

Assignment #1

Due Date: Oct. 7th, 2009 (Wed)

Reading Assignment:

Read pages 1-8 of the paper on CCC and answer the following question

- Sarle, W. (1983). [Cubic Clustering Criterion](#). *Technical Report A-108*, SAS Institute, Inc., 1983.

Suppose that K clusters are formed for observed data

$\{x_{ij}, i=1, \dots, n \text{ and } j=1, \dots, p\}$ and vector $\bar{x}_k = (\bar{x}_{k1}, \dots, \bar{x}_{kp})'$ is the center for the k -th cluster where $k=1, \dots, K$. Give the specific expression of R^2 in terms of (x_{ij}, \bar{x}_{kj}) .

Theoretical Problem 1

1. Show that the sample correlation coefficient, r , can be written as

$$r = \frac{ad - bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$$

for two binary variables with the following 2×2 contingency table

	0	1
0	a	b
1	c	d

2. Suppose the correlation matrix is

$$\begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ 1 & & & & & & & \\ .643 & 1 & & & & & & \\ -.103 & -.348 & 1 & & & & & \\ -.82 & -.086 & .100 & 1 & & & & \\ -.259 & -.260 & .435 & .034 & 1 & & & \\ -.152 & -.010 & .028 & -.288 & .176 & 1 & & \\ .045 & .211 & .115 & -.164 & -.019 & -.374 & 1 & \\ -.013 & -.328 & .005 & .486 & -.007 & -.561 & -.185 & 1 \end{pmatrix}$$

Use Single linkage and complete linkage to find the clusters *by hand*.

R project:

The famous `iris` data set (Fisher's or Anderson's) gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

Let's do a cluster analysis project with this data set. The goal is to cluster the flowers according to their `Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width` information. Eventually we want to see whether they fall into three clusters, which corresponds to the three species. and complete the following questions.

- Make a heat map of the above dissimilarity matrix
- Cluster the data using K-Means methods with $k = 3$.
- Cluster the data using an hierarchical clustering method of your choice. Plot the related dendrogram. Note that you might want to study R functions `hclust()` and `dist()` (as well as function `daisy()` in the `cluster` library) among others. Obtain the results for three clusters.
- Applying multidimensional scaling technique, plot the data using the first and second components. Specify the cluster membership for each observation (i.e., flower) with different colors and specify its species with different symbols. Make this kind of plotting for both results you got using K-Means and hierarchical clustering methods.

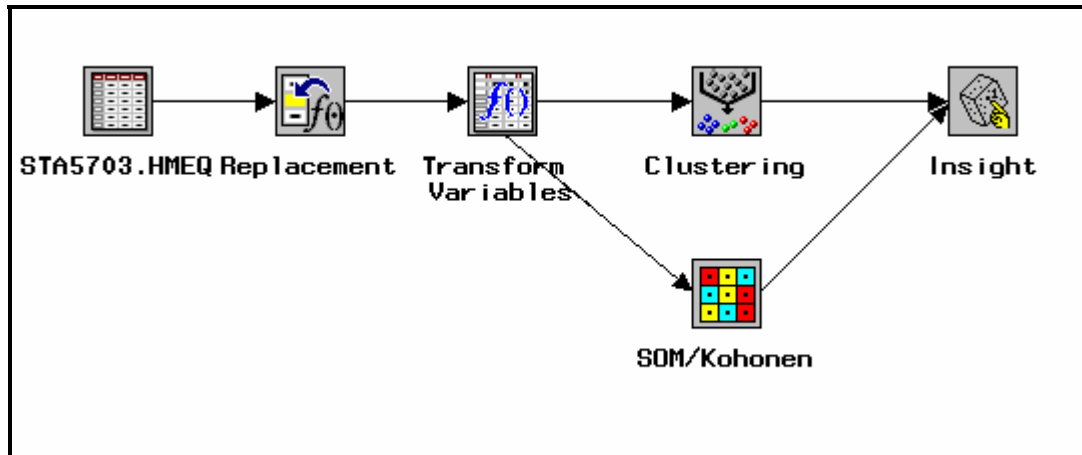
SAS Computer Project:

