

Working instructions

- This exam consists of **4 pages** with a total of **9 problems**. Note that after we cover the remaining topics in the course the final and repeat exams will contain more exercises.
- For the mock exam you are allowed to use everything you need. However, during the final exam you are only allowed to use an
 - A4 sheet of handwritten notes (two sides) and
 - **no other materials (e.g. books, cell phones, calculators) will be allowed!**
- Mock exam will not be graded.
- A sample solution will be available on Piazza.
- **For problems that say "Justify your answer" or "Show your work" you only get points if you provide a valid explanation.** Otherwise it's sufficient to only provide the correct answer.

Problem 1 Probabilistic inference

Consider the following probabilistic model. The likelihood is defined as

$$\begin{aligned} p(x_i | \tau) &= \text{LogNormal}(x_i | 1, \tau) \\ &= \frac{\sqrt{\tau}}{x_i \sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(\log x_i - 1)^2\right), \end{aligned}$$

where $x_i \in (0, \infty)$ is a data point (i.e. observed sample) and $\tau \in (0, \infty)$ is the unknown precision parameter. We place the following prior on τ :

$$\begin{aligned} p(\tau | a, b) &= \text{Gamma}(\tau | a, b) \\ &= \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau). \end{aligned}$$

We have observed N samples $\mathcal{D} = \{x_1, \dots, x_N\}$ drawn iid from the above model.

What is posterior distribution of τ given \mathcal{D} ? Write down the expression for $p(\tau | \mathcal{D}, a, b)$. Make sure that the distribution is correctly normalized. Show your work.

Note: Your solution should NOT contain integrals but it may contain the Gamma function $\Gamma(\cdot)$.

Problem 2 Probabilistic inference & Convexity

Consider the following probabilistic model. The likelihood is defined as

$$p(x_i | \theta) = \theta \exp(x_i - \theta \exp(x_i)),$$

where $x_i \in \mathbb{R}$ is a data point (i.e. observed sample) and $\theta \in (0, \infty)$ is the unknown parameter. We place the following prior on θ

$$p(\theta) = 2\theta \exp(-\theta^2).$$

We have observed N samples $\mathcal{D} = \{x_1, \dots, x_N\}$ drawn iid from the above model.

Prove or disprove the following statement:

The log-posterior density $\log p(\theta | \mathcal{D})$ is a concave function of θ on the set $(0, \infty)$.

Problem 3 Decision Trees

Assume you want to build a decision tree. Your data set consists of N samples, each with k features ($k \leq N$).

a) If the features are binary, what is the maximum possible number of leaf nodes and the maximum depth of your decision tree?

 0

b) If the features are continuous, what is the maximum possible number of leaf nodes and the maximum depth of your decision tree?

 0

Problem 4 Regression


Doe is a data scientist, and he wants to fit a polynomial regression model to his data. For this, he needs to choose the degree of the polynomial that works best for his problem. Unfortunately, John hasn't attended IN2064, so he writes the following code for choosing the optimal degree of the polynomial:


```
X, y = load_data()
best_error = -1
best_degree = None

for degree in range(1, 50):
    w = fit_polynomial_regression(X, y, degree)
    y_predicted = predict_polynomial_regression(X, w, degree)
    error = compute_mean_squared_error(y, y_predicted)
    if (error <= best_error) or (best_error == -1):
        best_error = error
        best_degree = degree

print("Best degree is " + str(best_degree))
```

Assume that the functions are implemented correctly and do what their name suggests (for example `fit_polynomial_regression` returns the optimal coefficients `w` for a polynomial regression model with the given degree).

0  a) What is the output of this code (string printed in the last line)? Explain in 1-2 sentences why this code doesn't do what it's supposed to do.

0  b) Describe in 1-2 sentences a possible way to fix the problem with this code.

Note: you don't need to write any code, just describe the approach.

