Assignment 1

COMP9418 – Advanced Topics in Statistical Machine Learning

Lecturer: Gustavo Batista

Last revision: Tuesday 24$^{\text{th}}$ September, 2019 at 12:52

Instructions

Submission deadline: Friday, October 11th, 2019, at 21:00:00.

Late Submission Policy: The penalty is set at 20% per late day. This is ceiling penalty, so if a group is marked 60/100 and they submitted two days late, they still get 60/100.

Form of Submission: This is a group assignment. Each group can have up to three students. Write the names and zIDs of each student in both the report and Jupyter notebook. **Only one member of the group should submit the assignment**.

The group should submit your solution in one single file in zip format with the name solution.zip. There is a maximum file size cap of 5MB, so make sure your submission does not exceed this size. The zip file should contain one Jupyter notebook file and one pdf file. The Jupyter notebook should contain all your source code. Use markdown text to organise and explain your implementation. The pdf file is a 2-page report summarising your findings. The report can include text and plots to illustrate your results.

Submit your files using give. On a CSE Linux machine, type the following on the command-line:

```
$ give cs9418 ass1 solution.zip
```

Alternative, you can submit your solution via the course website.

Recall the guidance regarding plagiarism in the course introduction: this applies to this homework, and if evidence of plagiarism is detected, it may result in penalties ranging from loss of marks to suspension.

The dataset and breast cancer domain description in the Background section are from the assignment developed by Peter Lucas, Institute for Computing and Information Sciences, Radboud Universiteit.

Introduction

In this assignment, you will develop some sub-routines in Python to create useful operations on Bayesian Networks. You will implement an efficient independence test, learn parameters from data, sample from the joint distribution and classify examples.

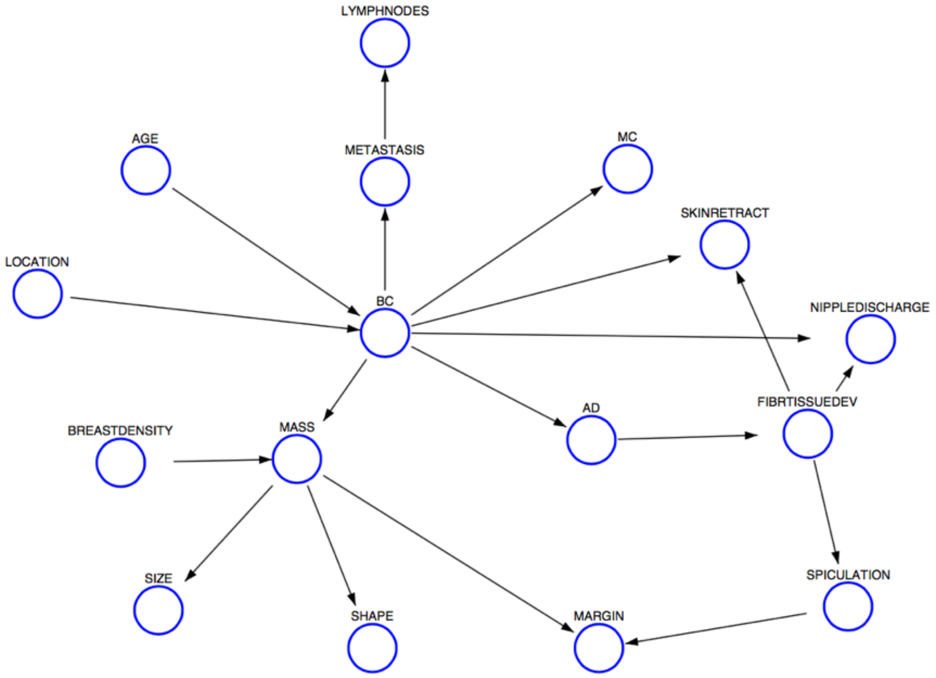
We will use a Bayesian Network for diagnosis of breast cancer. We start with some background information about the problem.

