Problem 1

*Adapted from *Predicting Life Expectancy in the United States* – page 419 in The Analytics Edge by Bertsimas, O'Hair, and Pulleyblank.

Clustering stock returns

The Nasdaq Stock Market is an American stock exchange that was founded in 1971. It is the second largest stock exchange in the world with a market cap of $10 trillion. When investing in stocks, investors typically build investment portfolios that include different stocks with different characteristics (e.g., risk, volatility, expected return, etc.). This approach allows investors to diversify their investment and hopefully obtain more stable returns.

This problem will focus on clustering monthly stock returns on the Nasdaq stock exchange from 2000-2009. The companies included in the file *NasdaqReturns.csv* were listed for the entire duration and their stock price did not fall below $1. The data dictionary is given in Table 1.

Table 1: Data dictionary for NasdaqReturns.csv

Variable	Description
StockSymbol	The stock market symbol that identifies the company
Industry	The industry the company is classified under
SubIndustry	The sub-industry the company is classified under
RetYYYY.MM	There are 120 return features; one for each month from 2000.00 to 2009.12. The value represents the percentage return from the stock during that month. For example, a value of 0.1 means that the stock increased in value by 10% during that month, while -0.02 means that the stock decreased by -2%.

1. You will begin by exploring the data.
   (a) How many companies are included in the data?
   (b) What are the industries and how many companies are included in each industry?
   (c) During December 2000, how many companies had a return of at least 10%? How many had a return less than -10%?
   (d) The 2007-2008 Financial Crisis caused a major global recession. October 2008 was one of the worst months in history for global stock markets with many record declines of over 10%. In October 2008, which three industries had the largest decline? What sub-industries actually increased during October 2008?
2. Next, you will train a hierarchical clustering model using the *RetYYYY.MM* features (i.e., do not cluster using StockSymbol, Industry, or SubIndustry). Please use the Euclidean distance metric.

(a) Why do we not need to normalize our data?

(b) Visualize the dendrogram (hint: use the code from the course website).

(c) Which linkage criteria provides the most balanced set of clusters? Use that criteria for the remainder of the problem.

(d) What do you think is an appropriate number of clusters? Why? Use that number for the remainder of the problem.

3. Now train a hierarchical clustering model using the *AgglomerativeClustering* function. Choose the number of clusters and the linkage metric according to your decisions above. For each cluster, determine the following:

   (a) Number of companies

   (b) Most common industry

   (c) Most common sub-industry

   (d) The centroid is a 120-dimensional vector that represents the average return over all companies in the cluster for each month (1 dimension = 1 month). What is the mean of the centroid (i.e., the average return over the entire 10 years)?

   (e) What is the standard deviation of the centroid (i.e., the monthly variation in return over the 10 years)?

   (f) What was the average return in February 2000? March 2000? What do you see? What event is responsible for this?

4. Now you will train a k-means clustering model using the same number of clusters and distance metric as your heirarchical model (note that the k-means function in sklearn can only use euclidean distance). For each cluster, determine the following (same as above):

   (a) Number of companies

   (b) Most common industry

   (c) Most common sub-industry

   (d) The centroid is a 120-dimensional vector that represents the average return over all companies in the cluster for each month (1 dimension = 1 month). What is the mean of the centroid (i.e., the average return over the entire 10 years)?

   (e) What is the standard deviation of the centroid (i.e., the monthly variation in return over the 10 years)?

   (f) What was the average return in February 2000? March 2000?

5. Are the clusters produced by heirarchical and k-means similar or different?

6. Imagine that you are an investor. How could you use the results of your clustering analysis to help build a stock portfolio?

Problem 2

Explaining *and Predicting* Life Expectancy in the United States during the 1970s

In this problem, you will continue to investigate the data from *StateData.csv*, which includes state-level data collected during the 1970s for all fifty US states. The data dictionary is given in Table 2.

Table 2: Data dictionary for StateData.csv

Variable	Description
Population	Population
Income	Per capita income
Illiteracy	Illiteracy rates (percentage of state population)
LifeExp	Life expectancy (in years)
Murder	Murder and non-negligent manslaughter rate per 100,000 population
HighSchoolGrad	High-school graduation rate
Frost	Average number of days (over the last 30 years) with a minimum temperature below freezing in the state capital or a major city
Area	Land area (square miles)
Longitude	Longitude of the center of the state
Latitude	Latitude of the center of the state
Region	The region that the state belongs to (Northeast, South, North Central, or West)

1. This question focuses on explaining life expectancy. Use all data to train your models.

   (a) Retrain the linear regression model with the seven features used in Homework 1 (Population, Income, Illiteracy, Murder, HighSchoolGrad, Frost, and Area). Which feature is the most important? Explain your choice.

   (b) Use the *DecisionTreeRegressor* function to train a decision tree model with the default settings using the same features. Visualize the tree using the *export_graphviz* function (see the lab for details). Which feature is the most important? Explain.

   (c) Which model do you think will have better out-of-sample performance (assuming they are applied as is)? Why?

   (d) Which model do you find easier to interpret? Explain.

2. This question focuses on predicting life expectancy. We will train both a regression tree and a linear regression model with lasso (L1-regularization). Follow the instructions carefully.

   (a) Use the *train_test_split* function to partition the data with a test size of 0.2 and a random state equal to 1. Train both a lasso (hint: use the *Lasso* function)

and CART model with default hyperparameters. Apply each model to the test set and report the R^2 values (hint: use *.score*). Which model performed better?

(b) Repeat part (a) with a random seed of 2. What are the new R^2 values? Which model performed better?

(c) Part (a) and (b) should produce very different results. Why do you think this is happening? Suggest an approach to fix the problem.

(d) Use the *KFold* function with a random state of 1 and shuffle = True to perform 10-fold cross validation for both CART and lasso (using default hyperparameters). What is the average R-squared for each model? Why are the models performing this way? Suggest a way to improve model performance.

(e) We will build off the 10-fold cross validation scheme from part (d). For each fold, use the *GridSearchCV* function to perform 10-fold cross validation using the training set to determine the best hyper-parameters. For CART, search over minimum leaf sizes in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and for lasso, search over alpha values in $\{0.001, 0.01, 0.1, 1, 10\}$. Apply the best CART and lasso models to the test set. What are the average R^2 values? Which model performed better?

(f) How do the results from part (d) compare to part (e). Why did this happen?

Problem 3

Suppose you train a KNN classifier, CART classifier, and regularized logistic regression using a large number of features (and a proportionally large number of observations). You suspect that some features are much more important than others, and that there may be many useless features. Assume that you have already removed all correlated features and that each model is trained to produce the best out-of-sample accuracy (i.e., we've minimized overfitting). Please answer the following questions:

1. Which model do you expect to have the worst out-of-sample prediction accuracy? Why?

2. Why is CART more likely to overfit as compared to KNN or logistic regression?

3. Suggest and describe two approaches for reducing overfitting that are specific to CART.

4. Suggest and describe one approach for reducing overfitting that can be applied to all three models.

5. How can you verify that your suggested approaches are actually reducing overfitting?