Assignment 2 - Learning Bayesian networks with pyAgrum

INTRODUCTION

The purpose of this assignment is to test and possibly expand your knowledge about learning Bayesian networks from data. Recall that learning Bayesian networks involves both structure learning, i.e., learning the graph topology from data, and parameter learning, i.e., learning the actual, local probability distributions from data.
There are basically two approaches to structure learning: (i) search-and-score structure learning, and (ii) constraint-based structure learning. Search-and-score algorithms search for a Bayesian network structure that fits the data best (in some sense). They start with an initial network structure (often a graph without arcs or a complete graph), and then traverse the search space of network structures. Constraint-based algorithms carry out a conditional (in)dependence analysis on the data. Based on this analysis an undirected graph is generated (to be interpreted as a Markov network). Using additional independence tests, some of the arcs can be directed. Typically, the result is a partially directed graph (a CPDAG, also called an *essential graph*).

A NOTE ON EVALUATING CLASSIFIERS

In this assignment, you are, among others, asked to evaluate a classifier, which is one of the ways that you can use a Bayesian network. Basically, classification is mapping the evidence from the case to a case label or category, and there are many other methods, such as classification tree and rules, and neural network, that can also be used for this purpose. This will be discussed more extensively in the course on Machine Learning.
Given a Bayesian network with joint probability distribution P and variable C of interest (often called the 'class variable'), the problem is to determine a value of C with maximum probability, or:

$$c_{\max} = \arg\max_{c} P(c \mid \mathcal{E})$$

with $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$ being the evidence (measurements, observations) for the case at hand.
For this assignment, it suffices to know that a popular way to evaluate the quality of classifiers is ROC (Receiver Operating Characteristic) analysis. This technique was originally developed in the UK during World War II to help British pilots in making a distinction between enemy (German) aeroplanes and their own aeroplanes: too many British pilots attacked their colleagues. For a brief introduction, we refer to the book *Artificial Intelligence: Foundations of Computational Agents* by Poole and Mackworth, in particular section 7.2.2 which you can find online at `https://artint.info/2e/html/ArtInt2e.Ch7.S2.SS2.html`.

Task 2.1: Trying the learning algorithms in pyAgrum

We start from the pyAgrum setup obtained during Assignment 1. We refer to the description of that assignment for instructions on how to start a pyAgrum Docker image. Use the pyAgrum tutorials at https://webia.lip6.fr/~phw//aGrUM/docs/last/notebooks/ (in particular, *Structural Learning*, *Learning BN as probabilistic classifier*, and *Learning Essential Graphs*) to see which learning algorithms are available, and how to estimate the quality of their performances. For example, how to practically compute the area under the ROC curve in pyAgrum is shown in the tutorial on *Learning BN as probabilistic classifier*. We encourage you to play with the code of the tutorials and its input data, in order to understand how the algorithms work, and how the function calls need to be made in pyAgrum.

You do not have to submit any of your results from this first task.

Task 2.2: Experiment with learning algorithms

From a scientific point of view, it is interesting to investigate the performance of various learning algorithms. In this assignment, you are asked to investigate some of these learning algorithms with respect to how well they can recover the true structure and how this impacts the classification performance. You will make use of two case studies:

1. **A1 network**: The Bayesian network you developed in Assignment 1 (to compare structure learning algorithms).
2. **Breast cancer**: You can make use of clinical data of patients with breast cancer disease (to investigate the relationship between learning structure and classification performance). See below for an elaboration of this dataset and some background.

BREAST CANCER DATABASE

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 life time. Every year in The Netherlands 13.000 women are diagnosed with breast cancer. 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 greatest risk factor for non-hereditary breast cancer: women with age of 50 or older has a higher chance for 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 recognize 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 a number of characteristics, 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, leading often 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 a number of physical symptoms: nipple discharge, skin retraction, palpable lump.

Breast cancer develops in stages. The early stage is referred as *in situ* ('in place'), meaning that the cancer remains confined to its original location. When it has invaded the surrounding fatty tissue and possibly has spread to other organs or the lymphs, 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. Computerized techniques appear to assist medical experts in this respect. Bayesian networks are especially useful given the uncertainty and complexity in mammographic analysis. Figure 1.1 presents a causal model for breast cancer diagnosis based on the knowledge presented above. All the nodes are assumed to be discrete and the values for each variable are given in Table 1.1.

