HW4: Resampling Methods

Programming assignment.

Directions.

• Comment the function to receive full credit.

• Provide a single Python file with the format name_391_hw4.py, where name is your full name.

• Send your file to nehemyl@uw.edu and zhangkh@uw.edu.

• Please specify your name in the heading of the email. This assignment is due on Thursday, May 4th at 11:59pm PST.

Write a function get_k_fold_ind that takes three input arguments: an integer \( n \) (total number of available observations), an integer \( k \) (number of folds), a one-dimensional Numpy array \( prop \) of length \( k - 1 \) (proportion of data in each fold), and returns \( k\_fold\_ind \) a list of Numpy arrays of length \( k \), which is a random partition of the indices \( 1, \ldots, n \). More precisely, \( k\_fold\_ind[0] \) contains a subset of the indices \( 1, \ldots, n \) of size \( \lfloor n \cdot prop[0] \rfloor \), and so on, \( k\_fold\_ind[k-2] \) contains a different subset of the indices \( 1, \ldots, n \) of size \( \lfloor n \cdot prop[k-2] \rfloor \), and \( k\_fold\_ind[k-1] \) contains the remaining indices. Here is an example of expected output.

```python
import numpy as np

n, k = 20, 5
prop = np.array([0.1, 0.3, 0.2, 0.1])

k_fold_ind = get_k_fold_ind(n, k, prop)

k_fold_ind
```

Your code should contain at most one for or while loop. You may use the Numpy functions random.shuffle and floor.
Directions. Show and explain all work to receive full credit. Homework is due on Thursday, May 4th at the beginning of class by 12:00 pm.

Problem 1. A data set consists of percentage returns for the S&P 500 stock index over 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. For each date, we have recorded the year that the observation was recorded, the percentage returns for each of the five previous trading weeks, \text{Lag1} through \text{Lag5}. We have also recorded \text{Volume} (the number of shares traded on the previous week, in billions), \text{Today} (the percentage return on the week in question) and \text{Direction} (whether the market was \text{Up} or \text{Down} on a given week). In this problem, a prediction is based on whether the predicted probability of a market increase is greater than or less than 0.5.

Download the file \text{Weekly.csv} that contains the data into your working directory. You can read the data using the following commands:

import pandas as pd

# Weekly data set
weekly_data_raw = pd.read_csv("Weekly.csv", usecols=range(1, 10))
mapping = {"Up": 1, "Down": 0}
weekly_data = weekly_data_raw.replace({"Direction":mapping})

X = weekly_data.ix[:, :-1]
y = weekly_data.ix[:, -1]

You will use logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA) to predict that the market will go up. To this end, you will use \text{scikit-learn}, a very popular Machine Learning library in Python. Here is an example on how you can fit logistic regression, LDA, and QDA using the whole data set, and then compute the overall fraction of correct predictions for the whole data.

from sklearn import discriminant_analysis

# instantiate a logistic regression model, and fit with X and y
model_lr = linear_model.LogisticRegression()
model_lr.fit(X, y)

# check the accuracy of logistic regression on the data set
score_lr = model_lr.score(X, y)

# instantiate an LDA model, and fit with X and y
model_lda = discriminant_analysis.LinearDiscriminantAnalysis()
model_lda.fit(X, y)

# check the accuracy of LDA on the data set
score_lda = model_lda.score(X, y)

# instantiate a QDA model, and fit with X and y
model_qda = discriminant_analysis.QuadraticDiscriminantAnalysis()
model_qda.fit(X, y)

# check the accuracy of QDA on the data set
score_qda = model_qda.score(X, y)

_for this problem, use the Python functions implemented in programming assignments, and the three scikitlearn classes mentioned above. Show your computations by displaying the function calls._

(a) In this problem, we will predict the direction of the market using Lag1 and Lag2. Using the validation set approach, estimate the test error of the three models. In order to do this, you must perform the following steps for each model:

i. Split the sample set into a training set and a validation set.
ii. Fit the model using only the training observations.
iii. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

Do not forget to set a random seed before beginning your analysis! Note: You may write one for loop.

(b) Repeat the process in (a) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

(c) Now consider the three models using Lag1, Lag2 and Lag3 as the predictors. Estimate the test errors for the models using the validation set approach. Comment on whether or not including Lag3 leads to a reduction in the test error rate.
Problem 2. We will now perform cross-validation on a simulated data set. 
For this problem, use the Python functions implemented in programming assignments. Show your computations by displaying the function calls.

(a) Generate a simulated data set as follows.

```python
import numpy as np

# Fixing random state for reproducibility
np.random.seed(1)

x = np.random.randn(100)
eps = np.random.randn(100)
y = x - 2 * x**2 + eps
```

In this data set, what is $n$ and what is $p$? Write out the model used to generate the data in equation form.

(b) Create a scatterplot of $X$ against $Y$. Comment on what you find.

(c) Compute the LOOCV error that results from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \varepsilon$
ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$
iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$.

(d) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

(e) Repeat (c) using 10-fold CV for 5 different random seeds corresponding to different splits of the data into ten parts. Plot the 10-fold CV error curves for the four models.

(f) Which of the models in (c) had the smallest 10-fold CV errors?