DUKE UNIVERSITY
The Fuqua School of Business

MGRECON 491 Data Mining — Gallant
Homework 3 — Term 1
Due October 10, 2008

The goal of this assignment is to determine a model that can be used by the same Czech bank considered in Homeworks 1 and 2 to cluster customers into groups on the basis of transactions data in hopes of learning background information that can help with formulation of bank policies and suggest future data mining activities. Keep a broad mind. For instance, one might discover that some customers seem to be using the banks services for criminal activities. This task uses ideas developed in connection with the Intrusion Detection Case and the Value at Risk Case in combination.

The Excel workbook clust.xls, which can be downloaded from the course web site, represents a summary of my initial results from working on the assignment myself and will serve as a guide for what a summary of your work should look like. The principal missing items are graphical or tabular representations of decision trees which is something you may want to include in your report for your best model. Your task is to continue working on the project where I left off. The workbook contains some of the data on customer transactions from Homework 2 in worksheet clust_data and the summary of my work in worksheets PCA, KM, and HC_Ward. I also put three additional workbooks, clust01.xls, clust02.xls, and clust03.xls, on the website to leave an audit trail so that you can verify the information in the summaries.

I began by analyzing all the transactions data from Homework 2 using the principal components tool. The first thing that I noticed was that the tool kept loading on the average variables and ignoring the count variables, so I deleted nearly all of the count variables just to get rid of clutter. The data that remained are in worksheet clust_data of workbook clust.xls. Then I ran principal components again on nearly everything in clust_data and found that principal components loaded on average balance, average cash deposits and withdrawals, average credit card withdrawals, and average wire deposits and withdrawals (rows 25 through 44 of columns D and E of worksheet PCA_Output1 of workbook clust01.xls). It did not load at all on average overdraft fee so I ignored average overdraft fee thereafter. This may have been a mistake as we shall see a few paragraphs hence. Here is the data dictionary for worksheet clust_data of workbook clust.xls:

account_id	account identifier and key for merges with other data bases
first	date of the first transaction in days since baseline
duration	date of last transaction minus the first in days
numtrans	number of transactions of all types over the period first to last
avbal	average balance over the period in Czech currency
avdposit	average deposit of all types over the period
avwdraw	average withdrawal of all types over the period
avccardwdraw	average withdrawal by credit card
avcashdposit	average cash deposit
avwiredposit	average wire transfer deposit
avcashwdraw	average cash withdrawal
avwirewdraw	average wire withdrawal
avinspment	average insurance payments
avstmntpment	average statement service charge
avintcredit	average interest credit
avodraftfee	average overdraft interest charge
avmortpment	average number of mortgage or rent payments
avpensndposit	average pension deposit
avloanpment	average loan payment

As you will recall, what principal components tries to do is find derived features that are weighted sums of the original features that account for most of the variance of the original features. To be more specific, you will see at the right of worksheet PCA of workbook clust.xls a block labeled "Principal Components" that consist of two sets of columns of factor loadings, one copied from XLMiner output (worksheet PCA_Output2 of workbook clust01.xls) and the other the same thing severely rounded. In terms of the rounded factor loadings, the derived features are

$$p1 = 0.8 \times \text{avbal} + 0.3 \times \text{avdposit} + \ldots + 0.2 \times \text{avcashwdraw}$$

$$p2 = 0.1 \times \text{avbal} + 0.1 \times \text{avdposit} + \ldots + 0.9 \times \text{avwiredposit}$$

These are the variables $p1$ and $p2$ shown in columns A and B of the worksheet (copied from PCA_Scores2 of clust01.xls). They are plotted in the figure labeled PCA_Plot2.

What the plot suggests to me are that there are probably at most three groups of customers. The loadings for $p1$ suggest that one major factor in distinguishing customers is the size of the account because $p1$ seems to be just a measure of account size. The variable $p2$ is hard to understand. My guess is $p2$ is just a dummy variable for wire deposits since wire deposits are rare in the data and have a lot of zeroes. The variable $p2$ probably acts almost like a zero-one dummy.

