Elements of Machine Learning, WS 2025/2026
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #2: *Classification*

Deadline: Wednesday, November 26, 2025, 23:59

This problem set is worth a total of 50 points, consisting of 3 theory questions and 1 programming question. Please carefully follow the instructions below to ensure a valid submission:

- You are encouraged to work in groups of two students. Register your team (of 1 or 2 members) on the CMS. You have to register your team for each assignment.
- All solutions, including coding answers, must be uploaded individually to the CMS under the corresponding assignment and problem number. On CMS you will find FOUR problems under each assignment. Make sure you upload correctly each of your solution against $Assignment\#X - Problem\ Y$ (where X- Assignment number and Y is the problem number) on CMS. In total you have to upload FOUR PDFs.
- For each **theoretical question**, we encourage using LaTeX or Word to write your solutions for clarity and readability. Scanned handwritten solutions will be accepted as long as they are clean and easily legible. Final submission format must always be in a single PDF file per theoretical problem. Ensure your name, team member's name (if applicable), and matriculation numbers are clearly listed at the top of each PDF.
- For **programming question**, you need to upload a PDF/HTML file to CMS under $Assignment\#X - Problem$ 4. For creating PDF/HTML, use the export of the Jupyter notebook. Before exporting, ensure that all cells have been computed. To do this:
  - Go to the "Cell" menu at the top of the Jupyter interface.
  - Select "Run All" to execute every cell in your notebook.
  - Once all cells are executed, export the notebook: Click on "File" in the top menu.
  - Choose "Export As" and select either PDF or HTML.

  The submission should include your name, team member's name, and matriculation numbers at the top of PDF/HTML document.
- Finally, ensure academic integrity is maintained. Cite any external resources you use for your assignment.
- No submission will be accepted over emails. Only submissions on CMS will be graded.
- If you have any questions follow the instructions here.

Problem 1 (Logistic Regression). **(20 Points)**
Consider a binary classification problem with the target variable $Y \in \{0, 1\}$ and the input variable $X \in \mathbb{R}^p$, which consists of p different features. Assume that we use logistic regression with a linear predictor $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ to model the target variable Y.

Hint: In the following, we use $\sigma(a) = \frac{e^a}{1+e^a} = \frac{1}{1+e^{-a}}$ to denote the *logistic function*.

(a) Explain the logistic regression model using $f(X)$, including the logistic function. To this end, write down the formula for the logistic function as it is used in logistic regression, explain its components and parameters, and describe the meaning of its output. (3 Points)

(b) We say that logistic regression is a *linear classifier* because $f(X)$ is a linear function of the input features. Does logistic regression also result in a *linear decision boundary*? State the conditions that points on the decision boundary must satisfy, and explain whether these conditions imply a linear boundary. (3 Points)

Hint: In the case of binary classification ($Y \in \{0, 1\}$), the decision boundary consists of the points $x \in \mathbb{R}^p$ where the model is equally likely to assign class 0 or class 1. We say that a decision boundary is linear if these points form a hyperplane. You can also base your argument on the two-dimensional case ($p = 2$), where a hyperplane is simply a line.

(c) Given a dataset of n independent observations $\{(x_i, y_i)\}_{i=1}^n$, where $y_i \in \{0, 1\}$, derive the expression for the (conditional) log-likelihood function of logistic regression. Your final result should be written in terms of the logistic function, i.e., $\sigma(f(X))$. Furthermore, explain how maximum likelihood estimation can be used to determine the parameters of the logistic regression model. (3 Points)

(d) The logistic regression cost function (i.e., the log loss function) is given by

$$\begin{aligned}\ell(\beta) &= \ell(\beta_0, \beta_1, \dots, \beta_p) \\ &= -\sum_{i=1}^{n} \left[y_i \log\left(p(y_i = 1 \mid x_i; \beta)\right) + (1 - y_i) \log\left(1 - p(y_i = 1 \mid x_i; \beta)\right)\right].\end{aligned} \tag{1.1}$$

(i) Explain how this loss function relates to the log-likelihood function from part (c). (1 Point)

(ii) Next, the goal is to find the partial derivatives of the logistic regression cost function with respect to the parameters $\beta_0, \beta_1, \dots, \beta_p$. To this end, first show that derivative of the logistic function is given by:

$$\frac{d}{dx}\sigma(x) = (1 - \sigma(x)) \cdot \sigma(x).$$ (1 Point)

(iii) Using the derivative of the logistic function from part (ii), show that

$$\frac{\partial}{\partial \beta_j} y_i \log\left(p(y_i = 1 \mid x_i; \beta)\right) = y_i \cdot \left(1 - \sigma\left(\beta^T x_i\right)\right) \cdot x_{ij}.$$ (1 Point)

Hint: In the formula, we assume that x_i is padded with a 1, i.e., $x_{i0} = 1$ such that $\beta^T x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.
Hint: Consider how $(p(y_i = 1 \mid x_i; \beta))$ is defined in logistic regression.

