

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
Examiner: Prof. Dr. Stephan Günnemann

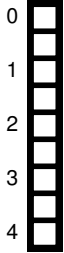
Date: Tuesday 16th February, 2021
Time: 11:00 – 13:00

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)



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

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

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

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

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.

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

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.

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

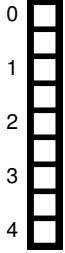
Summing the observations, we get $T = 26$ and so the positive most likely severity of the disease is $s^* = \sqrt{\frac{3}{26+1}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$.

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

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

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)



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

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

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

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

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)

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

0
1
2

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.

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 $(5, 1)$. The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

0
1

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 \blacksquare but is now labeled with \blacktriangle 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)

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 $(1, 3)$ 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 $(5, 1)$. The distance is $\sqrt{1^2 + 4^2} = \sqrt{17}$.

Problem 3 (Version A) (6 credits)

a)

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

$$\begin{aligned}\frac{\partial}{\partial w_{D+1}} J(\mathbf{w}) &= \sum_{i=1}^N (\mathbf{w}^T \tilde{\mathbf{x}}^{(i)} - y^{(i)}) + \lambda w_{D+1} \\ &= \mathbf{w}_{1:D}^T \sum_{i=1}^N \mathbf{x}^{(i)} + w_{D+1} \sum_{i=1}^N 1 - \sum_{i=1}^N y^{(i)} + \lambda w_{D+1}\end{aligned}$$

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

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

Solving for w_{D+1} we get

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

b)

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

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

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

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

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

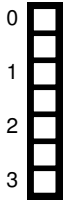
$$\mathcal{L}(\tilde{\mathbf{w}}) = \frac{1}{2} \sum_{i=1}^N \left(\tilde{\mathbf{w}}_{1:D}^T \mathbf{x}^{(i)} + \left(\frac{1}{N + \lambda} \sum_{j=1}^N y^{(j)} \right) - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\tilde{\mathbf{w}}_{1:D}\|_2^2 + \frac{\lambda}{2} \left(\frac{1}{N + \lambda} \sum_{j=1}^N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

The last part has zero gradient with respect to $\tilde{\mathbf{w}}_{1:D}$, so it does not influence the optimal $\tilde{\mathbf{w}}_{1:D}^*$ and we can drop it since $\tilde{\mathbf{w}}_{D+1}$ 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 $\hat{\mathcal{D}}$

$$\mathcal{L}(\hat{\mathbf{w}}) = \frac{1}{2} \sum_{i=1}^N (\hat{\mathbf{w}}^T \mathbf{x}^{(i)} - \hat{y}^{(i)})^2 + \frac{\lambda}{2} \|\hat{\mathbf{w}}\|_2^2.$$

showing that ridge regression on $\hat{\mathcal{D}}$ is equivalent to ridge regression on $\tilde{\mathcal{D}}$.

Problem 3 (Version B) (6 credits)



a)

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

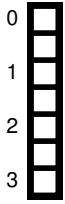
$$\begin{aligned}\frac{\partial}{\partial w_{D+1}} J(\mathbf{w}) &= \sum_{i=1}^N (\mathbf{w}^T \tilde{\mathbf{x}}^{(i)} - y^{(i)}) + \lambda w_{D+1} \\ &= \mathbf{w}_{1:D}^T \sum_{i=1}^N \mathbf{x}^{(i)} + w_{D+1} \sum_{i=1}^N 1 - \sum_{i=1}^N y^{(i)} + \lambda w_{D+1}\end{aligned}$$

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

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

Solving for w_{D+1} we get

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



b)

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

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

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

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

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

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

The last part has zero gradient with respect to $\tilde{\mathbf{w}}_{1:D}$, so it does not influence the optimal $\tilde{\mathbf{w}}_{1:D}^*$ and we can drop it since \tilde{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 $\hat{\mathcal{D}}$

$$\mathcal{L}(\hat{\mathbf{w}}) = \frac{1}{2} \sum_{i=1}^N (\hat{\mathbf{w}}^T \mathbf{x}^{(i)} - \hat{y}^{(i)})^2 + \frac{\lambda}{2} \|\hat{\mathbf{w}}\|_2^2.$$

showing that ridge regression on $\hat{\mathcal{D}}$ is equivalent to ridge regression on $\tilde{\mathcal{D}}$.

