Computational Intelligence 2008–2009
Practical Assignment II

1 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:

- search-and-score structure learning, and
- 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 by in each step either removing an arc, adding an arc, or reversing an arc. Read again the paper by Castello and Kočka [1] for a good overview of the principles and difficulties associated with this learning method. Recent search-and-score-algorithms take Markov equivalence into account, i.e., they search in the space of equivalence classes of Bayesian networks and the scoring method they use give the same score for equivalent networks. Bayesian networks with different graph topologies that are included in the same Markov equivalence class represent exactly the same conditional-independence information by d-separation. Examples of search-and-score algorithms are K2 and inclusion-driven learning. They usually are based on hill-climbing (greedy) search.

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, this network is converted into a Bayesian network. Constraint-based learning algorithms allow for the easy incorporation of *background knowledge*, i.e., prior knowledge on dependences or independences that hold for the domain under consideration. Examples of constraint-based learning algorithms are PC, NPC, grow-shrink, and incremental association. A good paper discussing the difficulties with constraint-based methods is the paper by Chickering and Meek [2].

For the underlying theory, consult the book "Bayesian Artificial Intelligence" [3] and the two paper referred to above, which can be downloaded from the CI seminar website. Also consult more specialised books and papers if required.

Below, you will find descriptions of a number of tasks that you need to perform. You are free to do something different in the context of Bayesian-network structure learning, e.g., to investigate learning *dynamic* Bayesian networks from temporal data, but it is required to obtain approval from the lecturer in that case.

2 Software

There are two environments for machine learning that can be used for this assignment:

- Matlab: a general purpose computing environment for applied mathematics;
- R: a general purpose computing environment for applied statistics.

Both environment are similar in the sense that they offer a script language (M versus R) to write programs in, and that they support an interactive style of programming. R is more specialised than Matlab as its focus is on statistics. A further difference is that Matlab is commercial software and R is in the public domain. As a consequence there is a slight preference for R. However, based on previous experience you may be tempted to use Matlab. Both environment are available for all operating systems. However, they support the sort of interaction typical for Unix and Linux, so it is more natural to use these packages under Unix (e.g. a Mac) and Linux.

Matlab is installed on all FNWI computers; it can be freely obtained by students for Unix, Linux and Windows from the CN&CZ department. R can be downloaded from

http://www.r-project.org/

Note that there are binary versions available for many different versions of operating systems. (However, installing the R system from the source code is not hard.) R includes extensive introductory documentation to get you started.

Both Matlab and R do not offer Bayesian network learning as default libraries, and you, thus, need to load special purpose Bayesian network learning packages for both systems:

- We advise you to use the following Bayesian network learning packages for R: bnlearn[1] and deal[2];
- For Matlab you can download the Bayes Net Toolbox (BNT)[3].

Remark: *Learning Bayesian networks from data is far from easy! Learning Bayesian networks is an active area of research and most software that is available is experimental in nature. Thus, do not expect perfect software packages in this area and be prepared to experiment with the software. Furthermore, the learning problem suffers from combinatorial explosion; so, do not expect that each database made available can be handled easily.*

3 Datasets

Of course, learning is impossible without the availability of data. On the CI website, three different datasets are made available that can be downloaded:

- **Alarm database or dataset 1** (large dataset): physiological data from patients under anaesthesia (See Section 4 of Practical I for an explanation);

[1] Download at: http://cran.r-project.org/web/packages/bnlearn/
[2] Download at: http://cran.r-project.org/web/packages/deal/index.html
[3] Download at: http://www.cs.ru.nl/~peterl/bntnew.tar.gz

- **NHL database or dataset 2** (small dataset): clinical data of patients with non-Hodgkin lymphoma (NHL) of the stomach (See Section 4 of Practical I for an explanation);

- **Breast cancer database or dataset 3** (large dataset): clinical data of patients with breast cancer disease (See below for an explanation).

