Elements of Machine Learning, WS 2024/2025
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #3: *Generalization, Regularization and Beyond Linearity*

UNIVERSITÄT DES SAARLANDES

Deadline: Wednesday, December 11th, 2024 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 at least ONE week before the submission deadline. 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 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 (Generalization). **(10 Points)**

1. Assume you are only given training points for a binary classification problem and a small validation set. Does it make sense to compute the validation error for all classification methods (Logistic Regression, LDA, QDA) and report minimal validation error over all methods as an estimate of the test error ? Justify your answer. (3 Points)
2. Is it possible that model selection using cross-validation overfits? If yes, describe with an example, if no, explain the reason why overfitting is impossible. (4 Points)
3. Why does K-fold CV result in a higher bias than LOOCV? (3 Points)

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

1. Lasso and Ridge regressions are used to predict a target Y from X as shown in Equation 2.1 and 2.2 respectively. In order to understand which of the two models is better suited for a task, the mathematical equations for these are written as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.1}$$

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2.2}$$

   (a) Discuss how the model coefficients (β_j) change as $\lambda \to 0$ and as $\lambda \to \infty$ in both Equation 2.1 and 2.2. (4 Points)

   (b) If we have significantly more independent features than observations and want to perform feature selection, which type of regularization method should we use? (Hint: L_1 or L_2?) What value of λ should be considered i.e. small or large? (3 Points)

2. Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a $\mathcal{N}(0, \sigma^2)$ distribution.

   (a) Write out the likelihood for the data. (2 Points)

   (b) Assume the prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b, written as:

$$p(\beta) = \frac{1}{2b}\exp\left(-\frac{|\beta|}{b}\right).$$

   Write out the posterior for β in this setting. (2 Points)

   (c) Show that the lasso estimate is the mode for β under this posterior distribution. (4 Points)

Problem 3 (Beyond linearity: Polynomial and Splines). **(15 Points)**

1. Cubic regression spline with one knot at ξ can be obtained using a basis of the form $x, x^2, x^3, (x-\xi)_+^3$, where $(x-\xi)_+^3 = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise. We can show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3$$

   is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

(a) Find a cubic polynomial (2 Points)

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

(b) Find a cubic polynomial (2 Points)

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(x)$ is a piecewise polynomial.

(c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ. (2 Points)

(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ. (2 Points)

(e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ. (2 Points)
Therefore, $f(x)$ is indeed a cubic spline.
Hint: Parts (d) and (e) of this problem require knowledge of single-variable calculus. As a reminder, given a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3,$$

the first derivative takes the form

$$f_1'(x) = b_1 + 2c_1 x + 3d_1 x^2.$$

2. Consider two curves, $\hat{g}_1$ and $\hat{g}_2$, defined by (5 Points)

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 \, dx \right),$$

$$\hat{g}_2 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(4)}(x)]^2 \, dx \right),$$

where $g^{(m)}$ represents the m-th derivative of g.

(a) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training RSS?

(b) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller test RSS?

(c) For $\lambda = 0$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training and test RSS?

Problem 4 (Coding Generalization, Regularization and Beyond Linearity). **(10 Points)**

In this assignment, you will work on selecting the best model using K-fold cross-validation. You will also explore methods for 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.