**NAME:**

## CS 583 – Assignment 5

### 1. [50 points] Estimating Parameters for Multiple Variables

We have three variables: X, Y, and Z. X and Z are binary with domain {T, F} and Y has three possible values: {R, G, B}. The Bayesian network has the following structure: $X \rightarrow Y \rightarrow Z$. Here are the counts for a dataset D. If a count is zero, it is not listed.

| X | Y | Z | Counts |
|---|---|---|---|
| T | R | T | 10 |
| T | R | F | 20 |
| T | B | T | 30 |
| F | R | F | 40 |
| F | B | T | 50 |

Note that we need to estimate $P(X)$, $P(Y \mid X)$, and $P(Z \mid Y)$ for this network.

**a.** **[15 pts]** What are the MLE estimates?

**b.** **[15 pts]** Assuming a uniform prior and K2 approach to Bayesian estimation, what are the predictive for next X, Y|X, and Z|Y?

**c.** **[20 pts]** Assuming a |D'| = probabilities 12, and P' is uniform, and a BDe approach to estimation, what are the predictive probabilities for next X, Y|X, and Z|Y?

**Constructing Bayesian Network [100pt]**

**Dataset Information**

**Attached file: auto-mpg.csv. Please use data from this curated dataset and not the one from the original source link. Please refer to lecture slides for example code and example IPython Notebooks.**

Original Source: https://archive.ics.uci.edu/ml/datasets/auto+mpg

1. Title: Auto-Mpg Data

2. Sources:
(a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.
(c) Date: July 7, 1993

3. Relevant Information:

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

5. Number of Instances: 100

6. Number of Attributes: 9, including the class attribute

7. Attribute Information:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

1. **[10 points] Prepare dataset for pgmpy**

Load the dataset into IPython notebook or script. Discretize continuous values. Use median value to separate continuous variables into 'high' and 'low' categories.

2. **[45 points] Structure Learning**

a. **[15 pts]** Perform structure learning using conditional independence tests (PC algorithm)
b. **[30 pts]** Perform structure learning using score-based method (Hill Climb Search). Use BIC score, BDs score, and K2 score.

3. **[45 points] Parameter Estimation**

Use the model obtained via the BIC score from the previous method.

a. **[22.5 pts]** Perform parameter estimation using the Maximum Likelihood algorithm. Print out CPDs and local independencies of the network.
b. **[22.5 pts]** Perform parameter estimation using the Expectation Maximization algorithm. Print out CPDs and local independencies of the network.