

**Review Topics for
Mid-term Exam
Eco 5385
Predictive Analytics for Economists
Spring 2014**

- The mid-term exam is scheduled for Thursday, March 6 at 5:00 pm in our classroom.
- The format of the test is going to be modeled after your QQs and Homework assignments. That is, you are going to have multiple choice questions, fill-in-the blank and short answer questions with a little computation, definitions, and interpretation of computer or spreadsheet output coming from XLMINER or SPSS Modeler. So I would study your graded exercises and QQs first. (The keys are available on the class website.) Also I would study your class notes and the Power Point slides that I have gone through in class and have sent to you. The chapters we have covered in the book are Chapters 1 - 3, Chapter 4 (apart from Section 4.7 on Principal Components which I will cover after the midterm), Chapter 5 with emphasis on predictive measures of accuracy but not the discussion of the evaluation of Binary classification models (that is, excluding Sections 5.1- 5.2). The midterm will also cover Chapter 6 (Multiple Linear Regression), K-Nearest Neighbors (Chapter 7), and Artificial Neural Networks (Chapter 11). The CART and CHAID trees of Chapter 9 for both prediction and classification problems will be topics covered **after** the midterm exam.
- In terms of the supplementary pdf files that I have gone over in class, please see the PPTs that you have been given (PPT1 – PPT8). The supplementary pdf files are listed within the PowerPoint presentations.

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?

- Professor Breiman of Stanford has said “data mining stands at the confluence of statistics and machine learning.” What does he mean by this by this 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?
- 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?