

DSC 433/533 – Class 18 – Clustering example for Churn dataset

This class exercise uses the Excel dataset **Churn.xls** (available on the data page of the course website).

- Run a *hierarchical clustering* analysis of the 7 variables in columns A to G of the spreadsheet. In particular:
 - Select XLMiner > Data Reduction and Exploration > Hierarchical Clustering.
 - Make sure “churn” is selected as the Worksheet, “Variable names in the first row” is checked, and “# Rows in data” is 3333. Also, “Data type” should be set to “Raw data.”
 - Select the variables “Account Length” through “CustServ Calls” and click “Next”.
 - Check “Normalize input data,” make sure “Euclidean distance” is selected, select “Ward’s method” for the “Clustering method,” and click “Next.”
 - Leave “Draw dendrogram” selected, but uncheck “Show cluster membership,” and click “Finish.”

You should obtain the following results on the “HC_Output1” worksheet:

| Stage | Cluster 1 | Cluster 2 | Distance | | |
|-------|-----------|-----------|-----------|---|------|
| 1 | 949 | 2496 | 0.006306 | | |
| 2 | 263 | 2247 | 0.017861 | | |
| 3 | 2499 | 3095 | 0.020969 | | |
| 3325 | 11 | 21 | 416.63985 | 8 | 12% |
| 3326 | 16 | 22 | 417.86253 | 7 | 0% |
| 3327 | 6 | 11 | 459.0788 | 6 | 10% |
| 3328 | 3 | 33 | 622.60618 | 5 | 36% |
| 3329 | 3 | 4 | 812.80828 | 4 | 31% |
| 3330 | 3 | 6 | 942.79366 | 3 | 16% |
| 3331 | 3 | 16 | 1263.4521 | 2 | 34% |
| 3332 | 1 | 3 | 2729.1969 | 1 | 116% |

The final two columns were added manually. For example going from 6 to 5 clusters, the distance (of the two clusters joined at each stage) increases from 459.0788 to 622.60618, i.e., $(622.60618/459.0788) - 1 = 36\%$. The “distance” indicates the difference between the two clusters that are being joined at each stage – the larger the distance the more different the clusters being joined. Thus, a 6-cluster solution seems reasonable here (why?).

The dendrogram on the “HC_Dendogram1” worksheet provides a graphical representation of the final 30 stages of the hierarchical clustering, in which 30 “sub-clusters” are sequentially joined together. For example, on the left of the dendrogram, at a distance of 193.6 (scale on the vertical axis), sub-cluster 1 joined sub-cluster 11 to form a new, larger cluster. You can see which cases are in each sub-cluster listed just below the dendrogram, e.g., case #1 is in sub-cluster 1 and case #17 is in sub-cluster 11.

- The results above suggest 6 useful clusters could be present. For a full analysis, a number of different cluster solutions could be investigated (e.g., 6, 5, 4, and 3), but for this exercise we’ll just investigate a 6-cluster solution. The hierarchical approach just used can be useful for identifying possible values for the number (k) of useful clusters that there are. However, it may be sub-optimal for actually assigning cases to clusters, since once two cases are joined together in a cluster, they remain forever joined, even if the cases might fit better with different clusters that form at a later stage (of the clustering process). A better approach for assigning cases to clusters is to use *k-means clustering*, which iteratively assigns cases to a fixed set of k clusters, until each case is assigned to the most suitable cluster (i.e., contains other cases that are the most similar to it):
 - Select XLMiner > Data Reduction and Exploration > K-Means Clustering.
 - Ensure “churn” is the selected Worksheet, “First row contains headers” is checked, and “# Rows in data” is 3333.
 - Select variables “Account Length” through “CustServ Calls” to be the “Input variables” and click “Next”.
 - Check “Normalize input data,” type “6” for the “# Clusters,” leave “# Iterations” at 10, leave “Fixed Start” selected, and click “Next.” [“Random starts” gives odd results, so don’t use this option.]

