ISTE-470 Assignment 1

##### Fall 2221 Name: Daniel Chung

Answer the following questions and submit them to the Assignment 1 dropbox by the specified due date.

Scoring: Questions 1-9 are worth 9 points each; Question 10 is worth 19 points.

1. List the 5 steps involved in extracting knowledge from data.

* Selection
* Preprocessing
* Transformation
* Data Mining
* Interpretation/Evaluation

2. Why is domain expertise not necessarily required by an analyst who is performing data mining on a domain-specific data set?

Although it helps to have domain expertise when analyzing data. It’s possible to learn what is required to understand the data sets without having to be an expert on the topic. The data tends to speak by itself when grouped and you can see and understand.

3. List the 5 data mining techniques discussed in the Intro to Data Mining lecture and for each, if they are supervised or unsupervised data mining approaches.

* Classification (supervised)
* Anomaly Detection (supervised)
* Estimation (supervised)
* Clustering (unsupervised)
* Association rule mining (unsupervised)

4. What is the difference between classification and estimation?

* Classification, deals with discrete outcomes
* Estimation deals with outcomes whose values are continuous in nature

5. Amazon tries to sell more products by presenting customers with advertisements for “similar” or “related” products. Which data mining technique(s) do you think Amazon uses to do this and why?

They use clustering, because they are grouping products that customers have looked for in the past and looking for similarities.

6. What is the “Curse of Dimensionality” and why is it a major problem in data mining?

The curse of dimensionality refers to the problems when working with high-dimensional data (when your data has too many features). This is a problem, because data mining algorithms have difficulty scaling, so if a data set has too many attributes computation complexity also rise.

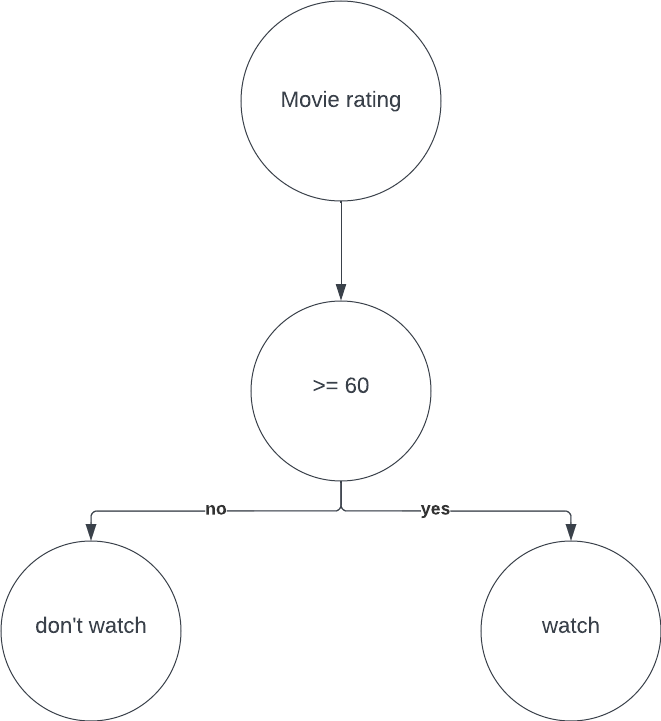
7. “Randomly sampling data from a large data set to create a model will always work well.” Do you agree with this statement? Why or why not?

Yes and no, randomly sampling large data will only work if the data set is big enough to allow so, if the data set is huge, then this can give you an unbiased subset that is easier to process, but if the dataset is small, then you might miss out on some important data.

8. Describe how stratified sampling differs from random sampling.

Stratified sampling will take a data set and split it into partitions and then get random samples from each partition. While random sampling will take a data set and then draw randomly from it. With stratified sampling you’re sure to get at least one data point from each partition, while random sampling you might miss one partition entirely.

9. A movie reviews website provides ratings on a scale of 1 to 100, where 1 is an unfathomably poor film and 100 is a classic work of art. Provide a discretization scheme that would allow users to quickly gauge whether a given movie is worth watching.



10. Using your code that you developed in Exercise 1 – Data Preparation in Python, create an output file using iris.arff called iris\_arff.arff and an output file using iris.csv called iris\_csv.arff. Open each respective output file in Weka and take a screenshot of the Weka Explorer window’s Preprocess tab to prove that it was successfully opened by the software. Insert and label each screenshot below and submit your code, along with this document, to the Assignment 1 dropbox.

