COMP 307 — *Introduction to AI*

Assignment 3: Uncertainty and Probability

12% of Final Mark — Due: 23:59 Monday 29 May 2017

1 Question Description

Part 1: Reasoning Under Uncertainty Basics [10 marks]

X	P(X)
0	0.300
1	0.700

Y	X	P(Y|X)
0	0	0.300
1	0	0.700
0	1	0.800
1	1	0.200

Z	Y	P(Z|Y)
0	0	0.600
1	0	0.400
0	1	0.800
1	1	0.200

Problem Description

The above tables gives the prior distribution $P(X)$, and two conditional distributions $P(Y|X)$ and $P(Z|Y)$. It is also known that Z is independent from X given Y. All the three variables (X, Y, and Z) are binary variables. Compute the table of their joint distribution based on the chain rule.

You should show your *working process* of the calculation in the form like $P(A = 0|P(B = 1) = \frac{P(A=0,B=1)}{P(B=1)}$, to demonstrate that you *know how to calculate* them.

Requirements

1. Create the full joint probability table of X and Y, i.e. the table containing the following four joint probabilities $P(X = 0, Y = 0)$, $P(X = 0, Y = 1)$, $P(X = 1, Y = 0)$, $P(X = 1, Y = 1)$. Also explain which probability rules you used.

2. If given $P(X = 1, Y = 0, Z = 0) = 0.336$, $P(X = 0, Y = 1, Z = 0) = 0.168$, $P(X = 0, Y = 0, Z = 1) = 0.036$, and $P(X = 0, Y = 1, Z = 1) = 0.042$, create the full joint probability table of the three variables X, Y, and Z. Also explain which probability rules you used.

3. From the above joint probability table of X, Y, and Z:
   (i) calculate the probability of $P(Z = 0)$ and $P(X = 0, Z = 0)$,
   (ii) judge whether X and Z are independent to each other and explain why.

4. From the above joint probability table of X, Y, and Z:
   (i) calculate the probability of $P(X = 1, Y = 0|Z = 1)$,
   (ii) calculate the probability of $P(X = 0|Y = 0, Z = 0)$.

Part 2: Naive Bayes Method [25 marks]

This part is to implement the Naive Bayes algorithm, and evaluate the program on the spam data set to be described below. The program should build a Naive Bayes classifier from the labelled data set and apply it to the unlabelled set.

Problem Description

The labelled data set is in the file `spamLabelled.dat`, which describes 200 emails, labelled as *spam* or *non-spam*. Each email is specified by 12 binary attributes, indicating the presence of features such as "Viagra", "MILLION DOLLARS", significant amounts of text in CAPS, an invalid reply-to address, and so on. Note that there are 2^{12}=4096 possible input patterns, compared to a data set of just 200 examples.

The layout of the data is that each row is an instance of features from one email, and columns correspond to the features, which are binary: the feature is either there or not. The last (right-most) column is the class: 1 = spam, 0 = non-spam.

The file `spamUnlabelled.dat` contains 10 new input patterns to be classified.

There's a good entry in wikipedia (`http://en.wikipedia.org/wiki/Naive_Bayesian_classifier` that discusses exactly the domain we're applying the algorithm to. I recommend you read this article. You don't need to worry about taking logarithms though, unless you run into precision problems.

As we discussed during the lectures, zero probabilities are a problem for the Naive Bayes method. For example, if the training data did not include a $C = 1$ instance with attribute F8 = 1, the simplest version of the algorithm will assume that $P(C = 1|F8 = 1) = 0$, and never predict $C = 1$ if $F8 = 1$. This is generally a bad idea because $P(C = 1|F8 = 1)$ is unlikely to be exactly zero, even if it is very low. The simplest solution is to initialise all the counts to 1, rather than 0, which means every $P(C|F)$ has at least a low probability. As discussed in the lecture, you should divide by the right number when you convert the counts into probabilities.

Requirements

Your job is to use the Naive Bayes method to classify the unlabelled instances in the `spamUnlabelled.dat` file. The method should use the training data in `spamLabelled.dat` to construct the classifier (Naive Bayes probability tables), and then apply the classifier to the data in `spamUnlabelled.dat`

Your program should take two file names as command line arguments, construct a classifier from the data in the first file, and then apply the classifier to the data in the second file.

You may write the program code in `Java, C/C++`, or any other programming language.

You should submit the following files electronically and also a report in hard copy.

- (15 marks) `Program code` for your Naive Bayes Classifier (both the source code and the executable program running on ECS School machines),
- (2 marks) `sampleoutput.txt` containing the output of your program on the unlabelled data set, and
- (8 marks) `A report` in PDF, text or DOC format. The report should include:
    1. the probabilities ($P(F_i|c)$ for each feature i.
    2. For each input vector F in the unlabelled set, given the input vector F, the probability $P(S|D)$, the probability $P(\bar{S}|D)$, and the predicted class of the input vector. Here D is an email represented by F, S refers to class *spam* and $\bar{S}$ refers to class *non-spam*.
    3. The derivation of the Naive Bayes algorithm assumes that the attributes are conditionally independent. Why is this like to be an invalid assumption for the spam data? Discuss the possible effect of two attributes not being independent.

