Written Assignment - Probabilites, Bayesian Networks & Decision Trees

Max points:

- CSE 4308: 100
- CSE 5360: 100

The assignment should be submitted via Canvas.

Instructions

- The answers can be typed as a document or handwritten and scanned.
- Name files as assignment5_<net-id>.<format>
- Accepted document format is .pdf.
  - If you are using Word, OpenOffice or LibreOffice, make sure to save as .pdf.
  - If you are using LaTEX, compile into a .pdf file.
  - Please do not submit .txt files.
- If you are scanning handwritten documents make sure to scan it at a minimum of 600dpi and save as a .pdf or .png file. Do not insert images in word document and submit.
- If there are multiple files in your submission, zip them together as assignment5_<net-id>.zip and submit the .zip file.

Task 1

12 points.

Consider the given joint probabilty distribution for a domain of two variables (Color, Vehicle) :

	Color = Red	Color = Green	Color = Blue
Vehicle = Car	0.1299	0.0195	0.0322
Vehicle = Van	0.1681	0.0252	0.0417
Vehicle = Truck	0.1070	0.0160	0.0265
Vehicle = SUV	0.3103	0.0465	0.0769

Part a: Calculate P (Color is not Green | Vehicle is Truck)

Part b: Prove that Vehicle and Color are totally independant from each other

---

Task 2

15 points.

In a certain probability problem, we have 11 variables: A, B_1, B_2, ..., B_{10}.

- Variable A has 7 values.
- Each of variables B_1, ..., B_{10} have 8 possible values. Each B_i is conditionally indepedent of all other 9 $B_j$variables (with j != i) given A.

Based on these facts:

Part a: How many numbers do you need to store in the joint distribution table of these 11 variables?

Part b: What is the most space-efficient way (in terms of how many numbers you need to store) representation for the joint probability distribution of these 11 variables? How many numbers do you need to store in your solution? Your answer should work with any variables satisfying the assumptions stated above.

Part c: Does this scenario follow the Naive-Bayes model?

Task 3

30 points

As in the slides that we saw in class (Prior and Posterior probabilties), there are five types of bags of candies. Each bag has an infinite amount of candies. We have one of those bags, and we are picking candies out of it. We don't know what type of bag we have, so we want to figure out the probability of each type based on the candies that we have picked.

The five possible hypotheses for our bag are:

- h_1 (prior: 10%): This type of bag contains 100% cherry candies.
- h_2 (prior: 20%): This type of bag contains 75% cherry candies and 25% lime candies.
- h_3 (prior: 40%): This type of bag contains 50% cherry candies and 50% lime candies.
- h_4 (prior: 20%): This type of bag contains 25% cherry candies and 75% lime candies.
- h_5 (prior: 10%): This type of bag contains 100% lime candies.

Given the following sequences of observations show how the Posterior probabilties change.

a. CCCCCL
b. CLCLCL
c. CCCLLL

Task 4

10 points

George doesn't watch much TV in the evening, unless there is a baseball game on. When there is baseball on TV, George is very likely to watch. George has a cat that he feeds most evenings, although he forgets every now and then. He's much more likely to forget when he's watching TV. He's also very unlikely to feed the cat if he has run

out of cat food (although sometimes he gives the cat some of his own food). Design a Bayesian network for modeling the relations between these four events:

- baseball_game_on_TV
- George_watches_TV
- out_of_cat_food
- George_feeds_cat

Your task is to connect these nodes with arrows pointing from causes to effects. No programming is needed for this part, just include an electronic document (PDF, Word file, or OpenOffice document) showing your Bayesian network design.

Task 5

10 points

For the Bayesian network of previous task, the text file at this link contains training data from every evening of an entire year. Every line in this text file corresponds to an evening, and contains four numbers. Each number is a 0 or a 1. In more detail:

- The first number is 0 if there is no baseball game on TV, and 1 if there is a baseball game on TV.
- The second number is 0 if George does not watch TV, and 1 if George watches TV.
- The third number is 0 if George is not out of cat food, and 1 if George is out of cat food.
- The fourth number is 0 if George does not feed the cat, and 1 if George feeds the cat.

Based on the data in this file, determine the probability table for each node in the Bayesian network you have designed for Task 3. You need to include these four tables in the drawing that you produce for question 3. You also need to submit the code/script that computes these probabilities.

Task 6

8 points

Given the network obtained in the previous two tasks, calculate P (Baseball Game on TV | not(George Feeds Cat)) using Inference by Enumeration

---

Task 7

15 points

Class	A	B	C
X	1	2	1
X	2	1	2
X	3	2	2
X	1	3	3
X	1	2	1
Y	2	1	2
Y	3	1	1

Y	2	2	2
Y	3	3	1
Y	2	1	1

We want to build a decision tree that determines whether a certain pattern is of type X or type Y. The decision tree can only use tests that are based on attributes A, B, and C. Each attribute has 3 possible values: 1, 2, 3 (we do not apply any thresholding). We have the 10 training examples, shown on the table (each row corresponds to a training example). What is the information gain of each attribute at the root? Which attribute achieves the highest information gain at the root?