LAOML 2025–2026 PROJECT 1

In this project you will construct artificial datasets to find out for which kind of data sets certain linear regression and classification methods do or do not work well.

Please hand in a Jupyter notebook with the Python or Julia code and brief explanations/discussions for each subquestion. For the code you should not use any machine learning libraries, but instead have your own functions that for instance use NumPy. Please make sure all the code in the notebook runs without us needing to modify anything. If you prefer to use some alternative notebook format or another open-source programming language, please discuss this with us. Use Brightspace to hand in the work. The hand-in date for this project is 10 November 2025. Please also prepare a short presentation (10min) where you discuss the experiments you performed and which results you found particularly interesting. The presentations/interviews will be on 12 and 13 November 2025.

You may use generative AI, but if you do, please make sure that you check (and discuss among each other) the AI generated material. Make sure that you fully agree with and understand everything you hand in; this will also be tested in the interview (10min) following the presentation. There will be an announcement on Brightspace to plan the time slot for the presentation and interview.

1. (a) Generate a 300×20 data matrix X, where each entry is uniformly random. Generate an outcome vector y, which is a linear combination of the columns of X with uniformly random weights, and some Gaussian noise added to each entry of y.

   (b) Write a function to divide the data set into a train and test set.

   (c) Write functions for OLS and Ridge regression and apply this to your synthetic data set. Discuss the performance on train and test sets.

   (d) Create a data matrix with many multicolinearities by adding a large number (say, 200) columns to X that are linear combinations of the original 20 columns with some Gaussian noise added to each entry. Run OLS and Ridge regression and discuss the performance on train and test sets. Is it hard to find a good value for λ?

   (e) Now instead of adding multicolinearities, add many superficial feature columns to X which have no relation to the outcome vector y. Again run OLS and Ridge regression and discuss the performance on train and test sets.

2. (a) Implement functions for logistic regression and hinge-loss classification.

   (b) Create a random data matrix X and construct an output vector y by generating and a random weight vector w and setting $y_i = \mathrm{sign}(x_i^\top w)$, where $x_i^\top$ is the i-th row of X. Use a test/train split and check the performance of OLS, Ridge regression, logistic regression and hinge-loss classification for binary classification. Do you see a large difference in performance between these methods?

   (c) Now create a data set (X, y) for binary classification (with $X \in \mathbb{R}^{n \times d}$ and $y \in \{-1, 1\}^n$) such that, given a test/train split, OLS and Ridge perform very badly but logistic regression and hinge-loss classification perform well. What kind of properties of your data set are responsible for this?