CSC520 Spring 2019 – Assignment 5

Due by midnight of **April** 26^{th} **2019**.

This assignment includes both conceptual and code questions. It must be completed individually. You may not collaborate with other students, share code, or exchange partial answers. Questions involving answers or code must be emailed to the instructor or TAs directly or discussed during office hours. Your answers to the conceptual questions must be uploaded to Moodle as a pdf file titled "Assign5_UnityID.pdf". Your sourcecode must be submitted as a self-contained zip. All code must be clear, readable, and well-commented.

Note: The code will be tested on the NCSU VCL system("CSC520_VCL"). You are advised to test your code there before submission.

Question 2: Bayesian Networks (30 pts)

Consider the following lexicon and rules :

Lexicon:

Hot– It is hot outside.
AC – The air conditioner is turned on.
Cooler– The evaporative cooler is turned on.
Cold – The room temperature goes down.
Humid – There is humidity in the room.

Rules:

If it is hot outside then we may turn on air conditioner or evaporative cooler.
The room temperature goes down if the AC or cooler is running.
The cooler causes humidity in the room

Prior and conditional probabilities known beforehand:

Hot	P(Hot)
T	0.69

Hot	Cooler	P(Cooler|Hot)
T	T	0.35
F	T	0.55

Hot	AC	P(AC|Hot)
T	T	0.68
F	T	0.42

Cooler	Humid	P(Humid|Cooler)
T	T	0.61
F	T	0.35

Cooler	AC	Cold	P(Cold|Cooler, AC)
T	T	T	0.40
T	F	T	0.35
F	T	T	0.20
F	F	T	0.39

1. (5 points) Using the knowledge base, show the corresponding Bayes Net.
2. (10 points) Using the ideas of conditional independence discussed in Ch. 14 of R&N, identify which variables in this network are conditionally independent of the others and under what conditions.
3. (15 points) Use the prior and conditional probabilities provided above to construct probability tables for each node of this network. Then use exact inference to answer the following questions. *Show your work*. What is the probability that
   (a) There is humidity in the room.
   (b) The cooler is not running and AC is running and room temperature is down and there is humidity in the room.
   (c) The room temperature goes down given it is hot outside.

Question 2: Decision Trees (40 pts)

You have been given two datasets represented as CSV files. Each dataset represents a distinct decision problem. Your task in this question is to:

1. (20 points) Implement the decision tree algorithm to analyze these datasets and to make predictions using Information Gain as discussed in class.
2. (10 points) Define and implement a different selection criteria that yields a different result for the datasets.
3. (10 points) In a written document you should:
   (a) Describe your two selection criteria.
   (b) Report on the performance of your implementation on each dataset and the resulting trees.
   (c) Analyze the performance to determine which criteria is better overall.
   (d) Determine if either implementation became overfitted. If so why, if not, why not?

Code

Your code should be implemented to take in two CSV files, a training and a test set, as well as the name of a class variable, it will then generate a decision tree and print the tree as well as a confusion matrix for the test set. You may print the tree as a series of if-then rules or in any other readable way. The commandline argument is:

```
java -jar dt.jar <Train> <Test> <Class>
```

Datasets

You have been given two datasets represented as csv files. Both datasets are described briefly below and may be found on the UCI Machine Learning repository:

- **Spect Heart** Data on heart classification via Single Proton Emission Computed Tomography (SPECT) images.

  http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

- **Mushroom Classification** Identification of poisonous mushrooms.

  http://archive.ics.uci.edu/ml/datasets/Mushroom