Both dataset 1 and 3 have been generated from existing Bayesian networks using sampling and are, thus, artificial. These datasets are large and contain complete data. Dataset 2 is a real dataset obtained from medical researchers and includes missing values (denoted by 'NA', i.e., 'Not Available'). For the three datasets there are Bayesian networks available designed by experts, that can be used for comparison. These networks are given below.

4 Diagnosis of breast cancer

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 11.000 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 presents a causal model for breast cancer diagnosis based on the knowledge presented the previous section. All the nodes are assumed to be discrete and, in terms of the values for each variable, we have:

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

5 What do you have to do?

In the next sections, it is described what is expected from you. The results you obtain should be described and explained in a comprehensive report, that will be assessed as part of the overall assessment for the course.

5.1 Comparison of two learning algorithms

For this part of the assignment you have to compare results obtained by a search-and-score- and a constraint-based learning algorithm for dataset 1 (alarm).

$\Rightarrow$ *The following is requested from you:*

*(1) Use dataset 1 (the **alarm database**) in order to investigate what the effect of size of the dataset is on the structure of the resulting Bayesian network. Do this for both classes of learning algorithms. Describe in the report (see below) what you observe.*

5.2 Comparison with manually constructed Bayesian networks

For the three datasets, there are Bayesian networks available that have been manually constructed. The alarm network is shown in Figure 2, the network of non-Hodgkin lymphoma of the stomach is shown in Figure 3, and the network of the breast cancer is shown in Figure 1.

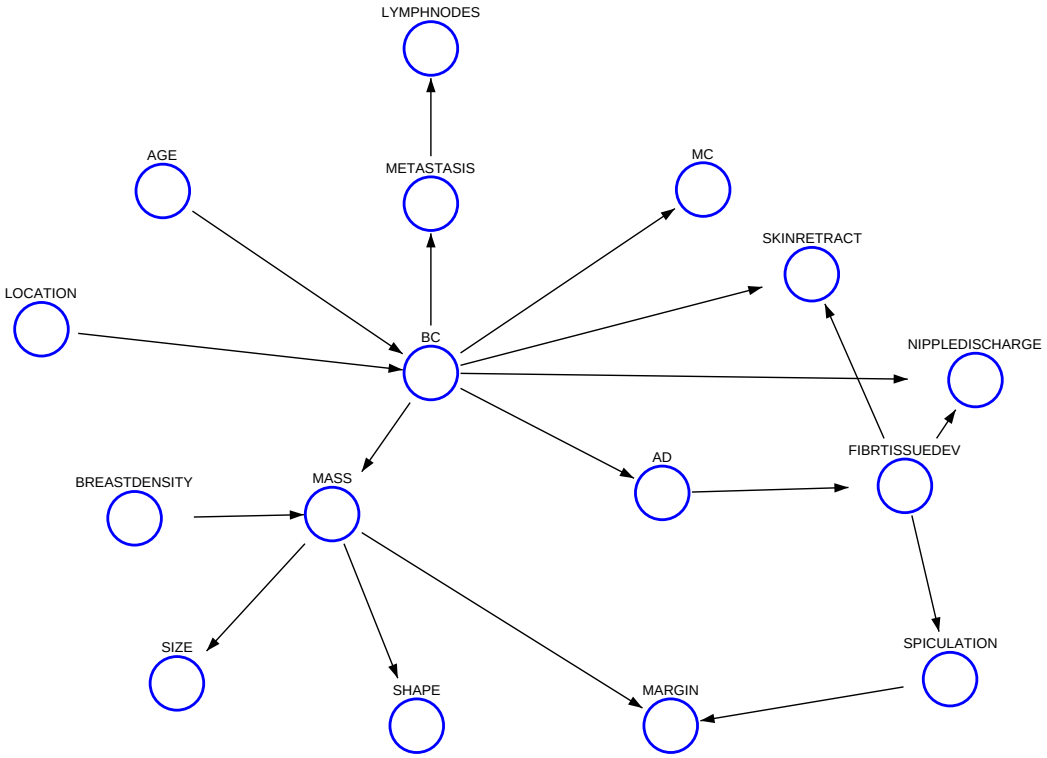

Figure 1: Bayesian network for breast cancer diagnosis.