Problem 4 (Version A) (6 credits)

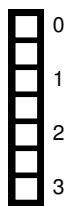
a)

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, $x_1 | y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 | y = c \sim \text{Bernoulli}(\alpha_c)$ 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 $p(y^{(b)} | \mathbf{x}^{(b)}) \propto p(\mathbf{x}_1^{(b)} | y^{(b)}) p(\mathbf{x}_2^{(b)} | y^{(b)}) p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

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



c)

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

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



d)

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

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



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, $x_1 | y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 | y = c \sim \text{Bernoulli}(\alpha_c)$ 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

$$\pi = \left(\frac{2}{7} \quad \frac{3}{7} \quad \frac{2}{7} \right)^T$$

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

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

0 ☐

1 ☐

b)

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

$$p(y^{(b)} | \mathbf{x}^{(b)}) \propto \left(e^{-\frac{1}{2} \frac{1}{2} \frac{2}{7}} \quad e^{-\frac{2}{3} \frac{1}{3} \frac{3}{7}} \quad e^{-\frac{9}{2} 1 \frac{2}{7}} \right)^T = \left(\frac{1}{7\sqrt{e}} \quad \frac{1}{7e^2} \quad \frac{2}{7e^{\frac{9}{2}}} \right)^T$$

0 ☐

1 ☐

c)

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

$$p(y^{(c)} | \mathbf{x}^{(c)}) = p(y) = \left(\frac{2}{7} \quad \frac{3}{7} \quad \frac{2}{7} \right)^T$$

0 ☐

1 ☐

d)

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

$$p(y^{(d)} | \mathbf{x}^{(d)}) = \left(\frac{1}{2} \frac{2}{7} \quad \frac{2}{3} \frac{3}{7} \quad 0 \frac{2}{7} \right)^T = \left(\frac{1}{7} \quad \frac{2}{7} \quad 0 \right)^T$$

Problem 4 (Version C) (6 credits)

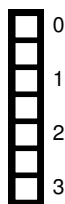
a)

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, $x_1 | y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 | y = c \sim \text{Bernoulli}(\alpha_c)$ 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 $p(y^{(b)} | \mathbf{x}^{(b)}) \propto p(\mathbf{x}_1^{(b)} | y^{(b)}) p(\mathbf{x}_2^{(b)} | y^{(b)}) p(y^{(b)})$, so we evaluate that for all three choices of $y^{(b)}$ and get

$$p(y^{(b)} | \mathbf{x}^{(b)}) \propto \left(e^{-\frac{9}{2}} \frac{1}{2} \frac{2}{7} \quad e^{-\frac{1}{2}} 0 \frac{2}{7} \quad e^{-\frac{9}{2}} \frac{2}{3} \frac{3}{7} \right)^T = \begin{pmatrix} \frac{1}{7e^{\frac{9}{2}}} & 0 & \frac{2}{7e^{\frac{9}{2}}} \end{pmatrix}^T$$



c)

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

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



d)

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

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



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, $x_1 | y = c \sim \mathcal{N}(\mu_c, 1)$. The discrete feature can be one of two values which we model with a Bernoulli distribution $x_3 | y = c \sim \text{Bernoulli}(\alpha_c)$ 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

$$\pi = \left(\frac{2}{7} \quad \frac{2}{7} \quad \frac{3}{7} \right)^T$$

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

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

0

1

b)

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

$$p(y^{(b)} | \mathbf{x}^{(b)}) \propto \left(e^{-8 \frac{1}{2} \frac{2}{7}} \quad e^0 0 \frac{2}{7} \quad e^{-2 \frac{2}{3} \frac{3}{7}} \right)^T = \left(\frac{1}{7e^8} \quad 0 \quad \frac{2}{7e^2} \right)^T$$

0

1

c)

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

$$p(y^{(c)} | \mathbf{x}^{(c)}) = p(y) = \left(\frac{2}{7} \quad \frac{2}{7} \quad \frac{3}{7} \right)^T$$

0

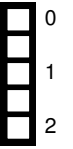
1

d)

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

$$p(y^{(d)} | \mathbf{x}^{(d)}) = \left(\frac{1}{2} \frac{2}{7} \quad 1 \frac{2}{7} \quad \frac{1}{3} \frac{3}{7} \right)^T = \left(\frac{1}{7} \quad \frac{2}{7} \quad \frac{1}{7} \right)^T$$

Problem 5 (Version A) (2 credits)



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

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

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

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

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

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

Sample Solution

Problem 6 (Version A) (3 credits)

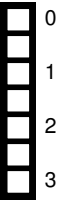
0	<input type="checkbox"/>
1	<input type="checkbox"/>
2	<input type="checkbox"/>
3	<input type="checkbox"/>

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.