Another thought I had was that account growth over time might be important. A derived variable that would measure this is

$$\text{growth} = [\text{avdposit} - \text{avwdraw}]/\text{duration}$$

Add this variable to the worksheet clust.xls. See if it helps with your analysis. If it doesn't, document the fact in your report.

Moving on, let us discuss the summary of the k-means analysis in worksheet KM. I used three means and ten random starts. K-means settled on the same answer labeled best several times so we can have reasonable confidence that the best clustering was found. From k-means we get a classification of each row of data into one of three groups. These are in the Cluster id variable in worksheet KM (which come from worksheet KM_Clusters1 of workbook clust03.xls). In the Value at Risk case I plotted $p1$ and $p2$ and colored the points according to the Cluster id variable. That doesn't work here (see worksheet KM_PC_Plot1 in clust03.xls). We need to do something else to get a graphical understanding of what k-means is doing.

We will use classification tools help understand the k-means output. There are two classification tools that are useful in this regard: linear discriminant analysis and classification trees. One can use linear discriminant analysis to classify the points and thereby get scores and loadings that serve the same purpose as the scores and loadings of principal component analysis. They are called cannonical scores and loadings should work better because they are actually derived from the k-means classification. To do this, I added Cluster id as a column to clust_data (which is worksheet KM_Data_Scores1 in workbook clust03.xls) and ran discriminant analysis. The cannonical scores are shown in the first columns A and B of worksheet KM of workbook clust.xls (which come from worksheet DA_TrainCanSco1 of clust03.xls) and Cluster id is in column C. To get a plot that would show groups as different colors I sorted columns A, B, and C on column C and then split them apart as seen in the worksheet and then plotted them as a scatter plot. Looking at the loadings to the right, one sees that the X and Y axes for the plot are

$$\begin{aligned} X &= 1.25 \times \text{avbal} - 1.1 \times \text{avwdraw} + 1.5 \times \text{avcashdposit} + 1.25 \times \text{avcashwdraw} \\ Y &= -3 \times \text{avdposit} - 2 \times \text{avccardwdraw} + 3 \times \text{avwiredposit} \end{aligned}$$

Taking into account the fact that deposits and withdrawals are about a tenth the size of average balance, X is probably a measure of account size. (One could check this by running a regression of avbal on X. If the R^2 is above 0.8, then the surmise would be confirmed.) The variable Y seems to be ranking customers on the basis of their credit card and wire deposit activity. Looking at the cluster centers shown below the cannonical variate loadings in worksheet KM of workbook clust.xls (from worksheet KM_Output1 in worksheet clust03.xls), this seems to be confirmed. In fact, no one in group 3 has a wire deposit and no one in groups 1 and 2 has had a credit card withdrawal. Apparently only customers with large average balances are allowed to use their credit card as a debit card by why they should have no wire deposits is a mystery. At any rate, it is easy to understand what linear discriminant analysis thinks that k-means is doing. A customer is in group 3 (yellow) if $X > 0.75$, in group 2 (pink) if $X < 0.75$ and $Y < 0.4$, and in group 1 (blue) if $X < 0.75$ and $Y > 0.4$ (see worksheet KM_Plot1 of workbook clust03.xls). Taking into account the fact that avccardwdraw is zero in groups 1 and 2 and that the magnitude of avwiredposit is larger than that of avdposit, it seems that the split between groups 1 and 2 is being made on the basis of avwiredposit. This is just to gain understanding. In practice we would actually score the customer's data using k-means scoring rather than using rules based on X and Y.

Another way to understand how k-means is grouping is to use classification trees. What one learns (worksheet CT_FullTree1 of workbook clust03.xls) is that k-means is mainly splitting on avbal and then trying to further pick customers apart on the basis of their deposit history. This confirms our interpretation of X but clouds our interpretation of Y. We next try a coarser classification tree by making the minimum number of records in a terminal node large. Putting the minimum number of records to 600 we get the rule group 2 if avbal < 35272.6, group 1 if avbal > 35272.6 and avcashdposit < 14954.9, and in group 3 if avbal > 35272.6 and avcashdposit < 14954.9 (worksheet CT_FullTree2 of workbook clust03.xls). The classification error rate for the tree is 9.64% (worksheet CT_Output2 of workbook clust03.xls) and for linear discriminant analysis it is 8.82% (worksheet DA_Output1 of workbook clust03.xls). One might trust the interpretation from linear discriminant analysis a bit more because of this.

