MIS2502: Exam 3 Study Guide (Spring 2023)
Instructor: Jeremy Shafer

**Overview**

* Exam 3 is a closed-book, closed-notes exam.
* The exam is composed of roughly 25 questions total with varied point value.
* Part 1 of the exam is made up of 16 multiple choice questions worth 3 points each.
* Parts 2, 3 and 4 are a combination of multiple choice and short answer questions where you must read a diagram, read a chart, or perform some calculation.
* There are also 5 bonus questions at the end of the exam. They are worth 2 points each.
* There are 100 points possible, on the exam, plus 10 additional points for the bonus questions.
* Calculator Policy: You can use a calculator that does not have the ability to communicate with other electronic devices. (You are not allowed to use your smartphone’s calculator.)
* Numeric answers should be rounded to the *third* decimal place. For example, the value 7.777777 would be rounded to 7.778.
* **When computing lift or confidence, use your rounded support values.**
* You will be required to turn in your printed copy of the exam, along with any related scrap paper you are done.

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 Advanced Data Analytics Techniques**

* Explain the three advanced data analytics techniques we covered in the course
	+ Decision Trees, Clustering, and Association Rules
	+ What kinds of problems can each solve?
	+ Identify business-oriented examples of each.
* Explain how data mining differs from analysis we did using SQL/NoSQL/ETL

**Using Jupyter and Python**

*You will not need to generate blocks of Python code for this exam. However, you should be familiar with the basic syntax.*

* The role of packages in Python
* Generate and explain basic syntax for Python, for example:
	+ Variable assignment
	+ Identify functions versus variables
	+ Identify how to access a variable (column) from a dataset (table)

CONTINUED

**Understanding Descriptive Statistics (Introduction to Python)**

* Be able to read and interpret sample (descriptive) statistics
* Be able to read and interpret results from simple hypothesis testing (e.g., the p-value of a t-test)

**Decision Tree Analysis (Decision Trees in Python)**

* Understand what classification is and when it is appropriate to use this technique
* Role and structure of input and predictor variables in a decision tree
* Understand the basic idea behind the decision tree algorithm
* Interpret a decision tree: determine the probability of an event happening based on predictor variable values
* Understand the meaning of the minimum split (MINIMUMSPLIT), and how it can alter the decision tree
* Compute error rate and correct classification rate based on a confusion matrix

**Cluster Analysis (Cluster Analysis Using Python)**

* Understand what cluster analysis is and when it is appropriate to use this technique
* Understand the basic idea behind K-means clustering algorithm
	+ K: the number of clusters, which we must specify in advance
	+ What is a centroid?
* Interpret within-cluster sum of squares error and between-cluster sum of squares error
	+ Within-cluster sum of squares error is also known as within-cluster SSE, or “WSS” in Python
	+ Between-cluster sum of squares error is also known as between-cluster SSE, or “BSS” in Python
	+ 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?
		- Higher separation, and easier to interpret
* Interpret normalized cluster means (centroid) for each variable
	+ Describe a particular cluster mean (centroid) in relation to the population average

**Association Rules (Association Rules Using Python)**

* Understand what association rule analysis is and when it is appropriate to use this technique
* Understand the basic idea behind association rule algorithm
* 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 and be able to explain the difference between support, confidence, and lift
	+ Can you have high confidence and low lift?
* Given a set of baskets, compute and interpret support, confidence, 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/Cable TV example from class)