Sample Solution

Problem 6 (Version B) (3 credits)



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.

Sample Solution

Problem 6 (Version C) (3 credits)

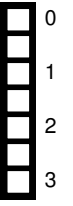
0	<input type="checkbox"/>
1	<input type="checkbox"/>
2	<input type="checkbox"/>
3	<input type="checkbox"/>

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.

Sample Solution

Problem 6 (Version D) (3 credits)

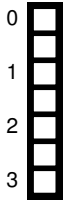


The output of conv1 will have shape [8, 64, 32]. Therefore,

- $C_{\text{in}} = 8$ 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.

Sample Solution

Problem 7 (All Versions) (5 credits)



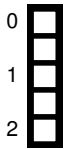
a)

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 \mathbf{w}_{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 $y_i(\mathbf{w}_{\text{soft}}^T \mathbf{x}_i + b) \geq 1$ since $\xi_i = 0$

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



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 \mathbf{w}_{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 $y_q(\mathbf{w}_{\text{soft}}^T \mathbf{x}_q + b) \geq 1 - \xi_q$. The optimal solution for

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

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

Substituting $\xi_q > 2$ we have $y_q(\mathbf{w}_{\text{soft}}^T \mathbf{x}_q + b) \geq -1$. 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 \mathbf{w}_{soft} with the corresponding margin $m_{\text{soft}} = \frac{2}{\|\mathbf{w}_{\text{soft}}\|}$, 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)

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

0
1
2

b)

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

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

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

$$\begin{aligned}\mathbf{w}^* \mathbf{X}' + b^* &= \mathbf{y}' \\ \mathbf{w}^* \sigma_1 \mathbf{u}_1 \mathbf{v}_{11} + b^* &= \sigma_1 \mathbf{u}_1 \mathbf{v}_{12} \\ \mathbf{w}^* \mathbf{v}_{11} + b^* &= \mathbf{v}_{12}\end{aligned}$$

From here we have: $b^* = 0$ and $\mathbf{w}^* = \frac{\mathbf{v}_{12}}{\mathbf{v}_{11}}$.

0
1
2
3

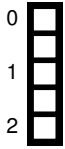
c)

Since we assume that \mathbf{X}' is full rank there are only two valid options: $K = D$ or $K = D + 1$. If $K = D$ then \mathbf{y}' can be expressed as a linear combination of \mathbf{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 .

0
1

Problem 9 (Version A) (6 credits)



a)

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

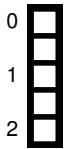
$$\mathbf{z}_{ik} = \begin{cases} 1 & \text{if } k = \arg \min_j (\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Using the objective we can also derive the centroid update as

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

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

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



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 $\mathbf{x} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 + c(0, 1)^T$ for $c \geq 0$. For any boundary we have $d(\mathbf{x}, \boldsymbol{\mu}_1) = d(\mathbf{x}, \boldsymbol{\mu}_2)$. We thus have

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

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

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

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

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

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

c)

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

[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 μ_2 and μ_3 is $\mathbf{x} = (\mu_2 + \mu_3)/2 + c(2, 1)^T$ for a certain range of c . With $\mu_2 - \mu_3 = (1, -1)^T$ we therefore have

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

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 Σ is invertible, Σ^{-1} must be positive definite. the solution for each version is (for any $a > 0$)

$$\Sigma_A^{-1} = \begin{pmatrix} a & 0 \\ 0 & 2a \end{pmatrix}, \quad \Sigma_B^{-1} = \begin{pmatrix} 2a & 0 \\ 0 & a \end{pmatrix}, \quad \Sigma_C^{-1} = \begin{pmatrix} a & 0 \\ 0 & 2a \end{pmatrix}, \quad \Sigma_D^{-1} = \begin{pmatrix} 2a & 0 \\ 0 & a \end{pmatrix}. \quad (8)$$

Problem 10 (Version A) (6 credits)

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

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

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

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.

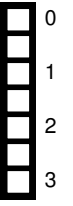
Since f is bounded, the aggregation function takes as input m numbers, g_1, \dots, g_m in the interval $[a, b]$. Changing one instance can change at most one g_i , and in the worst case the change can be anywhere in the interval $[a, b]$.

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.

A large grid of graph paper for solutions, with a diagonal watermark reading "Sample Solution". The grid is composed of small squares, and the watermark is written in a large, light blue font.

Sample Solution

Sample Solution