API 222 Problem Set 3

Machine Learning and Big Data Analytics: Spring 2024

Your Name

This problem set is worth 30 points in total. To get full credit, submit your code along with a write-up of your answers. This should either be done in R Markdown or Jupyter Notebook, submitted in one knitted PDF.

Conceptual Questions (11 pts)

1. Dimension Reduction (4 pts)

(a) What is the reason to scale (standardize) your variables when performing dimension reduction? (1 pt)

(b) You want to run a Principal Components Regression (PCR). Let's denote two of your features X1 and X2. The correlation between X1 and the outcome variable Y is 0 and the correlation between X2 and Y is 0.9. Which of the two variables will receive more weight in the the first factor loading of PCR regression? Briefly explain. (1 pt)

(c) You want to run a Partial Least Squares (PLS) regression. Let's denote two of your features X1 and X2. The correlation between X1 and the outcome variable Y is 0 and the correlation between X2 and Y is 0.9. Which of the two variables will receive more weight in the the first factor loading of PLS regression? Briefly explain. (1 pt)

(d) Is PCR a feature selection method? Why or why not? (1 pt)

2. Shrinkage (3 pts)

(a) What is the reason to scale (standardize) your variables when performing shrinkage methods? (1 pt)

(b) What are the trade-offs between using LASSO versus Ordinary Least Squares (OLS) in terms of bias and variance? (1 pt)

(c) What is the main advantage of LASSO over Ridge? (1 pt)

3. Tree-based Methods (4 pts)

(a) In 2-3 sentences, explain the idea of tree pruning and its importance in terms of the bias-variance trade-off. (1 pt)

(b) Under what conditions might we expect a tree-based classifier to outperform a linear classifier such as LDA?(1 pt)

(c) Briefly explain how Bagging can lead to both low variance and low bias compared to decision trees. (1 pt)

(d) Briefly explain how Random Forests results in lower variance compared to Bagging. (1 pt)

Data Questions (19 pts)

The DARWIN dataset contains handwriting data collected and is composed of 25 handwriting tasks. The protocol was specifically designed for the early detection of Alzheimer's disease (AD). The dataset includes data from 174 participants (89 AD patients and 85 healthy people).

The file "DARWIN.csv" contains the acquired data. The file consists of one row for each participant plus an additional header row. The first row is the header row, the next 89 rows collect patients data, whereas the remaining 84 rows collect information from healthy people.

The file consists of 452 columns. The first column shows participants' identifiers, whereas the last column shows the class to which each participant belongs. This value can be equal to 'P' (Patient) or 'H' (Healthy).

The remaining columns report the features extracted from a specific task. The tasks performed are 25, and for each task 18 features have been extracted. The column will be identified by the name of the features followed by a numeric identifier representing the task the feature is extracted. E.g., the column with the header "total_time8" collects the values for the "total time" feature extracted from task #8.

```r
# Load the data
# handwriting <- read.csv("DARWIN.csv")
```

1. Data preparation (3pts)

(a) Remove the columns that contain the participant's identifier and convert `class` to a factor type. (1 pt)

(b) Using the random seed 222, split the data (70/30) into a training set and a test set. (1 pt)

(c) How many observations are in the training data set? How many observations are in the test data set? (1 pt)

For the next set of questions, we will consider the prediction problem of classifying Alzheimer's disease patients from healthy people.

2. Decision tree (6 pts)

(a) Run a decision tree on your training data and use cross-validation to choose the optimal size. For your cross-validation use $k = 10$ and `prune.misclass` as the `FUN` argument. Hint: if you are not sure how to do this read the help file for `?cv.tree`. (1 pt)

```r
# Run a decision tree
library(tree)
```

(b) Show a plot of the deviance compared to the size. (2 pts)

(c) What is the optimal size? (1 pts)

(d) What is the misclassification error rate on the test set? Show this both for the original tree and for the pruned tree using the optimal size. (2 pts)

3. Random forest (5 pts)

(a) Run a random forest with 5000 trees using the training data. (1 pt)

```r
# Run a random forest
library(randomForest)
```

(b) What is your test misclassification error? (2 pts)

(c) What are the three most important variables in the random forest? (2 pts)

4. Boosting (4 pts)

(a) Recode the class variable so that it is 0 for healthy people and 1 for patients. (1 pt)

(b) Run a boosting model with 5000 trees and interaction depth of 4 using the training data. Hint: make sure you specify in your model that the distribution is "bernoulli". (1 pt)

```r
# Run a boosting model
library(gbm)
```

```
## Loaded gbm 2.1.9
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.co
```

(c) What is your test misclassification error? Hint: You will need to determine a threshold for which you decide a predicted value to be 1 (ALZ) or 0 (Healthy). Use 0 as this threshold. (2 pts)

5. Which model would you choose to predict Alzheimer's disease? Why? (1 pt)

6. Fun extra question (0 pt)

(a) Try running the threshold analysis for the boosting model to determine a better threshold than 0 to minimize misclassification error? (0 pts)