IS 621: Business Analytics and Data Mining

Spring 2015: Final Examination

Instructions

- Your responses to this examination are due by the end of day **Monday, May 25, 2015**. Any extension beyond this deadline requires special permission, since grades will be due shortly after this date.
- Only the last submission will be graded.
- This examination is meant to be a combination of test and educational experience. Much of it is review, but you are also walked through a few new items. With that in mind, a collaborative approach is acceptable. However, if you work with others, you are responsible for making sure you fully understand your answers in case you are asked to elaborate on one of them. In addition, please make sure to identify clearly those you work with.
- I will be available for our course meetup on Thursday, May 21, 2015, from 6:30 p.m. to 7:30 p.m. Eastern time, to discuss concerns or questions on the exam.
- Please note that in part 1 below you are only required to do one of the two questions. You may do both for extra credit if you like, but you will not be penalized if you choose not to do so!

Part 1: Classification and Regression

Option 1: Classification

The attached classification dataset contains training data for crime prediction for various neighborhoods of a major city. Your job is to build a classification model that takes given inputs and predicts whether the neighborhood will be at risk for high crime levels. A description of the variables is also attached for reference.

Your deliverables:

- A short (half page) description of your final model and how effective it is. You may assume you are addressing me as a fellow data scientist and you do not need to shy away from technical details.
- Assigned classifications and probabilities for the evaluation dataset (also attached but without answers!).

Option 2: Regression

The attached cigarette consumption dataset contains information from a 1970 study of cigarette usage in the United States. The dataset has been modified in two ways. First, it contains only three predictor variables of the original seven, since we will be building a simple model. Second, five states have been removed and placed into a separate evaluation file. You will provide estimates for those five states using your final model.

Suggested Analysis Steps

1. Build a regression model using all three predictor variables. How well does your model seem to do? (Evaluate the model on the basis of (a) mean squared error, (b) R^2, and (c) a plot of the residual errors against each predictor to look for any patterns.
2. Try leaving each of the predictor variables out one at a time. Do any of these new models seem to be better than the full model? Explain.
3. Try using each of the three predictor models as a single predictor in a simple linear regression model. How well does each one do? How much variation does each predictor seem to account for? Explain.
4. Which of your seven models would you use? Why?

Your deliverables:

- A short (half page) description of your final selected model and how effective it is. You may assume you are addressing me as a fellow data scientist and you do not need to shy away from technical details.
- Assigned estimations for the evaluation dataset (also attached but without answers!).
- You do not need to provide detailed answers to the four questions above. Those are a suggestion for how you might approach the problem.

Part 2: Clustering

In this second part, we will examine a new technique. (This is really a chance for me to sneak in a bit of clustering before we finish the semester! I'm hoping you enjoy this. Grading will be generous…) In particular, we will look at segmentation, also known as clustering. Generally speaking, we wish to take a dataset and break the observations into a number of clusters that capture similarity. We will first have a short tutorial, walking you through the basic steps of k-means clustering. We will then have a challenge problem at the end in which you will be asked to perform a cluster analysis.

Tutorial

We begin with a dataset of dietary habits of various countries in Europe in 1973 (notice both East and West Germany!). The dataset contains information on sources of protein in the diets of people in the various countries. The goal of the analysis is to group the countries based on patterns in protein consumption. (This example is from the book *Practical Data Science with R*, which is a decent book!)

We will try two clustering techniques.

1. Read the data into R as a dataframe called **protein**.
2. We will first perform hierarchical clustering. Hierarchical clustering creates a dendrogram that we will try to use to identify groupings that may seem natural.
    a. We will create a scaled version of the dataset so that each column is centered at mean 0 and standard deviation 1. (This is similar to what we do in k-nearest neighbor analysis.) Use the scale() function to scale each column accordingly. You should end up with a matrix called **proteinmatrix** of scaled values (the country names should not be included in this scaled matrix). Here is the code:

```
proteinmatrix <- scale(protein[,2:10])
attr(proteinmatrix,"scaled:center") # see the centers
attr(proteinmatrix,"scaled:scale") # see the std devs
```

    b. Create a distance matrix that gives the distances between observations. We will use Euclidean distance. Here is the code:

```
distances <- dist(proteinmatrix, method="euclidean")
```

    c. Run the clustering. Here is the code:

```
protein.hierarchical <- hclust(distances, method="ward.D")
```

    d. Plot the dendrogram. Here is the code:

```
plot(protein.hierarchical, labels=protein$Country)
```

    And that's it. Of course, we need to take a look and see if we see logical patterns in the dendrogram. Do you see any? It certainly looks to me like there could be a decent case for five clusters. The clusters even make sense. Do you agree?

    There's obviously much more to hierarchical clustering, but we'll leave it at this and move on to the next technique.

3. Next up is k-means clustering. This time, we will try to identify the logical clustering pattern by specifying a number of clusters and trying to break the data up into that number of clusters. Naturally, we might want to try more than one number of clusters! Again, we'll take it in steps.
    a. Again, we need the data scaled. We'll use identical code to the above scaling code:

```
proteinmatrix <- scale(protein[,2:10])
```

    b. Now we will use the kmeans() function to perform the clustering. Here is the code:

```
protein.kmeans <- kmeans(proteinmatrix, centers=5,
     iter.max=100, nstart=100)
```

    I've chosen five clusters (centers=5) to match what we found using hierarchical techniques. You should also explore three, four, and six clusters to see if you get more meaningful results.

    c. We wish to evaluate the clusters now. The various details available to us include various metrics, including the total sum of squares and the within-cluster sum of squares, etc. We can explore these features with the following code, though we won't worry about this in this course:

```
summary(protein.kmeans)
protein.kmeans$cluster
protein.kmeans$totss
protein.kmeans$withinss
protein.kmeans$size
```

    And so on…

    d. Let's assign the cluster labels back to the original data. The resulting output vector **protein.kmeans$cluster** can be appended to the original dataframe. Here is the code:

```
protein$cluster <- protein.kmeans$cluster
proteinsorted <- protein[order(protein$cluster),]
View(proteinsorted) # inspect the dataframe
```

    Do you notice any similarities between these cluster assignments and the results of the hierarchical approach? You should! (This is not always true, though. It's a bit lucky that this has happened. Usually we have to make judgment calls about which are best.)

Challenge Problem

Your assignment for the exam is to cluster the world's nations using three different variables. Instead of protein consumption, we'll take a look at population, median GDP growth over the past three years, and an economic trade index that measures openness to trade.

Using the country cluster data, apply the above techniques to determine a good clustering of nations based on these variables. You may restrict yourself to a number of clusters between 3 and 6 (inclusive).

Your deliverables:

- Cluster assignments for each country in a CSV file
- A brief (half page at most) description and assessment of the result you obtained.