In all cases it is necessary and possible to compare the structure of the learnt Bayesian networks with the structure that was constructed manually.

⇒ *The following is requested:*

(2) Define measures that can be used to determine the quality of a learning algorithm in terms of a known Bayesian network structure. Motivate the definition of these measures.

(3) Use these measures in order to evaluate the quality of the two learning algorithms.

For the network of non-Hodgkin lymphoma of the stomach we also have the marginal probability distributions (See Figure 3).

⇒ (4) *Compare the marginal probability distributions of the learnt NHL Bayesian network with those of the manually constructed network.*

5.3 Diagnostic Bayesian network structures

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

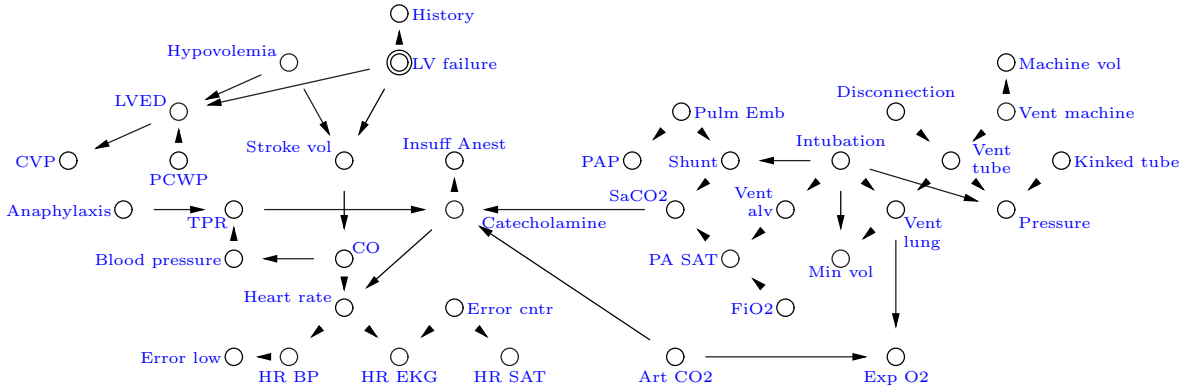

Figure 2: Alarm Bayesian network.

As the class variable acts as the *output* of a Bayesian network that models the random variables contained in the dataset, where the other variables represent the input (also called evidence or features), if present, a diagnostic Bayesian network has often a *special* structure. Popular special Bayesian-network structures for diagnostic applications are naive Bayesian networks (NBN) and tree-augmented Bayesian networks (TAN) (see Lecture 3 "Building Bayesian Network" for examples on these types of networks).

⇒ *The following is requested:*

(5) Learn a Bayesian network from the breast cancer dataset using a search-and-score or constraint-based algorithm.

(6) Develop a special purpose (e.g., NBN or TAN) Bayesian network (either by hand or automatically if the software offers such special-purpose learning algorithms).

(7) Compare the Bayesian network obtained by step (5) and the special purpose Bayesian network from (6) with the manually constructed network in Figure 1 in terms of network structure and accuracy, e.g., using measures such as misclassification error and area under the ROC curve.

6 Results

Write a report in which you motivate your choices (of software, algorithms and measures) and where you discuss the results you have obtained for parts 1 to 7.

References

[1] R. Castelo and T. Kočka, On inclusion-driven learning of Bayesian networks, *Journal of Machine Learning Research*, 527–574, 2003.

[2] D.M. Chickering and C. Meek, On the incompatibility of faithfulness and monotone DAG faithfulness. *Artificial Intelligence*, vol. 170, 653-666, 2006.

[3] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*, Chapman & Hall: Boca Raton, 2004.

