b)

The global sensitivity of f' does not depend on n and therefore does not change.

The global sensitivity of f' decreases as we increase m since we are dividing by m.

0 1

c)

We can obtain ϵ-DP by adding noise from the Laplace distribution with zero mean and variance $\frac{\Delta_1}{\epsilon}$ where $\Delta_1 = \frac{b-a}{m}$.

Problem 10 (Version B) (6 credits)

a)

Changing any single instance only modifies one of the groups G_i so it is sufficient to reason *only* about the sensitivity of the aggregation function operating on the groups.

Since f is bounded, the aggregation function takes as input m numbers, $g_1, \dots, g_m$ in the interval $[a, b]$. Changing one instance can change at most one g_i, and in the worst case the change can be anywhere in the interval $[a, b]$.

In the worst-case we have the following scenario:
Before changing a single instance: $g_1 = a, g_2 = a, \dots, g_{m/2} = a, g_{m/2+1} = b, \dots, g_{m-1} = b, g_m = b$
After changing a single instance: $g_1 = a, g_2 = a, \dots, g_{m/2} = b, g_{m/2+1} = b, \dots, g_{m-1} = b, g_m = b$

Here the median is $g_{m/2}$ and it has changed from a to b. Therefore, the global Δ_1 sensitivity of f' is $b - a$.

b)

The global sensitivity of f' does not depend on n and therefore does not change.

The global sensitivity of f' does not depend on m and therefore does not change.

c)

We can obtain ϵ-DP by adding noise from the Laplace distribution with zero mean and variance $\frac{\Delta_1}{\epsilon}$ where $\Delta_1 = b - a$.

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