(iv) Using the derivative of the logistic function from part (ii), show that

$$\frac{\partial}{\partial \beta_j} (1 - y_i) \log\left(1 - p(y_i = 1 \mid x_i; \beta)\right) = -(1 - y_i) \cdot \sigma\left(\beta^T x_i\right) \cdot x_{ij}.$$ (1 Point)

(v) Using the previous results, show that

$$\frac{\partial}{\partial \beta_j} y_i \log\left(p(y_i = 1 \mid x_i; \beta)\right) + \frac{\partial}{\partial \beta_j} (1 - y_i) \log\left(1 - p(y_i = 1 \mid x_i; \beta)\right) = (y_i - \sigma(\beta^T x_i)) x_{ij}.$$

(1 Point)

(vi) Finally, write down the full derivative for $\frac{\partial}{\partial \beta_j}\ell(\beta) = \frac{\partial}{\partial \beta_j}\ell(\beta_0, \beta_1, \ldots, \beta_p)$. (1 Point)

(e) Now that you have a theoretical understanding of logistic regression, apply it to the following small dataset. The dataset consists of 6 data points, each with two input features (i.e., x_1 and x_2) and a binary output label y:

i	x_0	x_1	x_2	y	$p(y_i = 1 \mid x_i; \beta) = \sigma(f(x_i))$	$\hat{y}$
1	1.0	2.0	2.0	0		
2	1.0	2.0	−2.0	1		
3	1.0	5.0	0.0	0		
4	1.0	6.0	−1.0	0		
5	1.0	5.0	−2.0	0		
6	1.0	8.0	−2.0	1		

Given the logistic regression model with parameters $\beta_0 = -3$, $\beta_1 = 0.5$, and $\beta_2 = -0.5$:

(i) Calculate the predicted probability $p(y_i = 1 \mid x_i; \beta)$.

(ii) Using a threshold of 0.5, determine whether each data point is classified as 0 or 1 (i.e., compute $\hat{y}$).

(iii) Report the fraction of points that are classified correctly (i.e., compute the accuracy).

(iv) Calculate the logistic regression loss in Equation (1.1) for the given dataset and parameters.

(5 Points)

Problem 2 (Linear Discriminant Analysis). **(10 Points)**
You are given the scatter plot of a dataset with two classes, i.e., the positive class ($y = 2$) and the negative class ($y = 1$), in Figure 1. Your task is to apply 2-dimensional LDA ($p = 2$), where the discriminant function for a class $k \in \{1, 2\}$ is given by

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k). \tag{2.1}$$

(a) To apply LDA, we first have to estimate the unknown parameters (μ_k, Σ) from the data. Start by estimating the mean vector μ_k for each of the two classes (i.e., for $k = 1$ and $k = 2$). (1 Point)

(b) Estimate the *pooled covariance matrix* Σ of both classes. (2 Points)

Hint: We use the pooled covariance matrix of both classes since LDA assumes that all classes share the same covariance matrix. For $K \geq 2$ classes, the pooled covariance matrix Σ is defined as

$$\Sigma = \frac{\sum_{k=1}^{K} (n_k - 1) \cdot \Sigma_k}{\sum_{k=1}^{K} (n_k - 1)}, \tag{2.2}$$

where n_k is the number of samples in class k and Σ_k is the empirical covariance matrix for class k.

(c) Assuming equal priors for both classes (i.e., $\pi_1 = \pi_2 = 0.5$) and your estimates above, show how you would classify a new data point with coordinates $(4, 3)^T$ using the discriminant function in Equation (2.1). Clearly show all your steps. (3 Points)

(d) Draw the decision boundary for this problem in Figure 1. To this end, state the condition that points on the decision boundary must satisfy. Then derive the decision boundary and draw it. If you cannot draw it due to technical limitations, it is also fine to describe it briefly. (2 Points)

Hint: In the case of binary classification ($Y \in \{1, 2\}$), the decision boundary consists of the points $x \in \mathbb{R}^p$ where the model is equally likely to assign class 1 or class 2.

