Probability distributions

For your reference, we provide the following probability density functions.

- Normal distribution

$$\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Gamma distribution

$$\text{Gamma}(x \mid \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x \in (0, \infty), \\ 0 & \text{else} \end{cases}$$

  where $\Gamma(\cdot)$ is the gamma function.

- Log-normal distribution

$$\text{Log-normal}(x \mid \mu, \tau) = \begin{cases} \frac{\sqrt{\tau}}{x\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(\log x - \mu)^2\right) & \text{if } x \in (0, \infty), \\ 0 & \text{else.} \end{cases}$$

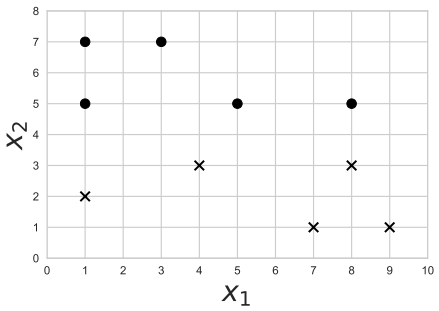
Probability Theory

Problem 1 [2 points] Given is the joint PDF of 3 continuous random variables $p(a, b, c)$. Write down how the following expressions can be obtained using the rules of probability.

(i) $p(a)$ (ii) $p(c \mid a, b)$ (iii) $p(b \mid c)$

K Nearest Neighbors

Problem 2 [3 points] Given is the data in the figure below, with binary labels marked as • and ×:



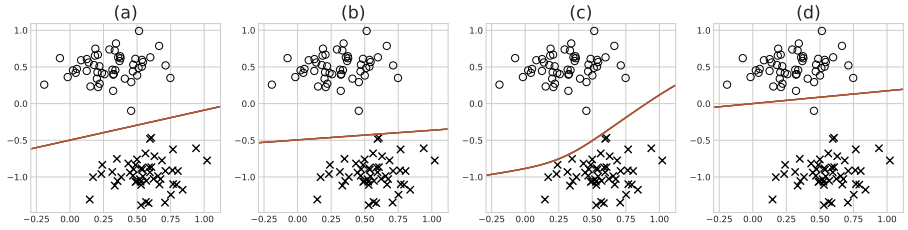
a) How many points get misclassified during the leave-one-out cross-validation (LOOCV) procedure when using a 1-NN classifier?

b) The original data can be represented as a matrix $\boldsymbol{X}_{orig} \in \mathbb{R}^{10\times2}$, where samples are stored as rows. We can apply a linear transformation to the data as $\boldsymbol{X}_{new} = \boldsymbol{X}_{orig} \cdot \boldsymbol{A}$, where $\boldsymbol{A} \in \mathbb{R}^{2\times2}$ is the transformation matrix.

   Find a linear transformation matrix $\boldsymbol{A} \in \mathbb{R}^{2\times2}$, such that all points get classified correctly during the LOOCV procedure using a 1-NN classifier. Provide a short explanation.

Classification

Problem 3 [3 points] We fitted 4 different classification algorithms on the same dataset and obtained 4 different decision boundaries.



Match each classifier to the respective decision boundary. Explain each of the answers with 1 sentence.

#	**Classification algorithm**	**Plot**
1.	Logistic regression	
2.	Hard margin linear SVM	
3.	Soft margin SVM with polynomial kernel	
4.	Perceptron	

Problem 4 [5 points] We have a binary classification problem with 1-dimensional data. The training samples from class 1 are $\{-1, 1, 3, 6, 1\}$ and the samples from class 2 are $\{5.5, 7, 9, 9, 9.5\}$.

We model class priors as a categorical distribution with parameters $\{\pi_1, \pi_2\}$. We model the class-conditionals as Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. Assume that the variances of class-conditionals are known to be $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$.

a) Find the maximum likelihood estimates (MLE) of the prior class probabilities $\{\pi_1, \pi_2\}$, as well as of the means $\{\mu_1, \mu_2\}$ of the class conditional densities.

b) What class will the point $x = 6$ be assigned to? Justify your answer. *Hint:* $\ln 1/\sqrt{2} \approx -0.35$

c) How many points $x \in \mathbb{R}$ lie on the decision boundary? That is, for how many points $x \in \mathbb{R}$ does it hold $p(y = 1 \mid x) = p(y = 2 \mid x)$? Provide a mathematical justification for your answer.

