Review Topics for
Final Exam
Eco 5385
Predictive Analytics for Economists
Summer I, 2014
Tom Fomby

The final exam is scheduled for **Tuesday, July 1, from 6:00 – 9:00 PM in 251 Maguire**. It is a comprehensive exam with about 33% of it consisting of the material covered on the mid-term exam and the remainder consisting of material covered since the mid-term exam. The format of the test will be similar to that of the Mid-term exam.

First, I recommend that you study all the QQs and Exercises, the Mid-term exam, the Mid-term Review Sheet, and the topics mentioned below. Also study all of the PPTs that I have covered in class, the pdf files that I have mentioned in the PPTs and in class, and read all of the related material in your textbook.

The first topic we covered after the mid-term was **classification methods** including CART, CHAID, and Logit models as well as the curves, charts, and measures used to evaluate the performances of these classification methods including (see relevant PPTs, pdfs, and your textbook):

- Classification (Coincidence) Matrices
- Type I Error (False Negative), Type II Error(False Positive)
- General Payoff Matrix
- Profit Matrix for Target Marketing
- Cumulative Profit Curve and the Maximum Cumulative Profit used to compare classifiers and set "penetration" rates.
- Symmetric and Asymmetric Loss Matrices
- Accuracy Rate and Error Rate
- Weighted Error Rate
- The role of Cutoff (Threshold) probabilities in the case of Asymmetric Loss
- Cumulative Gain Charts and Areas under Cumulative Gain Charts to compare classifiers
- Cumulative Lift Charts and Decile-by-Decile Lift Charts and how they are used to compare classifiers
- ROC curves and the areas under ROC curves to compare classifiers
- The AC_d measure associated with the ROC curve and how it is used to compare classifiers and to choose an "optimal" cutoff probability.

Following our discussion of classification methods and their evaluation tools, we studied the unsupervised learning methods of **cluster analysis** and (time permitting) **association rules.**

With respect to cluster analysis we discussed **hierarchical clustering** (both agglomerative and divisive) and the **dendrogram**. We then discussed K-Means clustering with random starts. I discussed the **Silhouette Measure of Cohesion and Separation** and the role it plays in determining **cluster quality across different numbers of clusters <u>within a cluster method</u>** and the **<u>relative quality</u> of clusters produced by different clustering methods** (as used by the Auto Cluster Node in SPSS Modeler in choosing a "best" cluster method). The **Silhouette measure** of a given cluster arrangement (average case silhouette) is given by

$$S = \sum_{i=1}^{N} \frac{S_i}{N}$$

where the i-th case's Silhouette measure is given by

$$S_i = \frac{B_i - A_i}{\max(A_i, B_i)} \ , \ \ i = 1, 2, \cdots, N.$$

Here N represents the number of cases used to derive the cluster arrangement, A_i is the distance of the i-th case from the centroid of the cluster the case belongs to and B_i is the distance of the i-th case from the nearest centroid of the other clusters. If S = 1 then all of the cases are located on their respective cluster centroids (good). If S = -1 then all cases are located on the cluster centers of some other cluster (bad). A value of S = 0 means, on average, cases are equi-distant between their own cluster centroid and the centroid of the nearest other cluster. According to Kaufman and Rousseeuw (1990), a "poor" cluster arrangement has an average case silhouette that is between -1 and 0.2 $(-1 \leq S < 0.2)$, a "fair" cluster arrangement has an average case silhouette that is between 0.2 and 0.5 $(-0.2 \leq S < 0.5)$, and a "good" cluster arrangement has an average case silhouette that is between 0.5 and 1.0 $((0.5 \leq S \leq 1.0)$. Of course, for a single clustering method, say, K-means, one could compare the average case silhouette measure across a range of clusters, say from K = 2, 3, … , K*, and determine the optimal number of clusters to apply for one's analysis of the data. The number of clusters that maximizes the average case silhouette measure while hopefully achieving at least a fair score is the best cluster arrangement **for that method**. Denote this set of within-method average case silhouette measures as $\{S_{(K)}$, K = 2, 3, … , K*}. In a similar manner, different clusters methods (as in the SPSS Modeler Auto Cluster node) can be compared examining their optimal average case silhouette measures. The clustering method that produces the best average case silhouette measure across its own number of clusters is judged to be the best clustering method in the sense of the Auto Cluster node.

In addition to the Silhouette Measure, we talked about **variable importance measurement in cluster analysis**. In essence the more a variable explains of the **Within Cluster Sum of Squares** in a cluster arrangement, the more important the variable is in the given cluster arrangement. It is very useful to know the important variables because it is these variables that will give the investigator the best chance of identifying the distinguishing characteristics of the various clusters. In turn this allows the investigator the best opportunity to put meaningful "labels" on the derived clusters for descriptive purposes.

The key terms for **association rules** are consequent, antecedent, support, confidence of a rule, lift of a rule and the logic of the A priori Algorithm of Agrawal and Srikant used in deriving the association rules. What are the steps in this algorithm?

Finally, (time permitting) I gave you a short introduction to **Text Analytics** and the **Singular Value Decomposition** (SVD) of rectangular matrices. See PPT 18 for the discussion on Text Analytics and the Singular Value Decomposition. In short, these tools are used to provide quantitative measures of major features in text data sets (the so-called SVD variables) that can be used as inputs in prediction and classification models.

You have been a great class. It has been a pleasure teaching you some topics that I definitely have fun discussing. Have a good rest of your summer and good luck in the future.
TF