COMP 307/AIML 420 — *Introduction to AI*

Assignment 3: Uncertainty and Probability

10% of Final Mark — Due: 23:59 Sunday 16 May 2021

1 Question Description

In the following, unless explicitly specify, a *capital letter* (e.g., A, B, X, Y) represents a *random variable*, and a *lowercase letter* (e.g., a, b, x, y) represents a value.

Part 1: Reasoning Under Uncertainty Basics [30 marks]

This part contains several questions about the basics of reasoning under uncertainty. You need to write your answers to each of these questions in your report, and **Show your working.**

For calculations, you need to show the steps in the form like $P(A = 0|B = 1) = \frac{P(A=0,B=1)}{P(B=1)}$, to demonstrate that you *know how to calculate* them.

For proving, you also need to clearly show each step of the proof.

Question 1 [10 marks]

The tables below give 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.

X	$P(X)$
0	0.35
1	0.65

X	Y	$P(Y\|X)$
0	0	0.10
0	1	0.90
1	0	0.60
1	1	0.40

Y	Z	$P(Z\|Y)$
0	0	0.70
0	1	0.30
1	0	0.20
1	1	0.80

1. Compute the table of the joint distribution $P(X, Y, Z)$. **Show the rule(s) you used, and the steps of calculating each joint probability**.
2. Create the full joint probability table of X and Y, i.e., the table containing the following four joint probabilities $P(X = 0, Y = 0)$, $P(X = 0, Y = 1)$, $P(X = 1, Y = 0)$, $P(X = 1, Y = 1)$. **Show the rule(s) used, and the steps of calculating each joint probability**.
3. From the above joint probability table of X, Y, and Z, calculate the following probabilities. **Show your working**.
   (a) $P(Z = 0)$,
   (b) $P(X = 0, Z = 0)$,
   (c) $P(X = 1, Y = 0|Z = 1)$,
   (d) $P(X = 0|Y = 0, Z = 0)$.

Question 2 [10 marks]

Consider three Boolean variables A, B, and C, $P(B) = 0.7$, $P(C) = 0.4$, $P(A|B) = 0.3$, $P(A|C) = 0.5$, and $P(B|C) = 0.2$, we also know that A is independent from B given C. Calculate the following probabilities. **Show your working.**

(i) $P(B, C)$

(ii) $P(\neg A|B)$

(iii) $P(A, B|C)$

(iv) $P(A|B, C)$

(v) $P(A, B, C)$

Question 3 [10 marks]

Dragonfly has a rare species, which always has an extra set of wings. However, common dragonflies can sometimes mutate and get an extra set of wings. A dragonfly either belongs to the common species or the rare species with the extra wings. There are 0.3% dragonflies belonging to the rare species with the extra set of wings. For the common dragonflies, the probability of the extra-wing mutation is 0.1%. Now you see a dragonfly with an extra pair of wings. What is the probability that it belongs to the rare species? **Show your working.**

Question 4 [for AIML420 ONLY, 10 marks]

Prove the following statements. **Show your working.**

(i) If $P(A|B, C) = P(B|A, C)$, then $P(A|C) = P(B|C)$

(ii) If $P(A|B, C) = P(A)$, then $P(B, C|A) = P(B, C)$

(iii) If $P(A, B|C) = P(A|C) * P(B|C)$, then $P(A|B, C) = P(A|C)$

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. You are recommended to read this article.

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

需要加两个email, 一个att全是0, 一个att全是1
,一个Class=spam, 一个NOSpam

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`

You should implement the Naive Bayes method from scratch (not call it from any machine learning library). 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.

- (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 conditional probabilities $P(F_i|C)$ for each feature i and each class label.
  2. For each instance in the unlabelled set, given the input vector $F = (f_1, f_2, \ldots, f_{12})$, the probability $P(C = 1, F)$ (enumerator of $P(C = 1|F)$, *score* of spam), the probability $P(C = 0, F)$ (enumerator of $P(C = 0|F)$, *score* of non-spam), and the predicted class of the input vector.
  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 conditionally independent.

Part 3: Building Bayesian Network [30 marks]

This part is to build a Bayesian Network for the problem 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%.
- 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.

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

Requirements

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 student's belief that Rachel's light is on ?

Part 4: Inference in Bayesian Networks [35 marks]

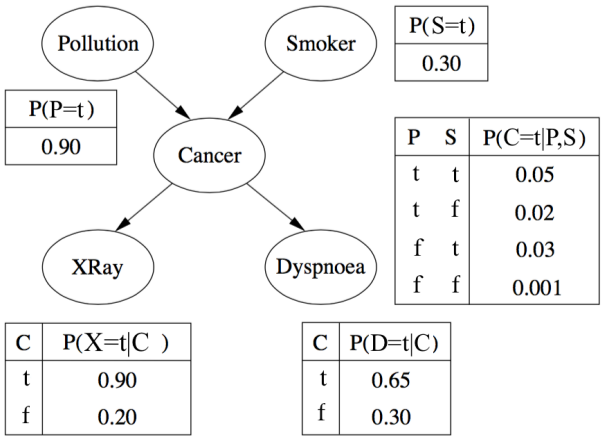
Problem Description

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 of 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 of each other, and dependent variables remain being dependent]. For each connection, explain why it is needed.

Part 5: Bayesian Network: Applications [For AIML420 ONLY, 20 marks]

Identify a real-world application (**different from the examples given in this assignment and the lectures**) that can be described using Bayesian network. There should be at least 5 random variables in this Bayesian network.

In your report, you should:

(i) Clearly define the random variables and their domains.

(ii) Clearly describe their relationships (using plain language).

(iii) Draw the Bayesian network that can reflect the described relationship.

(iv) Write the factorisation of the Bayesian network.

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 (**Executive program file and source files**) for Part 2. 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 report document that consist of **the answers of all the individual parts**. The document should mark each part clearly. The document can be written in PDF, text or the DOC format.

3. For drawing the diagram such as the Bayesian network, you need to **make the diagram very clear to be marked**. We highly recommend using drawing tools rather than by hand.

4.2 Submission Method

The programs and the PDF report should be submitted through the web submission system (accessible from the COMP307 or AIML420 course web site) **by the due time**. Please ensure you submit to the correct system based on which course you are enrolled in!

4.3 Late Penalties

The assignment must be submitted on time unless you have made a prior arrangement with the **course coordinator** or have a valid medical excuse (for minor illnesses it is sufficient to discuss this with the course co-ordinator.) The penalty for assignments that are handed in late without prior arrangement is one grade reduction per day. Assignments that are more than one week late will not be marked.

Remember that you have three late days for this course (which can be used fractionally), but that these apply across the whole course, not per-assignment! Please save some late days in case you need them later on!