**DATA 512 Predictive Analytics: Estimation and Clustering**

**Course Project 3: Clustering**

**Prof Larose**

**Name:**

Use the *cc.tr* and *cc.te* data sets. The datasets summarize the usage behavior of active credit card holders over six months.

**The project instructions are shown in bold.** This is to distinguish the instructions from your work. Your work should be in not-bold.

* Work neatly. Aim for a professional-looking presentation. I will be grading your level of professionalism, as well as your English expression.
* Make sure all graphs and tables fit neatly on the page.
* Neither add nor delete pages. Use single-spacing.

**For all text output, surround it with a text box. In Word, select the text output, then Insert > Text Box > Draw Text Box.**

Apart from this document, which you will save as a pdf and submit, you must submit your R script, containing the code you used to solve the problems. The R script should be neat and easily understandable by people who are not you. It should be well-annotated, describing what you are doing so that anyone could understand it.

This Project is brand new, and may have typos, errors, etc, that I have missed. Please report these to me asap. For this and other reasons, this Project is subject to change at any time (though of course I will be reasonable.)

I am aware that I am not asking about validating the regression assumptions here. Of course, you should do so in the real world. But this project is long enough, and I think you deserve a break.

**Standardize the predictors, but not the target, *balance*.**

**Make sure you do not include the target *balance* when creating clusters.**

**You may need to set the data sets to be data frames.**

**Install and library the *caret, psych, Kohonen,* and *NbClust* packages.**

**Set seed to 12345.**

**Good luck!**

**Prof Larose**

1. **Insert your Executive Summary here. (A strategy for this is given at the end.)**
2. **Import the *cc.tr* training data set. Use the CH Criterion to determine the optimal number of clusters, for *k* = 2 … 8. Provide the CH values for all candidate models. Let the winning value of *k* be denoted as , and the runner-up as . Report and .**

$All.index

2 3 4 5 6 7 8

2478.522 2154.693 1934.772 1757.754 1723.844 1655.559 1514.996

$Best.nc

Number\_clusters Value\_Index

2.000 2478.522

1. **Use the Mean Silhouette Criterion to determine the optimal number of clusters, for *k* = 2 … 8. Provide the Mean Silhouette values for all candidate models. Let the winning value of *k* be denoted as , and the runner-up as Report and .**

$All.index

2 3 4 5 6 7 8

0.5302 0.4802 0.3954 0.3826 0.3801 0.3913 0.3648

$Best.nc

Number\_clusters Value\_Index

2.0000 0.5302

= 0.5302 where k = 2

1. **We introduce a third measure of cluster goodness: the Predictive Clustering Criterion. The Predictive Clustering Criterion selects the value of *k* that obtains the best predictive metrics, when the clusters are used as the sole predictors.** 
   1. **Use *k-*means clustering to develop clusters, using . Perform a linear regression of *balance* versus the clusters only (no other predictors).**
   2. **Repeat (a) for the clusters only.**
   3. **Repeat (a) for the clusters only.**
   4. **If your is distinct from the other values of *k*, repeat (a) for . Otherwise skip (d).**
   5. **Construct a nice table in Word containing CH-metric, the mean silhouette, the *s* and the values, for each different value of *k*. The values of *k* should be in the columns of the table in increasing order. For each metric, highlight the best in bold green font and the worst in red font.**

**The following rather inelegant code will work to get this done, for *k* = 4. I would like to hear from an R expert who can show how to do this more elegantly.**

k4 <- kmeans(ccnbs, centers = 4, nstart = 25)

k4cluster <- k4$cluster

k4cluster <- as.factor(k4cluster)

k4cluster <- data.frame(k4cluster)

dummy\_model <- dummyVars(" ~ .", data = k4cluster)

k4cluster.d <- predict(dummy\_model, newdata = k4cluster)

reg4 <- lm(CC$balance ~ k4cluster.d)

summary(reg4)

1. **Thoroughly discuss your findings from the table in the previous problem. Then, weighing each of the criteria, attempt to find a consensus as to the optimal value of *k*.**
2. **Use *k-*means clustering to develop clusters, using . Then construct a nice table in Word containing the record count and the means for all variables and all clusters. Highlight the max and min means using green and red font. Then provide brief but comprehensive profiles of each cluster.**
3. **Use Kohonen networks clustering to develop clusters, using and hexagonal structure. Then construct a nice table in Word containing the record count and the means for all variables and all clusters. Highlight the max and min means using green and red font. Copy the nice table from the *k*-means clustering model. Compare the two clustering solutions.**
4. **Use *k-*means clustering to develop clusters, using . Then construct a nice table in Word containing the record count and the means for all variables and all clusters. Highlight the max and min means using green and red font. Then provide brief but comprehensive profiles of each cluster.**
5. **Use Kohonen networks clustering to develop clusters, using and hexagonal structure. Then construct a nice table in Word containing the record count and the means for all variables and all clusters. Highlight the max and min means using green and red font. Copy the nice table from the *k*-means clustering model. Compare the two clustering solutions.**
6. **For , compare the Kohonen clusters with the *k-*means clusters, similar to Video 14. For what proportion of records do the two clustering methods concur? Are there any systematic or interesting ways the two methods diverge?**
7. **For , compare the Kohonen clusters with the *k-*means clusters, similar to Video 14. For what proportion of records do the two clustering methods concur? Are there any systematic or interesting ways the two methods diverge?**
8. **Develop cluster profiles for the *k-*means model using . Construct a nice table in Word containing the record count and the means for all variables and all clusters. Highlight the max and min means using green and red font. Then, provide brief but comprehensive profiles of the clusters.**
9. **Import the *cc.te* test data set. Validate the *k*-means clustering solution for .**
10. **Validate the *k*-means clustering solution for .**
11. **Validate the *k*-means clustering solution for .**