CMSC 422 Assignment 4: Bayesian Models   Spring 2019

1. Naive Bayesian Classifier

Using the prior and conditional probabilities from the example involving stroke recovery discussed in class (see Lecture Slides), apply Bayes' Theorem by hand (a calculator may be used for arithmetic) to compute the probability of a poor outcome for a 50 year old person with a mild thrombotic infarction.

2. Estimating the Probabilities for a Naïve Bayesian Classifier

The data below describe people, some of whom developed lung cancer and some who did not, along with the level of pollution they had been exposed to and whether or not they were a smoker during the last several years.

Cancer	**Pollution**	**Smoker**
no	low	no
no	low	no
no	medium	yes
no	low	no
yes	high	yes
yes	low	no
yes	medium	yes
no	low	no
no	low	yes
no	high	yes
yes	medium	yes
no	low	no
yes	low	yes
no	low	yes
no	medium	yes
yes	medium	yes
yes	high	yes
no	medium	yes
no	high	yes
no	high	yes
yes	medium	no
no	medium	no
yes	high	no
no	high	no

a. Estimate *all* of the probabilities that would be needed for a naive Bayesian classifier that would be used to predict, given this data, whether or not a person would develop cancer based on their pollution exposure and smoking history.

b. Given a naive Bayesian classifier that is using the probabilities learned in (a), what would it predict as the probability that a person would develop cancer if they did not smoke, but were exposed to a high level of pollution?

c. Would this specific data set create any problems for a decision tree learner such as ID3? Why or why not?

3. Cancer Diagnosis

Suppose you wish to predict if a patient will develop lung cancer, based on their smoking habits, exposure to pollution, x-ray results, and dyspnea (difficult breathing). Below is an expanded set of training data related to lung cancer that now includes all four of these input features:

No.	**Cancer**	**Dyspnea**	**Pollution**	**Smoker**	**Xray**
1	no	none	low	no	negative
2	no	none	low	no	negative
3	no	none	medium	yes	positive
4	no	moderate	low	no	negative
5	yes	moderate	high	yes	negative
6	yes	none	low	no	negative
7	yes	none	medium	yes	positive
8	no	none	low	no	negative
9	no	none	low	yes	negative
10	no	severe	high	yes	negative
11	yes	none	medium	yes	positive
12	no	none	low	no	negative
13	yes	none	low	yes	positive
14	no	severe	low	yes	negative
15	no	moderate	medium	yes	positive
16	yes	severe	medium	yes	positive
17	yes	none	high	yes	positive
18	no	moderate	medium	yes	positive
19	no	severe	high	yes	positive
20	no	moderate	high	yes	positive
21	yes	moderate	medium	no	positive
22	no	none	medium	no	negative
23	yes	moderate	high	no	positive
24	no	none	high	no	negative

In this table, the leftmost column *No.* is not a predictive attribute and should not enter into your considerations below in any way. Ignore the fact that there is insufficient data to accurately predict probabilities. Assume values of all attributes are mutually exclusive and exhaustive, and that all possible attribute values are included in the data.

a. Using each attribute as a node, sketch a reasonable *causal* Bayesian network for this domain. In doing this, assume that the patient has no other pre-existing disease (no lung, blood, neuromuscular or heart disease, etc.) and no previous history of trauma.

b. For the probability table associated with the node *Cancer*, explicitly state the best numeric estimates for *all* probabilities in the probability table. Be sure to clearly indicate which probability is which.

c. For the probability table associated with the node *Pollution*, explicitly state the best numeric estimates for *all* probabilities in the probability table. Be sure to clearly indicate which probability is which.

4. Learning Bayesian Networks Using Weka

You will be provided with a file *cancerData.arff* containing 286 examples of breast cancer patients, including whether the cancer recurs or not (the output class targets). Each example has 9 input features describing a patient. Further details are given as comments in the first part of the data file.

Separately train the following three classifiers using the default parameter settings in Weka unless indicated otherwise: *NaiveBayes*, *BayesNet K2 algorithm* with maximum number of parents set to 3 (set under searchAlgorithm), *and BayesNet TAN algorithm*. Use 5-fold cross validation in all cases, not the 10-fold default. Turn in the following in hardcopy format:

a. State the percent correct during testing for each of the three classifier methods as reported by Weka.

b. For the classifier built with K2 only, what is the network's computed probability that tumor size will be between 30-34, given that the patient experienced recurrence events?

c. Based solely on the data in (a), if you had to select one of the three classifier methods (naive Bayes, K2, or TAN) for a similar problem in the future, which would you use? Justify your answer in a sentence or two.

d. Examine the Bayesian network induced by the K2 and TAN algorithms. Describe the extent to which the links in these networks do or do not capture *causal* relations between the attributes. If they largely represent causal relations, explain in just a few sentences how the algorithm(s) identified such relations in selecting links. If they are largely a mixture of causal and non-causal relations, explain in just a few sentences what it is that the network's structure captures.

e. In what ways are the Bayesian networks generated by K2 and TAN quite similar, and in what ways do they differ significantly?

What do I turn in?

The hardcopy and electronic submissions are due at different times. The hard copy portion is due at the start of class Thursday April 11, the electronic submission is due earlier at 11:30 pm Wednesday April 10.

Hardcopy: Your answers to the questions in problems 1 - 4.

Electronic submission: For problem 4, turn in a single zip file that includes the full transcripts from your three Weka sessions (*NaiveBayes*, *BayesNet K2*, and *BayesNet TAN*; you can cut-and-paste these from the Weka Explorer session window). Use the Computer Science Department project submission server at https://submit.cs.umd.edu to submit this zip file.