DS5220: Supervised Machine Learning and Learning Theory Fall 2022

Problem Set 2

Instructor: Hongyang Ryan Zhang *Due:* ***October 18, 2022, 11:59pm***

Instructions:

- You are expected to write up the solution on your own. Discussions and collaborations are encouraged; remember to mention any fellow students you discussed with when you turn in the solution.
- There are up to three late days for all the problem sets and project submissions. Use them wisely. After that, the grade depreciates by 20% for every extra day. Late submissions are considered case by case. Please reach out to the instructor if you cannot meet the deadline.
- Submit your written solutions to Gradescope and upload your code to Canvas. You are recommended to write up the solution in LaTeX.

Problem 1 (10 points)

(a) (2 points) State some use cases of Lasso regression and Ridge regression. Also comment when one is better than the other.

(b) (1 point) Explain in words the meaning of P-value and confidence interval.

(c) (1 point) Explain the idea behind maximum likelihood estimation for logistic regression.

(d) (2 points) Define and explain the difference between variance, covariance matrix, and correlation coefficient.

(e) (2 points) Explain the steps needed for performing the linear discriminant analysis. Try to state the steps in brief terms without using much notations.

(f) (2 points) Explain what is K-nearest neighbors. When is this approach better than alternative approaches (e.g., linear models)?

Problem 2 (30 points)

In this problem, you will develop a model to predict the classes of wine using the `load_wine` data set that is available in `sklearn.datasets`. The `load_wine` data set has 13 features and 3 classes. You can find the description of this data set at `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html`.[1] Note: Import the data set directly from sklearn library in python.

(a) (3 points) Import the `load_wine` data set from `sklearn.datasets` and convert it into a data frame. [Hint: Make sure to combine both data (`feature_names`) and classes (`target`) into pandas data frame]

(b) (6 points) Explore the data graphically in order to investigate the association between `target` and the other features. Which of the other features seem most likely to be useful in predicting `target`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings. [Hint: You may find `matplotlib.pyplot` helpful.]

(c) (3 points) Split the data into a training set and a test set with 80% observations randomly assigned to the training set and the rest 20% observations assigned to the test set.

(d) (6 points) Perform logistic regression on the training data in order to predict `target`. What is the test error of the model obtained? [Hint: You may find `LogisticRegression` from `sklearn.linear_model` and the functions `fit()` and `predict()` helpful.]

(e) (6 points) Perform LDA on the training data in order to predict `target`. What is the test error of the model obtained? [Hint: You may consider using the `LinearDiscriminantAnalysis` function from `sklearn.discriminant_analysis`.]

(f) (6 points) Perform KNN on the training data, with several values of K, in order to predict `target`. Report the test errors you observe. Which value of K performs the best for this data set? [Hint: You may find `sklearn.neighbors.KNeighborsClassifier` helpful.]

Problem 3 (20 points)

In this problem, we will consider the bootstrap sampling. We will derive the probability that a given data point is part of a bootstrap sampled set. Suppose that we obtain a bootstrap sampled set from a (training data) set of n observations: $x_1, x_2, \ldots, x_n$ and the set of z observations: $z_1, z_2, \ldots, z_n$.

(a) (4 points) Let z_1 be the first bootstrap sample. What is the probability that $z_1 \neq x_1$?

(b) (4 points) Let z_2 be the second bootstrap sample. What is the probability that $z_2 \neq x_1$? (Hint: The selection is independent)

[1] The data set can be downloaded here: `https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data`.

(c) (6 points) For any $n = 1, 2, \ldots$, let z_n be the n-th bootstrap sample. Let $S = \{z_1, z_2, \ldots, z_n\}$ be the set of bootstrap samples. When $n = 100$, what is the probability that x_1 is in S? (Hint: Use answer of part (b) to proceed)

(d) (6 points) For an arbitrary n, what is the probability that $x_1 \in S$? Based on this probability, what is the expected number of distinct data points in the set S? [Hint: Use answer of part (c) to proceed. Leave the answer in terms of n]

Problem 4 (40 points)

This question is based on the `Amazon's Best Selling Books` dataset from the years 2010 - 2020. This data set has the information about the title, rank, year, author, price and the ratings of the top 100 books for various years. You can remove `Year`, `Rank`, `Book_Title`, `Author`, and `Num_Customers_Rated`. [2]

(a) (2 points) Based on this data set, provide an estimate for the population mean of `Price` (the price of a book). Let's call this estimate $\hat{\mu}$. [Hint: You may find `numpy.mean()` helpful.]

(b) (3 points) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. [Hint: You can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations. You may find `numpy.std()` helpful.]

(c) (10 points) Now estimate the standard error of $\hat{\mu}$ using 1,000 bootstrap sampled sets. Let $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{1000}$ be the estimated mean from the 1,000 bootstrap sampled sets. Estimate the standard error of $\hat{\mu}$ using these 1,000 values. How does this compare to your answer from (b)? [Hint: You may find `sklearn.utils.resample` helpful. The standard error of $\hat{\mu}$ is the standard deviation of the 1,000 estimated means $\{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{1000}\}$ from all the bootstrap sampled sets.]

(d) (3 points) Based on your bootstrap estimate of the standard error from (c), provide a 95% confidence interval for the mean of `Price`. [Hint: You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2 \cdot \mathrm{se}(\hat{\mu}), \hat{\mu} + 2 \cdot \mathrm{se}(\hat{\mu})]$.]

Then, compare it to the results obtained using `scipy.stats.norm.interval()` (applied to `Price`).

(e) (2 points) Based on this data set, provide an estimate for the first 25% quantile of `Price`. Let's call this quantity $\hat{\mu}_{0.25}$ [You may find `numpy.quantile()` useful.]

[2] The data set can be downloaded from here: `https://www.kaggle.com/datasets/jiyoungkimpf/amazon-best-sellers-of-20102020-top-100-books`.

(f) (10 points) We would like to estimate the standard error of $\hat{\mu}_{0.25}$. While there is no simple formula to compute the standard error of $\hat{\mu}_{0.25}$, proceed by estimating the standard error of the first quartile using the bootstrap. Compare the standard error to the value of $\hat{\mu}_{0.25}$. Then, comment on your findings. [Hint: Follow similar steps to step (c).]

(g) (10 points) Consider a linear regression model to predict `Price` using `Rating` (Average Ratings by Users). Compute estimates for the standard errors of the intercept β_0 and coefficient β_1 of `Rating` in two different ways: (1) using the bootstrap, and (2) using the standard errors provided in the `scipy.stats.linregress()` function. Comment on your findings.