Review Topics for
Mid-term Exam
Eco 5385
Data Mining Techniques for Economists
Fall 2012
Tom Fomby

- The mid-term exam is scheduled for Thursday, October 18 during our regular class period. The exam is going to be focusing on (1) the **Introduction to Data Mining**, (2) **Exploration and Modification of the data**, and (3) **Prediction** in data mining and the tools used in prediction. The topics of **Classification** and Unsupervised learning methods like **Principal Components analysis** and **Cluster Analysis** are going to be reserved for the final exam. To the extent that you have remaining questions after going through this document, I will be in my office on Wednesday, October 17 (day before exam) from 4:00 – 5:30 PM to answer any questions that you have concerning the exam. Also, of course, I will be holding office hours in the hour and one-half before the exam.
- In terms of Exercises and QQs, the test will cover the topics in Exercises 1 – 6 and QQs 1 – 7. See the class website for keys to these exercises and Quick Quizzes. Also I would suggest that you consult the applicable chapters in your textbook and the "HELP" file in XLMiner for additional discussion of techniques covered in class.
- By **Introduction to Data Mining** I mean the ideas in Professor Leo Breiman's "two cultures" paper that can be found on the class website - the first culture being the "statistical hypothesis testing" culture and the second culture being the "prediction" culture or sometimes called the "cross-validation" culture. Be sure and understand the difference between these two cultures and the roles played by these cultures in quantitative economic analysis. For additional discussions of the history of Data Mining and some of the basic tasks that Data Mining takes on, go to Lectures 1 – 6 PPTs on the class website. We talked about **different types of variables** (target, input, interval, categorical, nominal, ordinal, and the like). You need to know these types. Also we talked about the principal tasks undertaken by data mining, namely **prediction** and **classification** (which are lumped together into what is called **supervised learning**) and **unsupervised learning** (lumping together tasks that do not involve either prediction or classification,) Your textbook covers reasons why data mining is becoming so important to businesses and government agencies in improving their decision making. Businesses and governments are overrun with data and are seeking ways to fruitfully use it.
- **Exploration and Modification of the data.** By **exploration of data**, I mean the construction of histograms of the data, Box-Plots to detect outliers in the data, the computation of summary statistics of the data like mean, median, and standard deviation, and the identification and treatment of missing observations in the data. By **modification of the data** I mean the ideas of binning data, the creation of "dummy" (binary categorical) variables from multi-valued, unordered categorical variables (also called "nominal" variables), the construction of "rank" variables

from ordered categorical variables (also called "ordinal" variables), and the formation of any transformations of the data that might appear useful (like taking logs of variables and forming cross-product and higher powers of variables of interest).

- **Prediction** covers those techniques used to predict interval target variables and how to build good prediction models through the use of data partitioning and the principle of cross-validation. We have covered the prediction techniques of MLR, K-NN, ANN, CART/CHAID, and Ensemble methods.
- The exam will be styled much like a QQ quiz but with more questions and there also could be some short answer questions and the analysis of some XLMINER output that I will provide with the exam. You can check out a few previous mid-term exams on the class website to get a feel of what to expect in terms of mid-term exam questions.
- In terms of the pdf files that I have discussed in class see the following: **Scoring Measures for Prediction Problems.pdf; Multiple Linear Regression and Subset Selection.pdf; K-Nearest Neighbors.pdf; K-NN Method.pdf, ANNs.pdf; Regression and Classification Trees.pdf; and Regression and Classification Trees_Fomby.pdf.**
- In addition I would suggest that you consider the following key phrases and concepts:

Some Key Phrases and Concepts

- Different types of variables: Interval variables; Categorical variables – nominal versus ordinal.
- Distinction between Prediction and Classification problems.
- Distinction between supervised learning and unsupervised learning.
- Various Tasks Associated with "Data Handling": Treatment of missing observations; detection of outliers and use of the Box-Plot; binning interval variables; creation of categorical variables from group designations; over-sampling in the case of rare classification events, etc.
- Terms such as input variable, output variable, cases.
- Data Partitioning: What is its purpose?
- What are the distinctive roles of the training, validation, and test data sets?
- What does the phrase "over-training a model" mean? How is over-training avoided? What are the consequences of over-training? Can you draw a graph that represents the consequences of over-training?
- Breiman says that "data mining stands at the confluence of statistics and machine learning." What does Breiman mean by his statement?
- List the prediction methods that XLMINER supports. What are their tuning parameters? What is a tuning parameter?
- Validation of the "goodness" of a proposed data mining method is usually carried out by "scoring" a "trained" model on a validation data set and then examining the accuracy of the model vis-à-vis its competitors. (This is called the technique of **Cross-Validation**.) How is "accuracy" measured in prediction problems in XLMINER? What are some of the other predictive

accuracy measures one could use? What loss function is implied by the RMSE accuracy measure in prediction problems?

- You should know the basic logic of the various data mining techniques.
- With respect to multiple linear regression (MLR), what is the difference between backward selection, forward selection, stepwise selection (all being **directed search** methods) and **comprehensive selection** procedures involving Adjusted R-square and Mallows Cp criteria
- Can you, in words, tell me how the K-nearest neighbors method works and the tuning parameter of this method?
- How are prediction trees built? What is the nature of the binary recursive method of building regression trees? Name two potential tuning parameters for building regression trees using the CART method. How does the CHAID model building procedure differ for the CART method of building regression trees?
- What do we mean by the **architecture** of an Artificial Neural Network? How do you choose an "optimal" architecture of an ANN model? What is the difference between an input layer, hidden layer(s), and an output layer? Hidden nodes, and activation functions. Given an XLMINER output, can you write out the formulas for an ANN model?
- Given some XLMINER output, could you discern which tuning parameter value of a particular data mining model is the best?
- What are **Ensemble Methods**? Why are Ensemble Methods frequently useful for prediction? How do they work? How do you construct a Granger-Ramanathan fixed-weight Ensemble prediction model? A Simple Average Ensemble Model? Are there diminishing returns to putting more and more models into an Ensemble model? If so, approximately how many models should you consider for your Ensemble?