(e) Describe the key assumptions made by LDA and QDA about the distribution of the data for each class. How do these assumptions affect the resulting classification models? (2 Points)

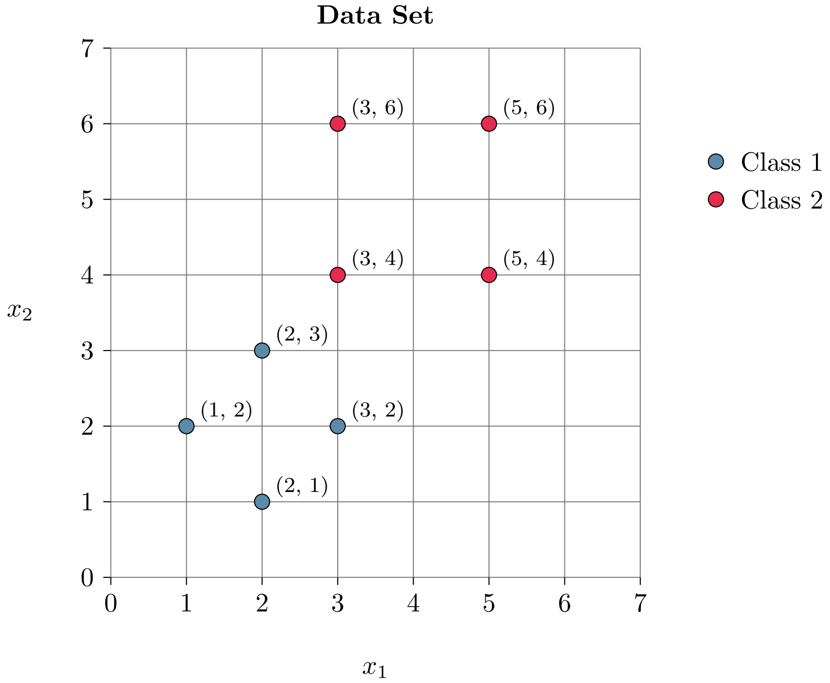

Figure 1: Training datapoints.

Problem 3 (Mixed Bag: Classification). **(10 Points)**

(a) What does it mean that the *Bayes classifier* is *optimal*? And if it is optimal, why do we rarely use it in practice? Explain in 2–3 sentences. (2 Points)

(b) Sam Notaltman has trained a binary classifier to predict whether an image contains copyrighted content (positive, $Y = 1$) or not (negative, $Y = 0$). On the test set, his classifier produces the following confusion matrix:

	Predicted Label	
True Label	Positive	Negative
Positive	426	$2,437$
Negative	$1,891$	$55,928$

(i) Calculate the sensitivity and specificity for the classifier on the test set. (1 Point)

(ii) Sam proudly proclaims that his model works nearly perfectly since it classifies a majority of the test data correctly. He now wants to deploy the model to prevent content theft. Do you think this is a good idea? Why or why not? Explain in 2–4 sentences. (1 Point)

(iii) Sam reveals that he used LDA as a classifier. He asks you whether QDA could make his test results even better. What is your answer, and why? (1 Point)

(c) You learned about the following classifiers: logistic regression, LDA, QDA, and k-NN. In the following, we always assume that there are only two classes ($Y \in \{0, 1\}$).

(i) Assume you have a very high-dimensional dataset with a relatively small sample size. Which method would you expect to work best and why? What are the problems with the other methods? (2 Points)

(ii) On your test set, LDA, QDA, and k-NN all give you approximately the same classification performance. Which model would you pick (to apply to new data), and why? (1 Point)

(iii) For highly imbalanced data, where one class has many more observations than the other one, in general, should you use a k-NN classifier? Why or why not? (1 Point)

(iv) Which method would you pick if you are unsure about the distribution of your samples? (1 Point)

Problem 4 (Coding: Classification). (**10 Points**)
In this assignment, you will work on classification. You will gain hands-on experience with logistic regression and explore the differences between LDA and QDA using an example dataset

Please refer to the file `assignment_2_handout.ipynb` and **only** complete the sections marked as tasks and the missing code denoted with `#TODO`. Once you have filled in the required parts, please revisit the submission instructions to learn how to submit your work.