Problem 5 Classification


We would like to perform binary classification on multivariate binary data. That is, the data points $\mathbf{x}_i \in \{0, 1\}^D$ are binary vectors of length D , and each sample belongs to one of two classes $y_i \in \{1, 2\}$. Consider the following generative classification model. We place a categorical prior on y


$$p(y = 1 \mid \boldsymbol{\pi}) = \pi_1, \quad p(y = 2 \mid \boldsymbol{\pi}) = \pi_2.$$


The class-conditional distributions are products of independent Bernoulli distributions

$$p(\mathbf{x} \mid y = 1, \boldsymbol{\alpha}) = \prod_{j=1}^D \text{Ber}(x_j \mid \alpha_j),$$
$$p(\mathbf{x} \mid y = 2, \boldsymbol{\beta}) = \prod_{j=1}^D \text{Ber}(x_j \mid \beta_j),$$

where $\boldsymbol{\alpha} \in [0, 1]^D$ and $\boldsymbol{\beta} \in [0, 1]^D$ are the respective parameter vectors for both classes.

0  a) Write down the expression for the posterior distribution $p(y \mid \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

0  b) Assume that $D = 3$, $\boldsymbol{\alpha} = [\frac{1}{3}, \frac{1}{3}, \frac{3}{4}]$, $\boldsymbol{\beta} = [\frac{2}{3}, \frac{1}{2}, \frac{1}{2}]$, $\pi_1 = \frac{1}{3}$ and $\pi_2 = \frac{2}{3}$. Write down a data point $\mathbf{x}_1 \in \{0, 1\}^3$ that will be classified as class 1 by our model. Additionally, compute the posterior probability $p(y = 1 \mid \mathbf{x}_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

0  c) Consider the case when $D = 2$, $\pi_1 = \pi_2 = \frac{1}{2}$, and $\boldsymbol{\alpha} \in [0, 1]^2$ and $\boldsymbol{\beta} \in [0, 1]^2$ are known and fixed. Show that the resulting classification rule can be represented as a linear function of \mathbf{x} . That is, find $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$, such that

$$\{\mathbf{x} \in \{0, 1\}^2 : \mathbf{w}^T \mathbf{x} + b > 0\} = \{\mathbf{x} \in \{0, 1\}^2 : p(y = 1 \mid \mathbf{x}) > p(y = 2 \mid \mathbf{x})\}.$$

Problem 6 Convexity

For some $1 < N \in \mathbb{N}$ let S be the set of the so called stochastic matrices.

$$S = \{\mathbf{M} \in \mathbb{R}_{\geq 0}^{N \times N} : \sum_{i=1}^N M_{ij} = 1 \text{ for all } j = 1, \dots, N \text{ and } \sum_{j=1}^N M_{ij} = 1 \text{ for all } i = 1, \dots, N\}$$

- a) Prove that S is convex.
- b) Prove that each perturbation matrix, that is a matrix $\mathbf{P} \in S$ with $P_{ij} \in \{0, 1\}$ for all $i, j = 1, \dots, N$, is a vertex of S .

Problem 7 Constrained optimization

We consider a linear regression model with the l_∞ -loss instead of using the standard l_2 -setting. We denote the given feature matrix as $\mathbf{X} \in \mathbb{R}^{N \times d}$, target values as $\mathbf{y} \in \mathbb{R}^N$ and parameter vector as $\mathbf{w} \in \mathbb{R}^d$. The corresponding task of fitting the model can be equivalently rewritten as follows using an auxiliary scalar variable v (where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^N$).

$$\begin{aligned} & \text{minimize}_{\mathbf{w}, v} && v \\ & \text{subject to} && \mathbf{y} - \mathbf{X}\mathbf{w} \leq \mathbf{1}v \end{aligned} \quad (1)$$

$$\mathbf{y} - \mathbf{X}\mathbf{w} \geq -\mathbf{1}v \quad (2)$$

- Write down the Lagrangian $L(\mathbf{w}, v, \boldsymbol{\alpha}, \boldsymbol{\beta})$ for this constrained optimization problem. Denote the Lagrange multipliers corresponding to constraints (1) and (2) as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ correspondingly.
- Obtain the Lagrange dual function $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ from the Lagrangian $L(\mathbf{w}, v, \boldsymbol{\alpha}, \boldsymbol{\beta})$.
- State the dual problem explicitly.
- What is the duality gap in this problem? Justify your answer.

Problem 8 Kernel

Prove or disprove that the function $k : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ defined as

$$k(\mathbf{x}, \mathbf{y}) = x_1 y_1 - x_2 y_2$$

is a valid kernel.

Problem 9 Deep learning

Suppose that we use the following activation function in a simple deep feedforward neural network (NN)

$$\sigma(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else.} \end{cases}$$

Describe in 1-2 sentences what problem will likely arise when training our NN using gradient descent.

[illegible]

