Lab : Bayesian networks — 3h

Cédric Buche

Objectives

1. Create a Bayesian network representing an expertise and define the probabilities.
2. Exploit the Bayesian network with Bayesian inference.
3. Use machine learning methods to automatically generate a Bayesian network from data.

Library

We use pgmpy, a python library for working with Probabilistic Graphical Models : `https://github.com/pgmpy/pgmpy`.

Documentation and list of algorithms supported : `http://pgmpy.org/`

Contents

1 Create a Bayesian Network

Consider the case of a lung disease specialist wishing to model his knowledge (very simplified) on the diagnosis of cancer or tuberculosis.

1.1 Create Nodes and Arcs

The development of a Bayesian network requires prior identification of the different variables needed to describe the domain. Each of these variables then corresponds to a node in the graph.

Our specialist starts by placing on his worksheet three numbers representing the patient's age (`Age`), whether the patient smokes (`Smoker`), and cancer (`Cancer`). Then he defines the probabilistic relations between these three variables:

- ▷ Age has a direct influence on smoking and cancer,
- ▷ Smoking has a direct influence on cancer.

Our specialist then adds the variable `Tuberculosis`, and another variable `TbOuCa` to perform a " or logical " between the `Tuberculosis` and the `Cancer`. This is not necessary, but it will simplify the graph later because of the symptoms common to both diseases.

Continue this work by placing the nodes :

- ▷ `VisitAsia`,
- ▷ `Bronchitis`,
- ▷ `RespiratoryDifficulty`,
- ▷ `Radiography` .

Then put the links :

- ▷ `VisitAsia` to `Tuberculosis`,
- ▷ `TbOuCa` to `Radigraphy`,
- ▷ `TbOuCa` to `RespiratoryDifficulty`,
- ▷ `Bronchitis` to `RespiratoryDifficulty` .

1.2 Type of variables

We need to define the type of variables for each node.

Set the variable `Age` according to:

- ▷ Young (15 - 30 years old)
- ▷ Adult (30 - 60 years old)
- ▷ Aged (60 - 99 years old)

Define the variables `Smoker` and `Cancer` according to two values: true / false.

1.3 Conditional probability tables

Once the structural part of the graph is created (nodes and arcs), it remains to fill the conditional probability table of each node.

Specify the direct probabilistic relationships between age, smoking and cancer, considering the following tables.

Age	Smoker = Yes	Smoker = No
Young	70	30
Adult	50	50
Aged	25	75

Table 1: Smoke Node: The younger the patient is, the more likely he smokes.

Age	Smoker	Cancer = Yes	Cancer = No
Young	Yes	5	95
	No	1	99
Adult	Yes	15	85
	No	0.1	99.99
Aged	Yes	5	95
	No	1	99

Table 2: Cancer Node: The older the patient is, the higher the probability of having cancer, smoking being an aggravating factor.

Continue to define the conditional probability table for each of the nodes (the values are subject to your expertise or wikipedia).

2 Exploit the Bayesian Network

The modeling being completed, we will be able to exploit the Bayesian network.

It is a question of using inference mechanisms to observe the probabilities of the various modalities of a variable, to assign a certain value to a variable, or to define a degree of likelihood on the values of the variables.

Suppose the specialist wants to check the "or logical" `TbOuCa`. It will perform a "monitoring" on this variable as well as its fathers (`Tuberculose` and `Cancer`). Here is part of the process to check a "or logical":

- ▷ If we observe the Yes modality of `Cancer`, then the Yes modality of `TbOuCa` must change to 100 %.
- ▷ Reciprocally, the observation of `Tuberculosis` at Yes implies a probability of 100 % for the Yes modality of `TbOuCa`.

Measure the impact of `TbOuCa` on the variables `Cancer` and `Tuberculosis`, for example by introducing moderate observations on `TbOuCa`.

3 Extract a network from a database

Using his career, the cancer specialist has created a database (file `Asia.txt`) containing patient information that he has consulted to leave a record of his successor. A line in the database corresponds to the diagnosis of a patient. We have 10,000 patients here.

The specialist who replaces him wants to determine all the associations between these variables in order to predict the risks for a patient to have cancer and to be able to judge whether it is necessary to perform an X-ray.

Launch the learning algorithms to exploit the database (learning structure and parameters).

Using inference, deduce the characteristics of people with cancer (`Cancer = yes`).

Conclusions?