CS 6140 Machine Learning: Homework 2

Hongyang R. Zhang

Due 2023/10/13 11:59pm

Instructions:

- Each student should write their solution independently and submit a single PDF file that includes all the solutions on Gradescope. The running codes, such as jupyter notebooks or python files, can be submitted as supplementary materials on Canvas. Include the names of classmates you discussed with at the beginning of the solution.
- There are up to three late days for all the problem sets and project submissions. Use them wisely. Late submissions are considered case by case. Please reach out to us if you need extra late days.

Problem 1 (30 points) In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set. The `Auto` data set has gas mileage, horsepower, and other information for cars. You can find the description of this data set at `https://rdrr.io/cran/ISLR/man/Auto.html`. To load the data,

```
import pandas as pd
from ISLP import load_data
df = load_data('Auto')
df = df[df['horsepower'].notna()]
```

where the last line is to remove the observations with missing value in `horsepower`.

(a) (4 points) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. (Hint: You could compute the `median` using the `median()` function. You could add the `mpg01` column in `df`.)

(b) (4 points) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings. (Hint: You may find

`pd.plotting.scatter_matrix()`

helpful and you can set the argument `figsize` as $(10, 10)$)

(c) (4 points) Split the data into a training set and a test set with 80% observations in the training set and 20% observations in the test set. (Hint: You may find

```
from sklearn.model_selection import train_test_split
```

helpful)

(d) (5 points) Perform logistic regression on the training data in order to predict `mpg01` using `cylinders`, `weight`, `displacement`, and `horsepower`. What is the test error of the model obtained? (Hint: You may find

```
from sklearn.linear_model import LogisticRegression
```

and `predict_proba()` helpful)

(e) (5 points) Perform LDA on the training data in order to predict `mpg01` using `cylinders`, `weight`, `displacement`, and `horsepower`. What is the test error of the model obtained? (Hint: You may find

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
```

helpful)

(f) (4 points) Perform QDA on the training data in order to predict `mpg01` using `cylinders`, `weight`, `displacement`, and `horsepower`. What is the test error of the model obtained? (Hint: You may find

```
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
```

helpful)

(g) (4 points) Perform KNN on the training data, with several values of K, in order to predict `mpg01` using `cylinders`, `weight`, `displacement`, and `horsepower`. What test errors do you obtain? Which value of K seems to perform the best on this data set? (Hint: You may find

```
from sklearn.neighbors import KNeighborsClassifier
```

library helpful)

Problem 2 (Classifying MNIST digits using principal component regression, 35 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) (10 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) (15 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) (10 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 3 (35 points) This question is based on the `College` data set. This data set has statistics for a large number of US Colleges from the 1995 issue of US News and World Report. You can find the description of this data set at `https://rdrr.io/cran/ISLR/man/College.html`.[1] Let us first create a variable of acceptance rate, `Accept.Rate`, that is the number of applications accepted (`Accept`) divided by the number of applications received (`Apps`). We will now try to predict the acceptance rate using all variables other than `Accept` and `Apps`. We can remove `Accept` and `Apps` from the data frame.

(a) (0 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.

(b) (5 points) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) (7 points) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained. [Hint: You may find `sklearn.linear_model.Ridge` useful.]

(d) (8 points) Fit a lasso regression model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates. [Hint: You may find `sklearn.linear_model.Lasso()`.]

How does the result compare to c)?

(e) (5 points) Fit a partial least squares model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation. [Hint: You may find `sklearn.cross_decomposition.PLSRegression` useful.]

(f) (10 points) Perform best subset selection using forward stepwise selection. How does your answer compare to the results in (d)? Print the results you obtained and comment on your findings. [Hint: Write for loops to implement the forward stepwise rule.]

[1]The data set can be downloaded here: `https://www.kaggle.com/ishaanv/ISLR-Auto?select=College.csv`.