MIS2502: Final Exam Study Guide

The exam will be a combination of multiple-choice and short-answer questions. It is a closed-book, closed-notes exam.

You will not be able to use a computer during the exam. You should bring a calculator!

The following is a list of items that you should review in preparation for the exam. Note that not every item on this list may be on the exam, and there may be items on the exam not on this list.

Data Mining and Data Analytics Techniques

- Explain the three data analytics techniques we covered in the course
    - Decision Trees, Clustering, and Association Rules
    - What kinds of problems can each solve? Provide a business-oriented example.
    - Make recommendations to a business based on the results of each type of analysis.
- Explain how data mining differs from OLAP analysis
    - Why would you use this instead of a data cube and a pivot table?

Using R and RStudio

You will not write blocks of R code for this exam. However, be familiar with basic syntax.

- Explain the difference between R and RStudio
- The role of packages in R
- Generate and explain basic syntax for R, for example:
    - Variable assignment
    - Identify functions versus variables
    - Identify how to access a variable (column) from a dataset (table)

Understanding Descriptive Statistics (Introduction to R)

- Be able to read and interpret a histogram
- Be able to read and interpret sample (descriptive) statistics

Decision Tree Analysis

- The data necessary to perform a decision tree analysis
- Role and structure of input and predictor variables in a decision tree
    - Why do decision trees typically have categorical outcome variables?
- Interpret a decision tree: determine the probability of an event happening based on predictor values
- Understand the meaning of the complexity factor, minimum split, and the size of the training partition and how it can alter the decision tree
- Compute/interpret Chi-Squared statistic for a split variable; determine which is stronger predictor

<<CONTINUED ON NEXT PAGE>>

Cluster Analysis (Cluster Analysis Using R)

- The data necessary to perform a cluster analysis
- Be able to read the output from a cluster analysis
    - And interpret a scatter plot of 2 dimensional data (i.e., the baseball example from the slides)
- Interpret what the boxplot and histogram tells you about each variable
- Interpret within cluster sum of squares and between cluster sum of squares
    - Relate them to cohesion and separation
    - What does it mean when those values are larger (or smaller)?
    - What happens to those statistics as the number of clusters increases?
    - What is the advantage of fewer clusters?
- Interpret standardized cluster means for each input variable
    - Describe a particular cluster in relation to the average for the entire data set

Association Rules (Association Mining Using R)

- The data necessary to perform an association rule analysis
- Be able to read and interpret the output from an association rule analysis
    - Find the strongest (or weakest) rule from a set of output
- Understand the difference between confidence, lift, and support
    - You should be able to explain the difference between them
        - Can you have high confidence and low lift?
- Given a set of baskets, compute and interpret confidence, support, and lift for an association rule.
- Given a table of aggregate purchase numbers for two products, compute and interpret the lift for the rule based on those two products (i.e., the Netflix/Comcast example from class).

Data Visualization

- Be able to assess an infographic or chart by applying data visualization principles.
    - Tell a story
    - Graphical integrity (lie factor)
    - Minimize graphical complexity (data ink)
- Explain how a visualization can be improved based on those principles.

And don't forget…

- ERDs, including cardinality, entities, and attributes
- Single table SQL queries and joins
- The difference between a transactional database and an analytical data store
- The ETL process and resolving data inconsistencies
- Semi-structured data, including JSON, XML, csv; how applications exchange data