DS5220 Supervised Machine Learning and Learning Theory: Homework 5

Due 2024/12/05 11:59pm

Instructions The homework must be submitted in **one pdf file**. Everyone needs to make a submission. Please write your nine-digit NUID and your northeastern.edu email. For the file names, please include your NUID. For example, "DS5220 HW5 (xxxxxxxxx)". Please specify the random seed at an appropriate place in your code chunk. It may not be the at the start, but specify it when you are executing codes that involve randomness. There are 100 points in total for this homework assignment.

Problem 1 (Classifying MNIST digits using principal component regression, 60 points) We will consider classifying the handwritten digits using principal component regression. Recall that principal component regression involves two steps. First, apply PCA to reduce the dimension of the dataset. Second, apply a regression model to the dimension-reduced dataset. To help you get started, we provided a handout notebook with instructions for loading the dataset into a numpy array.

(a) (20 points) Apply PCA to the training dataset with 20 principal components. Print the top 20 eigenvalues that correspond to the principal components. Also, print the explained variance ratios of the principal components.

(b) (20 points) Implement a principal component regression method by first applying PCA, then applying logistic regression to the dimension-reduced numpy matrix. Keep the number of principal components fixed at 20. Report the logistic regression model's training, validation, and test accuracy.

(c) (20 points) Select the number of principal components using the train-validation split in the handout. Report the number of principal components that achieve the highest validation accuracy. Then, report the training and test accuracy using this number of principal components in the principal component regression procedure. Comment on your findings.

Problem 2 [40 pts]

- [10 pts] How does the objective of causal inference differ from prediction or classification? Can you provide a real-life scenario to explain their differences?
- [10 pts] Explain the potential outcomes framework. In particular, explain the concepts of treatment groups and control groups in this framework.
- [10 pts] What is the difference-in-means estimator? And how does it relate to the average treatment effects?
- [10 pts] Explain the concept of selection bias, and based on that, describe the randomized assignment mechanism for running experiments. What is the selection bias of randomized experiments?