Elements of Machine Learning, WS 2024/2025
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #1: *Introduction and Linear Regression*

UNIVERSITÄT DES SAARLANDES

Deadline: Wednesday, November 13, 2024, 23:59 h

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 (Parametric and Non-parametric models). (10 Points)

Briefly describe parametric and non-parametric models (4 Points) and compare and contrast them in terms of their:

1. Complexity and flexiblity. (2 Points)
2. Assumptions about target data distribution. (2 Points)
3. Generalization ability with small and big datasets. (2 Points)

Problem 2 (Bias Variance Trade-off). (15 Points)

You are given D datasets of n data points (x_i, y_i) for $i = 1, \ldots, n$, where your goal is to estimate the true underlying function f that governs the relationship between house features X and house prices Y. Different linear models are used to estimate $\hat{f}$, and your task is to analyze the bias-variance tradeoff in each scenario (i.e. low or high).

1. Scenario 1 (Single Predictor, Medium Dataset) (5 Points)
   You use only house size feature to predict house prices with $n = 500$
   $$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{house\_size}$$
2. Scenario 2 (All Available Predictors, Small Dataset) (5 Points)

   You use all available predictors (including age of the seller) with small datasets with $n = 50$

   $$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{house\_size} + \hat{\beta}_2 \cdot \text{rooms} + \hat{\beta}_3 \cdot \text{neighborhood} + \hat{\beta}_4 \cdot \text{age\_of\_house} + \hat{\beta}_5 \cdot \text{age\_of\_seller}$$
3. Scenario 3 (Subset of Important Predictors with Interactions, Large Dataset) (5 Points)
   You use a subset of important predictors including the interaction terms with a large dataset $n = 5000$
   $$\begin{aligned}\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{house\_size} + \hat{\beta}_2 \cdot \text{rooms} + \hat{\beta}_3 \cdot \text{neighborhood} \\ + \hat{\beta}_4 \cdot \text{age\_of\_house} + \hat{\beta}_5 \cdot (\text{house\_size} \cdot \text{neighborhood})\end{aligned}$$

Problem 3 (Linear Regression). (15 Points)

Assume that we have the data generation process

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ is the vector of model coefficients, including the intercept β_0. $\boldsymbol{X} \in \mathbb{R}^{n \times (k+1)}$ is the design matrix, where each row corresponds to a data point and each column is a feature (including the intercept column of 1s). $\boldsymbol{Y} \in \mathbb{R}^n$ is the vector of observed values (responses) for each data point. $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the vector of error terms. It is assumed that $E(\boldsymbol{\epsilon}) = 0$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, meaning the errors have zero mean and constant variance.

We have the following dataset
We use the following linear estimator to model true data generator

(i)	x_{i1}	x_{i2}	y_i
1	1	2	5
2	2	1	6
3	3	3	9
4	4	2	10
5	5	3	13

Table 1: Sample data

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

The difference between the observed and predicted values is the residual, denoted as:

$$e_i = y_i - \hat{y}_i$$

1. Derive Residual Sum of Squares (RSS) is the sum of squared residuals for all data points. Make sure to customize it to our model. (3 Points)
2. Derive and compute the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ that minimize the residual sum of squares by taking the partial derivatives of the RSS with respect to each coefficient. (7 Points)
3. Compute R-square value for our model. (5 Points)

Problem 4 (Coding Linear Regression). (10 Points)

In this assignment, you will implement a Linear Regression model. You will gain hands-on experience with metrics, model training, and visualization through a practical task.

Please refer to the file `assignment_1_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.