I then repeated all the steps in the k-means cluster analysis using Wards hierarchical clustering method instead. The steps are the same. The only thing to call attention to is that one sets # Clusters to 3 in the "Hierarchical Clustering - Step 3 of 3" dialog box. Also, hierarchical clustering will only consider 4000 variables. It takes the first 4000 of the data in your input worksheet. At some point in your analysis you will have to drop down to 4000 variables. You may want to do that by copying worksheet clust_data to an new worksheet, randomizing the rows by sorting the entire worksheet on a column of random numbers as explained in Homework 1, then deleting the last 500 rows. The plot in worksheet HC_Ward of workbook clust.xls is interesting because of the mistake that I made by including avodraftfee when I did the linear discriminant classification. Linear discriminant analysis seems to load heavily on it but not as heavily as it might seem because avodraftfee is about three orders of magnitude smaller than variables like avbal and avdposit. Nonetheless, use avodraftfee in your analysis. See if it helps with your analysis. If it doesn't, document the fact in your report.

Your assignment is to complete my analysis. Do one k-means analysis the same as I described above using the variables avbal, avdposit, avwdraw, avccardwdraw, avcashdposit, avwiredposit, avcashwdraw, avwirewdraw, growth, avodraftfee. Do another using avbal, avdposit, avwdraw, growth, avodraftfee and another using avbal, avdposit, avwdraw, growth. Form a conclusion as to which works best. Remember that clustering is a subjective activity and deciding which works best is a subjective judgement on your part. You will have to defend your choice. An effective defense usually consists of finding loadings that make sense and would be operational for classifying customers for some business purpose. One business purpose would be asking tellers to suggest the purchase of a CD to a group that seems to load up on either a high balance or a high flow of funds through the account. You are free to consider some other set of variables or to derive variables of your own and use them if you can document their superiority.

When you have selected a set of variables use the hierarchical clustering tools to repeat one of the analysis. Use Ward's method.

As with Homeworks 1 and 2, the best way to organize your thoughts for presenting your work is to think of yourself as a consultant advising this Czech bank. This carries with it the presumption that the technical level of your audience is lower than yours so you will have to explain the ideas behind the tools that you use and your results. Present your results as a PowerPoint presentation that takes no more than twenty minutes to present. Teams will present their work to the class. Here are the main points to address in your presentation.

- Present your clustering results.
  - Describe the tools that you used to obtain these these results.
  - Describe the entire analysis.
  - For each clustering tool that you used, include a chart the shows the clusters plotted against linear discriminant analysis canonical scores and explain what information it conveys.
  - For each clustering tool, include the loadings for the canonical scores and interpret them.
  - For each clustering tool, describe what the classification tree suggested about interpretation of the clusters. Include the tree itself in your presentation if you think it is helpful. You may want to limit the size of the tree so that it is interpretable.
- Explain the aims and limitations of your analysis.
  - Explain what cluster analysis tries to accomplish.
    - Explain its subjective nature and why users can disagree on the interpretations of results.
    - Interpret your results and suggest business purposes to which they may be put and what they may suggest for future data mining activities.

What to turn in: Please submit your PowerPoint presentation and your Excel workbooks on a CD. The Excel worksheet files clust01.xls, etc. will provide an audit trail where I can trace the results shown in your presentation, if necessary. Your presentation should describe: (a) the goal of the analysis and what you learned about the bank's consumer business from your analysis, (b) why you chose the cluster model that you did in terms that a layman would understand. Try to follow good principles of statistical presentation i.e., describe the data and try to make its message as clear as possible. Feel free to comment on what you feel might be the underlying causes for the patterns you have observed.

Your presentation should begin with one or two slides that highlight your key conclusions and recommendations especially the bottom-line clustering model. Next, it should include some slides that address the background questions addressed. Be sure to include a slide or two describing the data variables and the units in which they were measured. The presentation should include a few slides that illustrate your solution and the process followed in reaching it.