MIS2502: Exam 3 Study Guide (Spring 2023)

Instructor: Leila Hosseini

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

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? Provide a business-oriented example.
* 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)

**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 maximum depth (max\_depth) and minimum split (min\_split), 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 have to specify in advance
	+ What is a centroid?
* Interpret within-cluster sum of squares error (WSS error) and between-cluster sum of squares error (BSS error)
	+ Within-cluster sum of squares error is also known as within-cluster SSE, or “withinss” in Python
	+ Between-cluster sum of squares error is also known as between-cluster SSE, or “betweenss” 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**

* 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)