Elements of Machine Learning, WS 2024/2025
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #2: *Classification*

Deadline: Wednesday, November 27, 2024, 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 at least ONE week before the submission deadline.
- 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 THREE PDFs (theoretical problems) and ONE ZIP file (programming problem).
- 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 ZIP file to CMS under $Assignment\#X - Problem$ 4. Each ZIP file must contain a PDF or HTML exported from Jupyter Notebook and the .ipynb file with solutions. Make sure all cells in your Jupyter notebook contain your final answers. 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 both PDF/HTML and .ipynb file document.
- Finally, ensure academic integrity is maintained. Cite any external resources you use for your assignment.
- If you have any questions follow the instructions here.

Problem 1 (Introduction to Logistic Regression). (10 Points)
Consider a binary classification problem where the target variable $Y \in \{0, 1\}$ and the input variable is a single feature $X \in \mathbb{R}$. Assume that we use a linear predictor $f(X)$ to model target variable Y.

(a) Explain the logistic regression model using $f(X)$, including the logistic function. Write down the form of the logistic function and interpret its output. (3 Points)

(b) 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 log-likelihood function for logistic regression. Then, explain how maximum likelihood estimation can be used to find the parameters of the logistic regression model. (4 Points)

(c) Compare and contrast generative and discriminative classifier models in terms of their output, loss functions, and optimization. (3 Points)

Problem 2 (Logistic Regression). (15 Points)

1. You learned about the logistic regression cost function (i.e the log loss function) which is given as:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log p(x_i; \beta) + (1 - y_i) \log \left(1 - p(x_i; \beta)\right) \right] \tag{2.1}$$

where $\beta = \{\beta_0, \beta_1, \beta_2\}$ and $x_i = \{1, x_{i1}, x_{i2}\}$ is a vector of the input values padded with a constant term

(a) Derive step by step the gradient of the logistic regression cost function (5 Points)

(b) Describe in your own words how the log loss cost function ensures that the model's predictions are as accurate as possible. (Not in more than one paragraph.) (5 Points)

(c) Now that you have a theoretical understanding of logistic regression, apply it to the following small dataset. The dataset consists of 8 data points with two input features (i.e x_1, x_2) and a binary output label: (5 Points)

i	x_0	x_1	x_2	y
1	1	1.0	2.0	0
2	1	2.0	3.0	0
3	1	3.0	4.0	0
4	1	4.0	5.0	1
5	1	5.0	6.0	1
6	1	6.0	7.0	1
7	1	7.0	8.0	1
8	1	8.0	9.0	1

i. Given the logistic regression model with parameters $\beta_0 = -3$, $\beta_1 = 0.5$ and $\beta_2 = 0.5$, calculate the predicted probability $p(x_i; \beta)$. (3 Points)

ii. Using a threshold of 0.5, determine whether each data point is classified as $y = 1$ or $y = 0$. (2 Points)

Problem 3 (Linear & Quadratic Discriminate Analysis). (15 Points)

1. You are given the scatter plot of datasets in Figure 1. Assume that we have two classes i.e. the positive and negative classes:

(a) Estimate the mean and covariance for each of the two classes (i.e., for $y = 0$ and $y = 1$). (2 Points)

(b) Assuming equal priors for both classes (i.e $\pi_1 = \pi_2 = 0.5$) and your estimates above, show how you will classify this new data point $(3.5, 2)$ using the discriminate function in Equation 3.1. Clearly show all your steps. (6 Points)

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{3.1}$$

(c) In your own words, describe the key assumptions made by LDA and QDA about the distribution of the data for each class. (2 Points)

(d) LDA can also be used as a dimensionality reduction technique in multiclass classification problems. Explain how LDA can reduce the dimensionality of a dataset while preserving class separability. (3 Points)

(e) When would you prefer to use LDA over QDA, and when would you choose QDA over LDA? Discuss in terms of sample size, number of features, and model complexity. (2 Points)

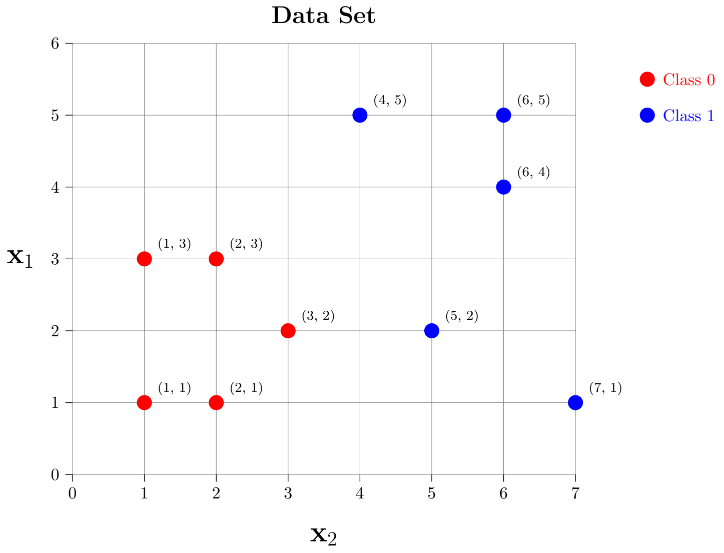

Figure 1: Training datapoints.

Problem 4 (Coding Classification). (10 Points)

In this assignment, you will work on Classification. You will gain hands-on experience with Logistic Regression and see differences of LDA and QDA on an example dataset.

Please refer to the file `assignment_2_handout.ipynb` and **only** complete the sections marked in red and missing codes denoted with `#TODO`. Once you have filled in the required parts, revisit submission instructions to check how to submit it.