STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 6

Instructions:

- Due date: Dec 4, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 35
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: Code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.
- Section 1 of the report must be limited to *five* pages. Also, only those software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

---

1. (12 points) Consider the `Hitters` data from the previous project. We would like a regression model with log(`Salary`) as response (due to skewness in `Salary`) and 19 predictor variables. All data will be taken as training data. For all the models below, use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

   (a) (2 points) Fit a tree to the data. Summarize the results. Unless the number of terminal nodes is large, display the tree graphically and explicitly describe the regions corresponding to the terminal nodes that provide a partition of the predictor space (i.e., provide expressions for the regions $R_1, \ldots, R_J$). Report its estimated test MSE.

   (b) (2 points) Use LOOCV to determine whether pruning is helpful and determine the optimal size for the pruned tree. Compare the best pruned and un-pruned trees. Report estimated test MSE for the best pruned tree. Which predictors seem to be the most important?

   (c) (2 points) Use a bagging approach to analyze the data with $B = 1000$. Compute the estimated test MSE. Which predictors seem to be the most important?

(d) (2 points) Use a random forest approach to analyze the data with $B = 1000$ and $m \approx p/3$. Compute the estimated test MSE. Which predictors seem to be the most important?

(e) (2 points) Use a boosting approach to analyze the data with $B = 1000$, $d = 1$, and $\lambda = 0.01$. Compute the estimated test MSE. Which predictors seem to be the most important?

(f) (2 points) Compare the results from the various methods. Which method would you recommend? How does your recommendation compare with the method you recommended in the previous project?

2. (12 points) Consider the `MNIST` dataset from `keras` package. It contains a training set of 60,000 28×28 grayscale images of 10 handwritten digits (from 0 to 9), along with a test set of 10,000 images. We would like to build a feedforward neural network model to identify the digit on the image. This is a multiclass classification problem with 10 output classes. For fitting the models below, use ReLU activation for the hidden layers, softmax activation for the output layer, and minibatches of size 128.

   (a) (1 point) Fit a neural network model with 1 hidden layer with 512 hidden units and 5 epochs. Report its training and test errors.

   (b) (1 point) Repeat (a) with 1 hidden layer with 512 hidden units and 10 epochs.

   (c) (1 point) Repeat (a) with 1 hidden layer with 256 hidden units and 5 epochs.

   (d) (1 point) Repeat (a) with 1 hidden layer with 256 hidden units and 10 epochs.

   (e) (1 point) Repeat (a) with 2 hidden layers, each with 512 hidden units, and 5 epochs.

   (f) (1 point) Repeat (a) with 2 hidden layers, each with 512 hidden units, and 10 epochs.

   (g) (1 point) Repeat (a) with 2 hidden layers, each with 256 hidden units, and 5 epochs.

   (h) (1 point) Repeat (a) with 2 hidden layers, each with 256 hidden units, and 10 epochs.

   (i) (1 point) Repeat (a) with L2 weight regularization with $\lambda = 0.001$.

   (j) (1 point) Repeat (a) with 50% dropout.

   (k) (2 points) Make a tabular summary of the results from all the above models and compare them. Which model would you recommend?

3. (11 points) Consider the `Boston Housing Price` dataset from `keras` package. It contains median price of homes in a Boston suburb in the mid-1970s, together with 13 numerical neighborhood characteristics. This relatively small dataset has 506 examples, split between a training set of size 404 and a test set of size 102. We would like to build a feedforward neural network model to predict the median home price based on the neighborhood features. Since the features are on different scales, they need to be standardized before fitting any model. Use the mean and standard deviation from the training data to standardize features in both training and test sets before doing any analysis. For fitting the models below, use ReLU activation for the hidden layers, no activation for the output layer, and minibatches of size 16. In addition, as described in the handout, use mean *absolute* error (MAE) computed using 4-fold CV as the performance accuracy measure.

   (a) (3 points) Fit a neural network model with 2 hidden layers, each with 64 hidden units, and 200 epochs. Make a plot of validation MAE against epoch. Would you recommend early stopping based on this plot? How many epochs would you suggest? Fit a model with the suggested number of epochs. Reports its validation MAE. Use this suggested number of epochs for all the models below.

   (b) (2 points) Fit a neural network model with 1 hidden layer with 128 units. Report its validation MAE.

   (c) (2 points) Add L2 weight regularization to the model with 2 hidden layers, each with 64 hidden units. Report its validation MAE.

(d) (2 points) Add L2 weight regularization to the model with 1 hidden layer with 128 hidden units. Report its validation MAE.

(e) (2 points) Compare the above models. Which model would you recommend? Compute MAE of the recommended model from the test data. Comment on the results.