Dimensionality Reduction

Problem 5 [5 points] Outline the main steps of applying PCA to reduce the dimensionality of a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ from D to K.

Deep Learning

Problem 6 [4 points] Consider the following classification problem. There are two real-valued features x_1 and x_2, and a binary class label. The ground truth class labels are generated according to the following rule:

$$y = \begin{cases} 1 & \text{if } x_2 \geq |x_1|, \\ 0 & \text{else} \end{cases}$$

a) Can this function be perfectly represented by a feed-forward neural network with no hidden layers and 1 softmax output layer? Why or why not?

b) Design a two layer feed-forward network (that is, one hidden layer followed by an output layer, two weight matrices in total) that represents this function. You are allowed to use the hard thresholding activation function $\sigma(x)$ as the elementwise nonlinearity.

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

Specify the number of neurons and values of weights in each layer.

Optimization

You are given the following objective function

$$f(x_1, x_2) = 0.5x_1^2 + x_2^2 + 2x_1 + x_2 + \cos(\sin(\sqrt{\pi})).$$

Problem 7 [2 points] Compute the minimum (x_1^*, x_2^*) of $f(x_1, x_2)$ analytically.

Problem 8 [2 points] Perform 2 steps of gradient descent on $f(x_1, x_2)$ starting from the point $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$ with learning rate $\tau = 1$.

Problem 9 [2 points] Will the gradient descent procedure from Problem 8 ever converge to the true minimum (x_1^*, x_2^*)? Why or why not? If the answer is no, how can we fix it?

Problem 10 [2 points] Given two convex functions $g_1 : \mathbb{R} \to \mathbb{R}$ and $g_2 : \mathbb{R} \to \mathbb{R}$, prove or disprove that the function $h(x) = g_1(g_2(x))$ is also convex.

Regression

John 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.
(e.g., `fit_polynomial_regression` returns the optimal coefficients `w` for a polynomial regression model with the given degree.)

Problem 11 [3 points] What is the output of this code? Explain in 1-2 sentences why this code doesn't do what it's supposed to do.

Problem 12 [2 points] Describe in 1-2 sentences a possible way to fix the problem with this code. (You don't need to write any code, just describe the approach.)

SVM & Constrained optimization

Problem 13 [5 points] The goal is to minimize the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = 4x^2$ subject to $f_1(x) = -2x + 1 \leq 0$ using constrained optimization methods.

a) Write down the Lagrangian $L(x, \alpha)$ for this constrained optimization problem. Denote the necessary Lagrange multiplier as α.

b) Obtain the Lagrange dual function $g(\alpha)$ from the Lagrangian $L(x, \alpha)$.

c) State the dual problem explicitly and solve it to obtain the value for the Lagrange multiplier α.

d) What is the duality gap in this problem? Justify your answer.

e) Obtain the solution to the original problem from the solution of the dual problem.

Problem 14 [3 points] Prove or disprove that the function $k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is a valid kernel.

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

Variational inference

Problem 15 [3 points] You are given a dataset consisting of N positive samples $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$, $x_i \in \mathbb{R}^+$. You model the data-generating distribution (i.e., the likelihood) using a log-normal distribution, that is

$$p(x_i \mid \mu, \tau) = \text{Log-normal}(x_i \mid \mu, \tau),$$

where μ is the *known* and *fixed* mean parameter, and τ is the *unknown* precision parameter. You choose a Gamma distribution as the prior for τ, that is $p(\tau \mid a, b) = \text{Gamma}(\tau \mid a, b)$.

You would like to approximate the posterior $p(\tau \mid \boldsymbol{x}, \mu, a, b)$ using variational inference. Which of the following families of variational distributions $q(\tau)$ will yield the best approximation (in terms of KL-divergence)?

a) $q(\tau) = \text{Log-normal}(\tau \mid \nu, \beta) = \frac{\sqrt{\beta}}{\tau\sqrt{2\pi}} \exp\left(-\frac{\beta}{2}(\ln \tau - \nu)^2\right)$

b) $q(\tau) = \text{Normal}(\tau \mid \nu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\tau - \nu)^2\right)$

c) $q(\tau) = \text{Gamma}(\tau \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau)$

d) $q(\tau) = \text{Inverse-gamma}(\tau \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{-\alpha-1} \exp\left(-\frac{\beta}{\tau}\right)$

Show your work! Just stating a), b), c) or d) is not enough!