**WEEK 2 HOMEWORK­­**

**INSTRUCTIONS**

* Every learner should submit his/her own homework solutions. However, you are allowed to discuss the homework with each other (in fact, I encourage you to form groups and/or use the forums) – but everyone must submit his/her own solution; you may not copy someone else’s solution.
* The homework will be peer-graded. In analytics modeling, there are often lots of different approaches that work well, and I want you to see not just your own, but also others.
* The homework grading scale reflects the fact that the primary purpose of homework is learning:

|  |  |  |
| --- | --- | --- |
| **Rating** | **Meaning** | **Point value (out of 100)** |
| 4 | All correct (perhaps except a few details) with a deeper solution than expected | 100 |
| 3 | Most or all correct | 90 |
| 2 | Not correct, but a reasonable attempt | 75 |
| 1 | Not correct, insufficient effort | 50 |
| 0 | Not submitted | 0 |

# Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

**Question 4.2**

The *iris* data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Iris> ). *The response values are only given to see how well a specific method performed and should not be used to build the model.*

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

# Kai: Note to Menno: Question:

# My question here is whether we should scale the data. I’m thinking YES, given that the range for length and width can differ in range, my question is then which kind of scaling should I use?

# I’ve seen two methods (written in R code)

1. Scaling within each column
   1. for (i in 1:4) { scdata[,i] <- (iris[,i]-min(iris[,i]))/(max(iris[,i])-min(iris[,i])) }
2. Scaling using R’s scale() function
   1. irisScale<- scale(iris[,-ncol(iris)])

# I’m not very sure what the difference is between methods 1 and 2.

I tried both, method #1 seems to give values from between 0 and 1.

# Kai: I also have one more question on the R script on how to pipe in the data columns into my kmeans function

**Question 5.1**

Using crime data from the file uscrime.txt (<http://www.statsci.org/data/general/uscrime.txt>, description at <http://www.statsci.org/data/general/uscrime.html>), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

# Kai: In my R code I’ve tested for normality based on: Shapiro.test and the qqplot, I’m not that sure how to interpret the results for these though..

Also I’ve run the grubbs test in R, though I know that we can specify type=10 or type=20

**Question 6.1**

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

**Question 6.2**

1. Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can get the data that you need from the file temps.txt or online, for example at <http://www.iweathernet.com/atlanta-weather-records> or <https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html> . You can use R if you’d like, but it’s straightforward enough that an Excel spreadsheet can easily do the job too.
2. Use a CUSUM approach to make a judgment of whether Atlanta’s summer climate has gotten warmer in that time (and if so, when).