Data:

The data set (HMEQ), available in SAS EM library `SAMPPIO`, was obtained from a financial services company who extends credit lines to homeowners. This data set includes 12 predict variables and one target variable. The target variable, `BAD`, is a binary variable with a value 0 when the homeowner repaid the loan and a value 1 when the client defaulted on the loan. There are a total of 5,960 homeowners in the data set. The output form *PROC CONTENTS* is as follows:

-----Alphabetic List of Variables and Attributes-----					
#	Variable	Type	Len	Pos	Label
1	BAD	Num	8	0	0 is good credit and 1 is bad credit
10	CLAGE	Num	8	56	Age of the Oldest Trade Line in months
12	CLNO	Num	8	72	
13	DEBTINC	Num	8	80	
9	DELINQ	Num	8	48	Delinquent Trade Lines
8	DEROG	Num	8	40	
6	JOB	Char	7	95	Job Category
2	LOAN	Num	8	8	Loan Amount
3	MORTDUE	Num	8	16	Amount Due on Existing Mortgage
11	NINQ	Num	8	64	
5	REASON	Char	7	88	Reason for the Loan: HOMEIMP or DEBTCON
4	VALUE	Num	8	24	Property Value
7	YOJ	Num	8	32	Years at current Job

Each observation represents an applicant for a home equity loan. The target is a binary variable indicating whether the applicant eventually defaulted or was ever seriously delinquent. This adverse outcome occurred in 1189 of the cases (19.95%).



Data Replacement node

Task:

- List the missing rate for each variable.
- Do not create imputed indicator variables for missing values
- Replace missing values for both JOB and REASON with default constant “Unknown”
- Set the default imputation method for interval variable to “*mean*”
- Change the imputation method to “*median*” for DEROG, DELINQ, and NINQ

Question: What are the imputed values for VALUE, DEROG, DELINQ, NINQ, MORTDUE, YOJ, CLAGE, DEBTINC, CLNO, JOB, and REASON?

Transformation Node

Task:

- **Create a new variable LTV (LOAN/VALUE)**
- **Perform Log Transformation on the following variables:** MORTDUE, YOJ, and CLAGE
- **Perform Optimal Binning for Relationship with Target variables On** DEBTINC and CLNO (use initial number of bins = 4)

Questions:

- How did you create the variable LTV?
- Produce a Screen Shot of the current screen.
- How many “*Input Variables*” in the data now?

Clustering Node - Perform a cluster analysis of the data.

Analysis I:

- (a) What options, other than the default ones, do you select?
- (b) Why do you want to exclude “BAD” as well?
- (c) Do you want to use categorical variables in cluster analysis if you want to use K-means method? Explain. What is the most important variable in the cluster selection process if you use categorical variable?
- (d) What is the optimal number of clusters based on CCC = 3 if you use all variables except “JOB” and “BAD”?
- (e) What cluster has the highest frequency?
- (f) What cluster has the largest radius?
- (g) What cluster has roughly normally distributed MORTDUE?
- (h) What cluster has the highest loan amount difference between BAD=0 and BAD=1?
- (i) Get a screenshot of part of the tree profile. What variables seem to be important in allocating the clusters?
- (j) Use the tools available in the Clustering Node to explore each cluster.

Analysis II:

In this data set, the optimal number of clusters is five that can be judged by the target variable frequency. Rerun the cluster analysis using appropriate options and find the best five clusters. Report the relative frequency table for each cluster and the overall frequency for the target variable “BAD” to show that you indeed find the best possible clusters. Note that it is possible some cases do not belong to any cluster due to missing target values.

SOM Node (Self-Study)

- (a) Do some self study on self organizing maps (SOM). Explain its main idea using several sentences.
- (b) Perform an SOM analysis of the data and explore the clusters. What variables are important?

Insight Node

- (a) Use Insight node, in particular, the Boxplot/Mosaic method, to further explore the clusters found by the Clustering node and the SOM node. In particular, focus on the important variables identified earlier.
- (b) Carefully describe the characteristics of each cluster in detail.