Background

Breast cancer is the most common form of cancer and the second leading cause of cancer death in women. Every 1 out of 9 women will develop breast cancer in her lifetime. Although it is not possible to say what exactly causes breast cancer, some factors may increase or change the risk for the development of breast cancer. These include age, genetic predisposition, history of breast cancer, breast density and lifestyle factors. Age, for example, is the most significant risk factor for non-hereditary breast cancer: women with age of 50 or older have a higher chance of developing breast cancer than younger women. Presence of BRCA1/2 genes leads to an increased risk of developing breast cancer irrespective of other risk factors. Furthermore, breast characteristics, such as high breast density are determining factors for breast cancer.

The main technique used currently for detection of breast cancer is mammography, an X-ray image of the breast. It is based on the differential absorption of X-rays between the various tissue components of the breast such as fat, connective tissue, tumour tissue and calcifications. On a mammogram, radiologists can recognise breast cancer by the presence of a focal mass, architectural distortion or microcalcifications. Masses are localised findings, generally asymmetrical in relation to the other breast, distinct from the surrounding tissues. Masses on a mammogram are characterised by several features, which help distinguish between malignant and benign (non-cancerous) masses, such as size, margin, shape. For example, a mass with irregular shape and ill-defined margin is highly suspicious for cancer, whereas a mass with round shape and well-defined margin is likely to be benign. Architectural distortion is focal disruption of the normal breast tissue pattern, which appears on a mammogram as a distortion in which surrounding breast tissues appear to be "pulled inward" into a focal point, often leading to spiculation (star-like structures). Microcalcifications are tiny bits of calcium, which may show up in clusters, or in patterns (like circles or lines) and are associated with extra cell activity in breast tissue. They can also be benign or malignant. It is also known that most of the cancers are located in the upper outer quadrant of the breast. Finally, breast cancer is characterised by several physical symptoms: nipple discharge, skin retraction, palpable lump.

Breast cancer develops in stages. The early stage is referred to as in situ ("in place"), meaning that cancer remains confined to its original location. When it has invaded the surrounding fatty tissue and possibly has spread to other organs or the lymph, so-called metastasis, it is referred to as invasive cancer. It is known that early detection of breast cancer can help improve the survival rates.



[25 Marks] Task 1 – Efficient d-separation test

In this part of the assignment, you will implement an efficient version of the d-separation algorithm. Let us start with a definition for d-separation:

Definition. Let **X**, **Y** and **Z** be disjoint sets of nodes in a DAG G. We will say that **X** and **Y** are d-separated by **Z**, written dsep(**X**,**Z**,**Y**), iff every path between a node in **X** and a node in **Y** is blocked by **Z** where a path is blocked by **Z** iff there is at least one inactive triple on the path.

This definition of d-separation considers all paths connecting a node in X with a node in Y. The number of such paths can be exponential. The following algorithm provides a more efficient implementation of the test that does not require enumerating all paths.

Algorithm. Testing whether $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in a DAG G is equivalent to testing whether $\mathbf{X}$ and $\mathbf{Y}$ are disconnected in a new DAG G', which is obtained by pruning DAG G as follows:

1. We delete any leaf node W from DAG G as long as W does not belong to $X \cup Y \cup Z$. This process is repeated until no more nodes can be deleted.
2. We delete all edges outgoing from nodes in $\mathbf{Z}$.

Implement the efficient version of the d-separation algorithm in a function `d_separation(G,X,Y,Z)` that return a boolean: true if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ and false otherwise. Comment about the time complexity of this procedure.

[05 Marks] Task 2 – Estimate Bayesian Network parameters from data

Estimating the parameters of a Bayesian Network is a relatively simple task if we have complete data. The file `bc.csv` has 20,000 complete instances, *i.e.*, without missing values. The task is to estimate and store the conditional probability tables for each node of the graph. As we will see in more details in the Naive Bayes and Bayesian Network learning lectures, the Maximum Likelihood Estimate (MLE) for those probabilities are simply the empirical probabilities (counts) obtained from data.

