Homework 2

Statistical Machine Learning • STAT GR 5241 • Spring 2025

Assigned: February 24

Due: March 7

Instructions:

- *You may work with others on this homework assignment but all solutions must be written up and submitted individually.*
- *All homework assignments must be submitted in pdf format and should be at most* **6 pages** *in length. Any material beyond 6 pages will not be graded.*
- *You must submit all code used to complete this homework assignment as an appendix (this can go beyond the 6 pages). Failure to submit code will result in 20% reduction.*
- *You are permitted 4 total late days on homework assignments throughout the semester. Any late days taken beyond these 4 will incur a 20% reduction per day.*
- *This homework assignment has 80 total points available and* **any points earned above 70 will count as extra credit**.

1. (25 points) Conceptual. Suppose a team of researchers is investigating genetic, neurological, and environmental risk factors for autism using a massive database of medical cases from four hospital systems in the US (e.g. classification task to predict autism). To test their predictive models, the team has been asked to use a separate data set to determine how well they can predict which children have autism. This test data set however, consists of medical cases from two hospital systems that are in Europe. Based on other studies, the team knows that several key features such as ethnicity, demographic information, and environmental factors, are different in the European test population as compared to the American training population. Given that the training and testing sets are so different, how would you recommend the team set up their machine learning process? Which types of methods should the team use and why? How should they tune hyperparameters? Which error metrics should be used for reporting the prediction error? Please provide a detailed outline of your proposed procedure and justify all choices. (Assume that the team is not allowed to mix the training and test sets and that the key differences in features between the training and test sets are known.)

   Note: This is an open-ended question in which there are many possible correct approaches. Please outline and discuss as many possible solutions as you can. Extra credit points can be given for especially strong answers.

2. (25 points) Cross-Validation. For this problem, please use the Communities and Crime data set available at `https://archive.ics.uci.edu/dataset/183/communities+and+crime` that you used in Homework 1 (you should use your same preprocessed data).

   (a) (15 points) Code up K-fold cross-validation (from scratch) to select the kernel hyperparameter and the amount of regularization in kernel ridge regularization. Your function should work for both polynomial and RBF kernels.

   (b) (5 points) Use your cross-validation function to tune hyperparameters for the kernel and regularization amount for kernel ridge regression for both polynomial and RBF kernels. Compare the hyperparameters your function choose to those of built-in cross-validation functions in `sklearn`.

(c) (5 points) Compare the prediction errors for your final RBF and polynomial kernel ridge models to that of linear regression models from Homework 1. Which model is the best and do non-linear models offer improvements on this data set?

3. (30 points) Multi-class classification. For this problem, you will be using the MNIST handwritten digit dataset training and test sets that can be downloaded from `http://yann.lecun.com/exdb/mnist/`; you should limit your consideration to the digits 3, 5, and 8.

   - (20 points) Compare and contrast the following classification methods:
     - Logistic Regression: One-vs-Rest.
     - Multinomial Regression.
     - Naive Bayes.
     - Linear Discriminant Analysis.
     - Linear SVM: One-vs-Rest.

     Which method performs best in terms of test accuracy? Why? Provide a confusion matrix for classifying the three digits. Which digit is most often misclassified? Provide visualizations to help explain or interpret the results of various methods. Reflect on your results. (Note that you may use a validation set or cross-validation to tune all hyperparameters).
   - (10 points) Regularization. Apply group-lasso regularized multinomial logistic regression to select features that separate the three digits. Provide a visualization of the selected features and interpret your results.