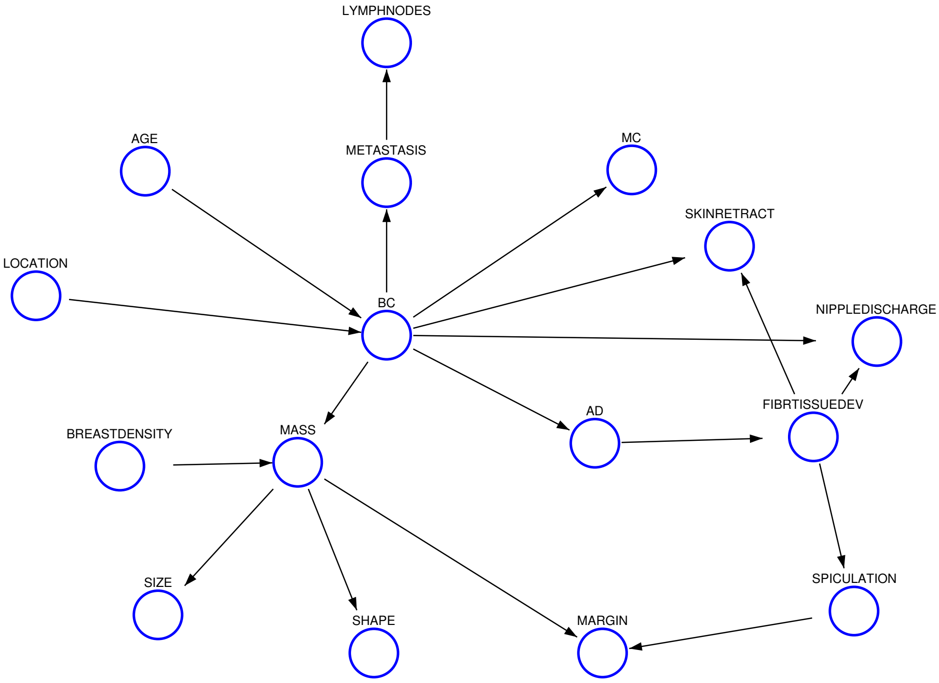

FIGURE 1.1 Bayesian network for breast cancer diagnosis.

TABLE 1.1 Definition of the variables from the breast cancer database

Node	Values
Age	$< 35, 35 - 39, 50 - 74, > 75$ (in years)
Location	UpOut, UpIn, LowOut, LowIn (in quadrants)
Breast Cancer	No, Invasive, InSitu
Density	Low, Medium, High
Mass	No, Malignant, Benign
Size	$< 1, 1 - 3, > 3$ (in cm)
Shape	Other, Round, Oval, Irregular
Margin	Ill-defined, Well-defined
Architectural Distortion	Yes, No
Fibrous Tissue Development	Yes, No
Skin Retraction	Yes, No
Nipple Discharge	Yes, No
Microcalcification	Yes, No
Metastasis	Yes, No
Lymph Nodes	Yes, No
Spiculation	Yes, No

WHAT DO YOU HAVE TO DO?

Investigating the effect of sample size of learning structure

For this part of the assignment you have to compare results obtained by a search-and-score- and a constraint-based learning algorithm for your own network.

⇒ *The following is requested from you:*

(1) *Generate data from your own network (say 1000 samples) and use these data to generate network structures. Investigate what the effect of size of the dataset is on the structure of the resulting Bayesian network (eg. generate networks using 100, 500 and 1000 samples). Do this for both classes of learning algorithms: select MIIC as a constraint-based algorithm and greedy hill climbing algorithm as a score-based algorithm. Evaluate the networks in terms of how well it learns the original Bayesian network in terms of structure (e.g. by a manual comparison or a visualisation with differences).*

Learning a classifier

The breast cancer network was constructed based on expert knowledge. Here you will compare this Bayesian network with a Bayesian network learned purely from data for classification purposes.

⇒ *The following is requested:*

(2) *Learn a Bayesian network from the available breast cancer data using the greedy hill climbing algorithm. Compare the learnt work to the manually constructed network in terms of structure (e.g. the structural Hamming distance).*

(3) *The breast cancer dataset includes the variable 'Breast Cancer' that is used to classify patients into one of three different categories:*

1: No *breast cancer*
2: Invasive *breast cancer*
3: Insitu *breast cancer*

*Compare the classification performance between the manually constructed and learnt network in terms of area under the ROC curve. It is sufficient for this assignment to treat it as a binary classifier (*no breast cancer *vs* breast cancer*).*

(4) *Investigate if you can steer the learning algorithm closer to the expert (manually constructed) network, by adding structural constraints. How does this impact the classification performance? (Note: one experiment is sufficient!)*

Task 2.3: Write a brief report (in Dutch or English)

CONTENT OF THE REPORT

Your report should consist of:

- Introduction (Context and a brief description of the aim of the two studies. What is investigated?)
- Methods (How are the investigations executed?)
- Results (What are your findings?)
- Conclusions and discussion (Your conclusions and some reflection on the results. Eg. If you observe unexpected results, can you think of an explanation?)

You can improve upon the structure of the report according to your own taste, but try to remain concise and not exceed a total of 2000 words (about 4 pages). You may add extra pages for an appendix with relevant figures and/or code.