- Leave “Show data summary” and “Show distances from each cluster center” selected, and click “Finish.” You should obtain the following data summary on the “KM_Output1” worksheet:

| Cluster | #Obs | Average distance in cluster |
|-----------|------|-----------------------------|
| Cluster-1 | 547 | 1.942 |
| Cluster-2 | 603 | 1.949 |
| Cluster-3 | 504 | 1.973 |
| Cluster-4 | 526 | 1.977 |
| Cluster-5 | 431 | 2.181 |
| Cluster-6 | 722 | 2.255 |
| Overall | 3333 | 2.052 |

The “average” account length in each of these clusters is also given:

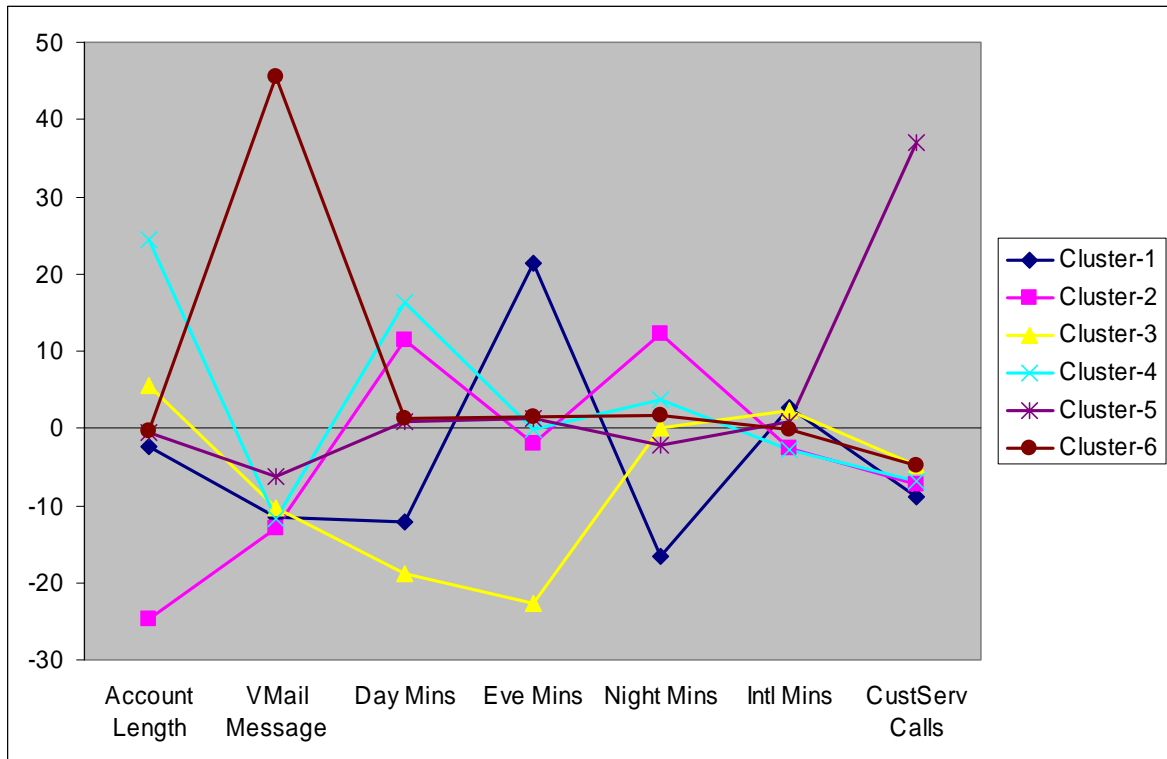
| Cluster | Account Length |
|-----------|----------------|
| Cluster-1 | 96.867669 |
| Cluster-2 | 61.003408 |
| Cluster-3 | 111.059412 |
| Cluster-4 | 143.371437 |
| Cluster-5 | 100.000004 |
| Cluster-6 | 100.465928 |

This is not quite the same as the arithmetic average of the variable for each cluster member, but instead represents a “geometric” center of the cluster with respect to the variable.

