Elements of Machine Learning, WS 2025/2026
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #3: *Generalization and Regularization*

UNIVERSITÄT DES SAARLANDES

Deadline: Wednesday, December 10th, 2025 23:59 hrs

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 (Generalization). (**15 Points**)

1. Why should we do cross-validation on different splits of data (e.g., k-fold) instead of a fixed 50-50 split of training-validation set? Explain in at most three sentences. (3 Points)
2. Assume we want to evaluate a linear regression model (fitted with least-squares), and we decide to run LOOCV and k-fold CV on a dataset with 1000 data points:
   (a) If we have an implementation of k-fold CV, can we use it to perform LOOCV? Justify your answer in one or two sentences. (2 Points)
   (b) If we use different random seeds and run LOOCV and 5-fold CV several times, how will the estimation vary? Summarize in one sentence for each method. (2 Points)
3. Assume we are running the bootstrap on a dataset of n observations:
   (a) How many unique observations are expected in a bootstrap sample? Explain how you reached your answer. You may explain with plain language in no more than 3 sentences. (Use n in your answer without approximating with $n \to \infty$.) (3 Points)
   (b) After fitting a model on a bootstrap sample, what data should we use to estimate the model performance? (2 Points)
   (c) Does the error estimation from the bootstrap tend to be high or low? Justify your answer in one or two sentences. (3 Points)

Problem 2 (Regularization). (**15 Points**)

1. Consider dataset with n samples: $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = [y_i] \in \mathbb{R}^n$. We try to fit them with linear regression:
$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i,$$
where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are parameters and $\varepsilon_i, \ldots, \varepsilon_i$ are independent and identically distributed from $\mathcal{N}(0, \sigma^2)$.
   (a) Write down the likelihood of the dataset ($\mathcal{L}(\mathbf{y}|\mathbf{X}, \beta)$, likelihood of output given input and parameters). (2 Points)
   (b) Why does least-squares give the same estimation as maximum-likelihood here? Explain with formulae that show the connection. (2 Points)
   (c) How to trade off data likelihood and the number of parameters k ($= p + 1$) in model selection? Give an example of a selection criterion with this tradeoff and explain in at most two sentences how it shows the tradeoff. (4 Points)
2. Ridge and Lasso regressions are commonly used to regularize the parameters.
   (a) Write down the loss function of Ridge and Lasso (you may use the same notations as in the previous question, and use λ for the tuning parameter). (2 Points)
   (b) Discuss the trend of β when λ increases from 0 to $+\infty$. Use one or two sentences for each method. (2 Points)
   (c) Given high-dimensional data with many features that are colinear to each other, which method should we choose? And how to choose a lambda λ? Justify your answer in no more than 3 sentences. (3 Points)

Problem 3 (Calculation and Model Selection). **(10 Points)**

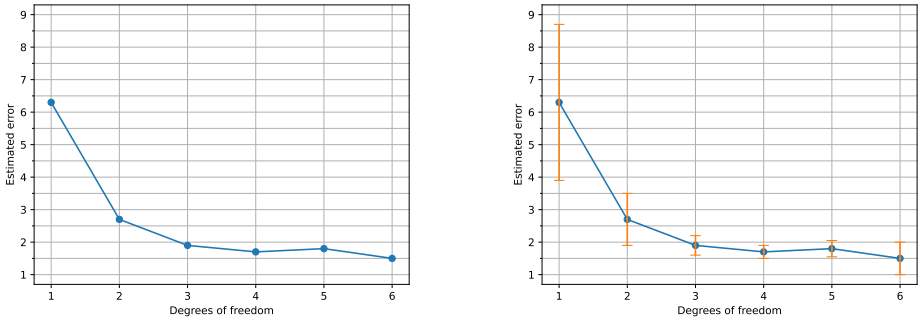
1. Assume you're trying to fit and evaluate a linear regression model $y = \beta_0 + \beta_1 X + \varepsilon$ on the following dataset. Please do 5-fold CV using RSS as metrics and report the final estimate ($CV_{(5)}$). (1 Point)
   Also, to show how you derived your answer, write down the data split (use X values to denote samples, e.g., train: 1,2,3; test: 4,5) and the corresponding RSS on the test set for each of the 5 folds. (5 Points)
   All numbers should be rounded to 2 digits after the decimal point.

X	y
1	2.665
2	6.556
3	9.189
4	12.481
5	13.786

2. Assume you are trying to select from several models with different degrees of freedom based on their performance on some dataset, and collected the estimations from a 5-fold CV process in Figure 1. Answer the following questions:

   (a) Based on just the estimated error plotted in Figure (a), which model is the best choice? Please name it by its dof.
   Are you confident with this selection? Please justify your answer in no more than two sentences. (2 Points)

   (b) Now you have plotted also the estimated standard deviation of the error in Figure (b). Will the new information change your selection? If yes, give your *new selection* and justify it in at most two sentences. If no, explain why the selection stays the same in at most two sentences. (2 Points)



Plot (a): Error estimated from 5-fold CV. Plot (b): Error estimation from 5-fold CV with standard deviation.

Figure 1: Estimated error vs. degrees of freedom.

Problem 4 (Coding Generalization, Regularization & Model Selection). (**10 Points**)

In this assignment, you will work on selecting the best model using k-fold cross-validation. You will also explore methods for regularization and selecting hyperparameters to enhance the generalizability of your trained models.

Please refer to the file `assignment_3_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.