CME 250 Homework 4

Instructor: Sherrie Wang (sherwang@stanford.edu)

Available: Friday, February 15
Due: Friday, March 1, 5:00pm

Please submit Part 1 of the homework via Google Forms, and submit Parts 2 and 3 below via Gradescope.

For each question, clearly identify your final answer and show any work or code used to arrive at the answer. Answers without the corresponding code or justification will not receive credit. You may use any programming language of your choice.

Part 2. Applied Exercise

Question 1. Breast Cancer Dataset

Download `hw4_breast_cancer.csv` from the course website. This is the well-known Breast Cancer Wisconsin (Diagnostic) Dataset, which you can read more about `here`. The data should contain 569 samples, 30 predictive features (things like the radius, texture, concavity of the tumor) and the response variable `diagnosis`.

We will use PCA to reduce the 30 features to 2 dimensions and visualize the results.

(a) Standardize Features (2 points)

Notice that the features vary widely in their units and order of magnitudes (e.g. take a look at radius vs. area vs. concavity values). Before proceeding with PCA, we need to standardize the data. Subtract each column of the data by its mean and divide by its standard deviation. (However, for this exercise do not whiten the data. Whitening means removing correlations between features.)

(b) Principal Component Analysis (5 points)

Use PCA to project the data into 2 dimensions and plot each sample in this new feature space. That is, plot principal component #1 along the x-axis and principal component #2 along the y-axis. Use one color to plot benign tumors (`diagnosis == 'B'`) and another to plot malignant tumors (`diagnosis == 'M'`).

What does this plot tell you about the characteristics of benign vs. malignant breast cancers?

Part 3. Classification on Real Data

Question 2. Young People Survey

Download `hw4_yps.csv` from the course website, and use this version of the Young People Survey for the remainder of this problem. This version of the data contains the mean-imputed values we computed in Homework 3.

(a) Random Forest Classifier [4 points]

Using a random seed of `0` for reproducibility, split 20% of the data into a test set, leaving the rest for training and validation.

Let's fit some random forest classifiers to predict `Gender`. First, set the number of trees to 100. We will hold all other hyperparameters at their default value in whichever random forest implementation you are using, and focus on tuning the minimum samples at each leaf.

Using `GridSearchCV` in Python or analogous functions in R and Matlab as you did on Homework 3, perform 5-fold cross-validation to find the best value for minimum samples per leaf in the set $\{1, 2, 5, 10\}$. Here "best" is defined as highest accuracy on validation set `Gender` prediction.

What is the best minimum samples per leaf that you found?

(b) Feature Importance [3 points]

Using existing functions and properties, plot a bar graph to show the 10 most important features for predicting `Gender`. Plot feature names on the x-axis and their corresponding feature importance values on the y-axis. Please sort the 10 features in decreasing order of importance in the plot.

Do these top 10 features make sense?

(c) Generalization Error [1 point]

Using a random forest with optimal minimum samples per leaf found in part (a), report the prediction accuracy on the test set.