Implement a function `learn_bayes_net(G, file, outcomeSpace, prob_tables)` that learns the parameters of the Bayesian Network G. This function should output a dictionary `prob_tables` with the all conditional probability tables (one for each node), as well as the outcomeSpace with the variables domain values.

We are working with a small Bayesian Network with 16 nodes. What will be the size of the joint distribution with all 16 variables?

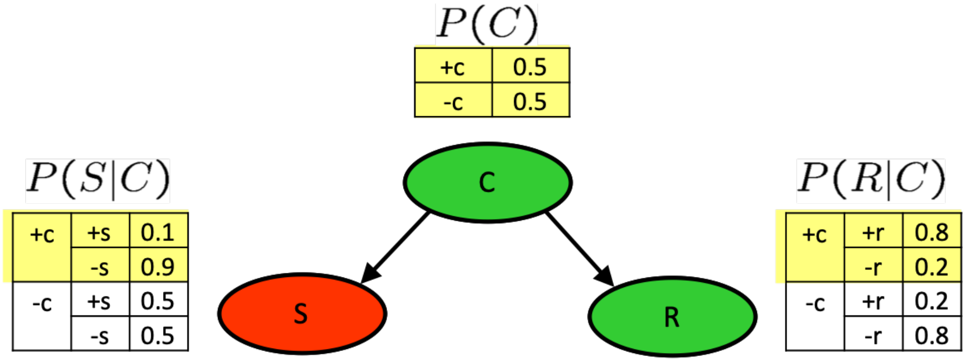
As we have implemented most of this function in the tutorials, Task 2 has a value of 5 marks.

[25 Marks] Task 3 – Sampling

We can sample a Bayesian Network to create instances according to the joint distribution. This procedure has many applications; one of them is to answer probabilistic queries in an efficient but approximated way.

A simple sampling procedure is known as *forward* or *ancestral* sampling. It consists of traversing the graph G in topological ordering. We use a random number generator to draw a value of each variable X according to $P(X|Parents(X))$.

The next figure illustrates this idea. One possible topological order for this graph is C, S and R. We can use a random number generator to draw a number between 0 and 1. For the node C, we use a cutoff 0.5. If the number is less than the cutoff then `C = +c`, otherwise `C = -c`. Let us suppose the random number is 0.3 and, therefore, we take `C = +c`. We continue in topological ordering and sample value for the variable S according to $P(S| + c)$. The cutoff is now 0.1. Suppose the random number generator returns 0.7. Thus, we assign the value `-s` to S. We continue to variable R and sample value according to $P(R| + c)$ leading to a cutoff of 0.8. We use the random number generator again and sample `R = +r`. In the end, the generated sample is `+c, -s, +r`. We can repeat this process to generate more instances.



$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

Use forward sampling to generate 1000 samples from the Breast Cancer Bayesian Network. Comment about the time complexity of the procedure and accuracy of the estimates. What happens as you add more observed variables in the query in terms of accuracy and effective sample size?

[25 Marks] Task 4 – Classification

This particular Bayesian Network has a variable that plays a central role in the analysis. The variable *BC* (Brest Cancer) can assume the values `No`, `Invasive` and `InSitu`. Accurately identifying its correct value would lead to an automatic system that could help in early breast cancer diagnosis.

Use the Bayesian Network to classify cases of the dataset. Propose an experimental setup to estimate the classification error. Compare the classification error of the Bayesian Network with your favourite Machine Learning classifier.

[20 Marks] Task 5 – Report

Write a two-page report (around 1000 words) summarising your findings in this assignment. Some suggestions for the report are:

- Which were the main challenges and how you solved these issues?
- Answer the questions of each task.
- Discuss the complexity of the implemented algorithms.
- Include plots to illustrate your results.