Part 3: Bayesian Networks [30 marks]

This part is to build a Bayesian Network and the problem/domain is described below

Problem Description

Dr. Rachel Nicholson is a Professor, who lives far away from her university. So, she prefer to work at home and she only comes to her office if she has research meetings with her postgraduate students, or teaching lectures for undergraduate students, or she has both meetings and teaching:
The probability for Rachel to have meetings is 70%, the probability of Rachel has lectures is 60%.
If Rachel has both meetings and lectures, the probability of Rachel comes to her office is 95%.
If Rachel only has meetings (without lectures), the probability of Rachel comes to her office is 75% because she can Skype with her students.
If Rachel only has lectures (without meetings), the probability of Rachel comes to her office is 80% because she can Skype with her students.
If Rachel has neither meetings nor lectures, there is a only 6% chance that she comes to the office.
When Rachel is in her office, half the time her light is off (when she is trying to hide from others to get work done quickly).
When she is not in her office, she leaves her light on only 2% of the time since the cleaners come for cleaning.
When Rachel is in her office, 80% of the time she logged onto the computer.
Because she sometimes work from home, 20% of the time she is not in her office, she is still logged onto the computer.

Requirements

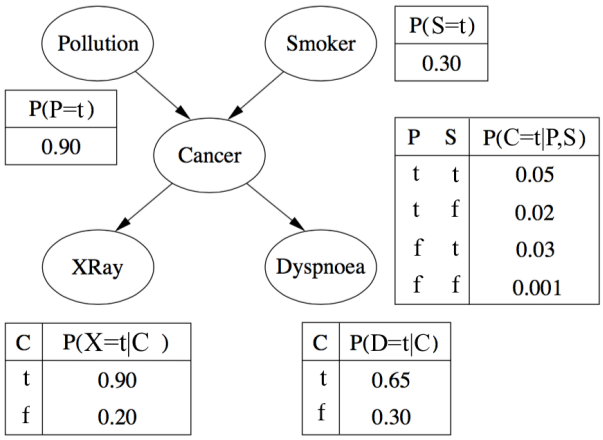
Note regarding the calculation, you should show your *working process* of the calculation to demonstrate that you *know how to calculate* them.

1. Construct a Bayesian network to represent the above scenario. (*Hint: First decide what your domain variables are; these will be your network nodes. Then decide what the causal relationships are between the domain variables and add directed arcs in the network from cause to effect. Finally, you have to add the prior probabilities for nodes without parents, and the conditional probabilities for nodes that have parents.*)
2. Calculate how many free parameters in your Bayesian network ?
3. What is the joint probability that Rachel has lectures, has no meetings, she is in her office and logged on her computer but with lights off.
4. Calculate the probability that Rachel is in the office.
5. If Rachel is in the office, what is the probability that she is logged on, but her light is off.
6. Suppose a student checks Rachel's login status and sees that she is logged on. What effect does this have on the students belief that Rachels light is on ?

Part 4: Inference in Bayesian Networks [35 marks]

Problem Description

The following Bayesian Network represents two causes and two effects related to Lung Cancer. Each variable takes the value true (t) or false (f). We will abbreviate the five variable names using their leading letters: P, S, C, X, and D. The probabilities shown are all for the "is true" outcome, e.g. read P(P=t) =0.90 as the probability that the variable Pollution takes the value true is 0.90. The probability that it is false is not shown, but is easily derived.



Requirements

Note regarding the calculation, you should show your *working process* of the calculation to demonstrate that you *know how to calculate* them.

1. Using *inference by enumeration* to calculate the probability $P(P = t|X = t)$ (i) describe what are the evidence, hidden and query variables in this inference, (ii) describe how would you use variable elimination in this inference, i.e. to perform the join operation and the elimination operation on which variables and in what order, and (iii) report the probability,
2. Given the Bayesian Network , find the variables that are independent to each other or conditionally independent given another variable. Find at least three pairs or groups of such variables.
3. If given the variable order as <Xray, Dyspnoea, Cancer, Smoker, Pollution>, draw a new Bayesian Network structure (nodes and connections only) to describe the same problem/domain as shown in the above given Bayesian Network. [hint: considering the above (conditionally) independent variables, the network should keep the original dependence between variables, which are that (conditionally) independent variables should remain being independent to each other, and dependent variables remain being dependent]. For each connection, explain why it is needed.

2 Relevant Data Files and Program Files

The relevant data files, information files about the data sets, and some utility programs files can be found from the following directory:

`/vol/comp307/assignment3/`

3 Assessment

We will endeavour to mark your work and return it to you as soon as possible, hopefully in 2 weeks. The tutor(s) will run a number of helpdesks to provide assistance.

4 Submission Guidelines

4.1 Submission Requirements

1. Programs for all individual parts. To avoid confusion, all the individual parts should use directories `part1/, part2/, ...` and all pieces of programs should be stored in their corresponding directories. Within each directory, please provide a `readme file` that specifies how to compile and run your program. A script file called `sampleoutput.txt` should also be provided to show how your program run properly. If you programs cannot run properly, you should provide a `buglist` file.

2. A document that consisting of the report of all the individual parts. The document should mark each part clearly. The document can be written in PDF, text or the DOC format.

4.2 Submission Method

The programs and the PDF/Text/DOC version of the document should be submitted through web submission system from the COMP307 course web site **by the due time**.

The hard copy of the document is required to submit to the COMP 307 handin box in the 2nd floor corridor of the Cotton building by the due time.

4.3 Late Penalties

All assignments must be submitted on time unless you have made a prior arrangement with the lecturer or have a valid medical excuse (for minor illnesses it is sufficient to discuss this with the lecturer.) Assignments that are handed in late without prior arrangement will be marked if time permits, but the mark may not contribute to your final grade.