- To profile the clusters (as we began to do in the last step by looking at the cluster centers for one of the variables), it helps to put the cluster centers on a comparable scale using, for example, a calculation based on the “one-sample t-statistic”. For each variable, calculate $(\text{value} - \text{mean}) / (\text{sd} / \sqrt{\text{cluster-size}})$, where “value” is the cluster center, “mean” is the average over the whole sample, “sd” is the standard deviation over the whole sample, and “cluster-size” is 547 for cluster-1, 603 for cluster-2, etc. For example, the calculations for the “Account Length” variable are as follows:

| | C | D |
|----|-----------|--|
| 27 | Cluster | Account Length |
| 28 | Cluster-1 | 96.867669 |
| 29 | Cluster-2 | 61.003408 |
| 30 | Cluster-3 | 111.059412 |
| 31 | Cluster-4 | 143.371437 |
| 32 | Cluster-5 | 100.000004 |
| 33 | Cluster-6 | 100.465928 |
| 34 | Cluster | Account Length |
| 35 | Cluster-1 | $-2.46503242 = (D28 - D\$41) / (D\$42 / \text{SQRT}(547))$ |
| 36 | Cluster-2 | $-24.7036166 = (D29 - D\$41) / (D\$42 / \text{SQRT}(603))$ |
| 37 | Cluster-3 | $5.63451712 = (D30 - D\$41) / (D\$42 / \text{SQRT}(504))$ |
| 38 | Cluster-4 | $24.3655987 = (D31 - D\$41) / (D\$42 / \text{SQRT}(526))$ |
| 39 | Cluster-5 | $-0.55511564 = (D32 - D\$41) / (D\$42 / \text{SQRT}(431))$ |
| 40 | Cluster-6 | $-0.40409463 = (D33 - D\$41) / (D\$42 / \text{SQRT}(722))$ |
| 41 | mean | $101.064806 = \text{AVERAGE}(\text{churn!A2:A3334})$ |
| 42 | sd | $39.8221059 = \text{STDEV}(\text{churn!A2:A3334})$ |

You should be able to insert 9 rows below row 33, type these formulas into cells D35-D42, and copy them across columns E to J to do similar calculations for all the variables. Then highlight cells C34-J40 and use the “chart wizard” to create a line graph like the following:



This suggests the following descriptions of the clusters:

| | |
|-----------|---|
| Cluster-1 | Lo-med acct length, v.lo vmail, lo day, v.hi eve, v.lo night, hi-med intl, lo custserv |
| Cluster-2 | V.lo acct length, v.lo vmail, hi day, lo-med eve, v.hi night, lo-med intl, lo custserv |
| Cluster-3 | Hi-med acct length, v.lo vmail, v.lo day, v.lo eve, med night, hi-med intl, lo custserv |
| Cluster-4 | V.hi acct length, v.lo vmail, v.hi day, med eve, hi-med night, lo-med intl, lo custserv |
| Cluster-5 | Med acct length, lo vmail, med day, med eve, lo-med night, med intl, v.hi custserv |
| Cluster-6 | Med acct length, v.hi vmail, med day, med eve, med night, med intl, lo custserv |

4. The “KM_Clusters1” provides the cluster id numbers for each case in the dataset. If you copy this to the “churn” worksheet you can see how the clusters (distinct customer segments) might provide useful business information with respect to other variables in the dataset. For example, you can use PivotTables (see hw 1) to find the proportion of churners, international plan subscribers, and voice mail subscribers in each cluster:

| | churn | intlplan | vmailplan |
|-----------|--------|----------|-----------|
| Cluster-1 | lo | med | |
| Cluster-2 | hi-med | lo-med | |
| Cluster-3 | lo | lo | |
| Cluster-4 | hi | hi | |
| Cluster-5 | v.hi | med | 15% |
| Cluster-6 | lo | hi-med | 100% |

Thus, we might want to focus retention efforts on clusters 5 (lots of customer service calls), 4 (long account lengths, high day mins), and 2 (short account lengths, high night mins). Also, we might want to recommend the international plan to cluster 3 (high intl mins but low intl plan subscribers), and voice mail to cluster 5 with only 15% voice mail subscribers (similar usage patterns to cluster 6 with 100% voice mail subscribers).