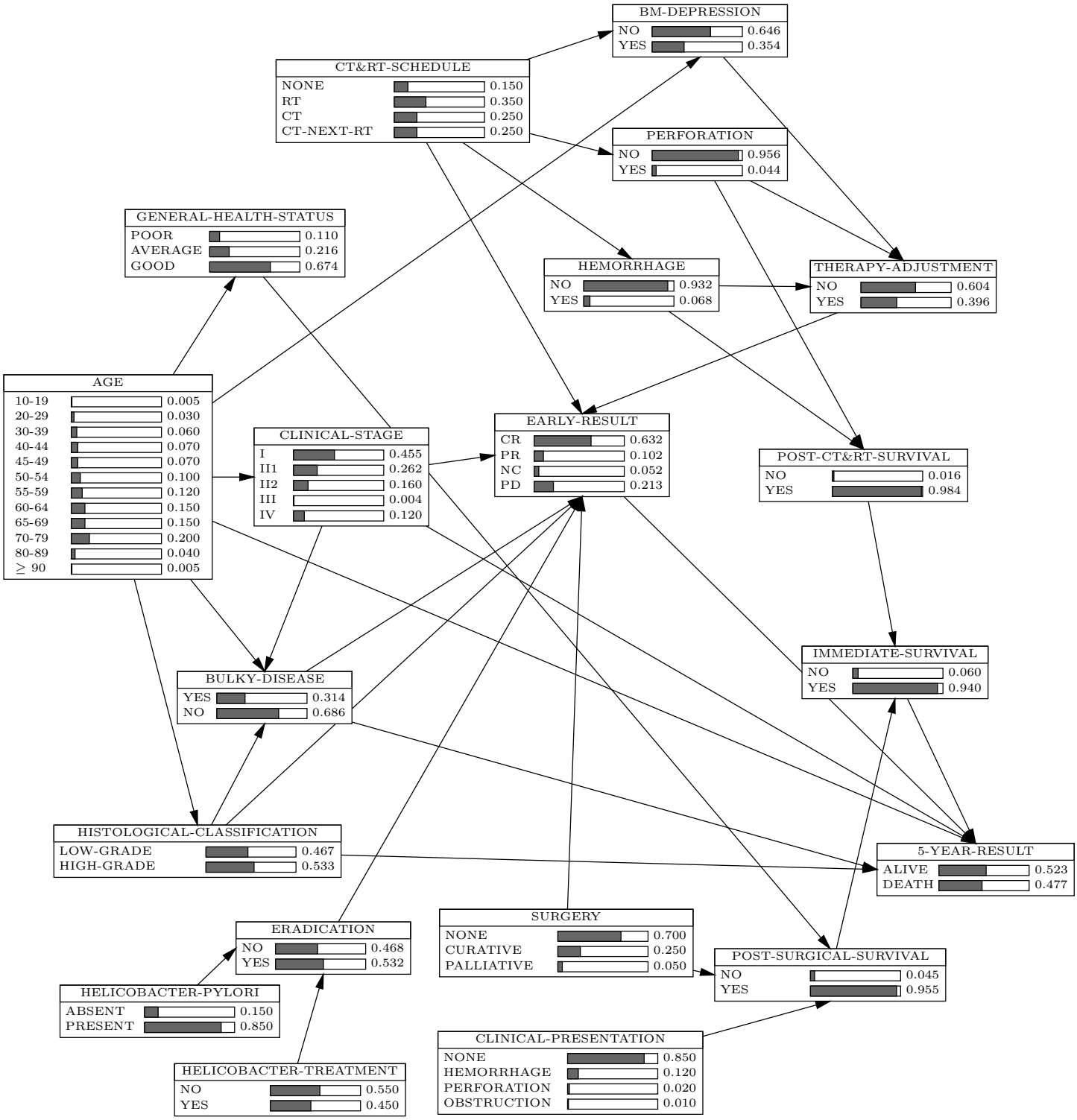

Figure 3: Bayesian network with prior probability distributions for gastric NHL.