Artificial Intelligence Laboratory 3: Bayesian Network

DT8012, Halmstad University

December 2, 2019

Introduction

This lab is designed to introduce you the use of Bayesian Network (BN). By the end of this lab session you will:

- Learn how to estimate BN parameters from data.

- Learn how to use BN for reasoning and decision making.

For this lab, you will use two datasets to construct BN:

- **Artificial Smart Grid Data**

- **Real Smart Grid Data**

Before you arrive at the lab:
You should read through the entire document and look up anything you don't quite remember from the lectures.

After the lab:
Within two weeks after this lab session you must hand in your **report**. Your report should explain your results, and **how** you achieved them. Make sure you include any relevant information to your explanation.
The deadline for your report submission is **December 19th**. The reports which are sent after December 15th will be graded at the end of February or at the end of May. This means there are only three fixed dates for grading. If you submit your reports after these dates, you need to wait until the next time that the course is given to get your grade.

Environment setup

For this lab, we work with the Bayesian Network Software for Artificial Intelligence, **Bayes Server**. To download the software you need to go to the following webpage: https://www.bayesserver.com/

In the Download page, you need to fill your personal information and then click on the "Email download details". After a few minutes, the download link will be sent to your email. Click on the link and install the software.

For this lab, we are using the trial mode of the software. Therefore, when you open the Bayes Server software, select the "Check this box to trial the Express edition". Then click OK.
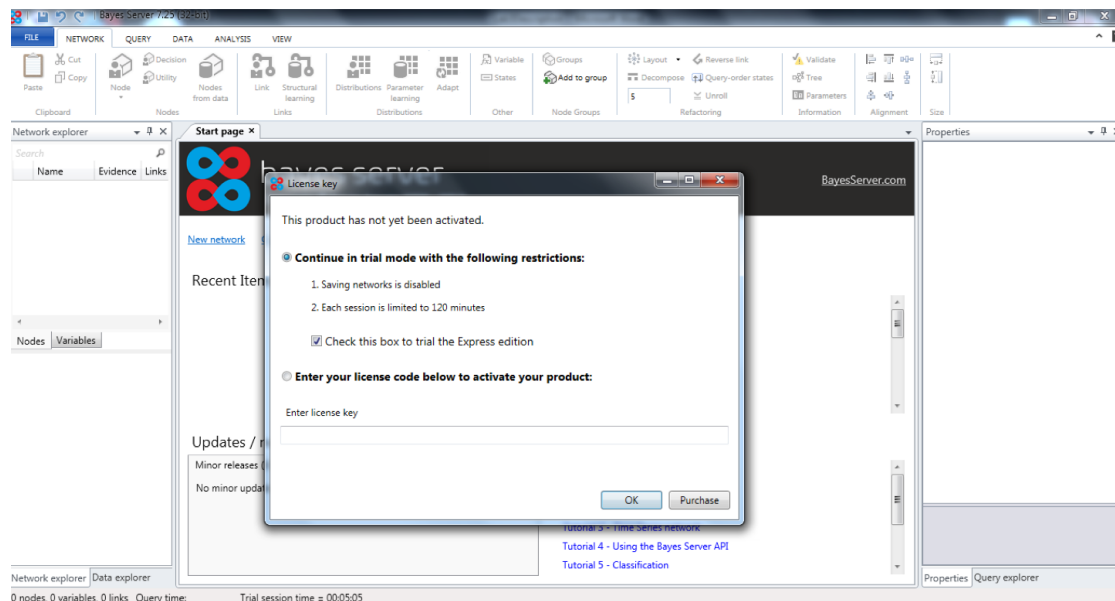

Figure 1: Bayes Server - License key

Note that in the trial version you cannot save the created networks. Moreover, every session is limited to 120 minutes.

In the learning center (https://www.bayesserver.com/docs/), you can find help, tutorial, and videos about how to use the software.

In this lab, we are mainly working with structure learning and parameter learning:

- https://www.bayesserver.com/docs/learning/structural-learning

- https://www.bayesserver.com/docs/learning/parameter-learning

Datasets

In the **Lab3.zip**, you can find the records of failures in smart grids with the specific conditions in which the failures have happened:

- **data_artificial.xlsx**: an artificial failure dataset, containing 700 observations.

- **data_real.xlsx**: a real historical failure dataset from a smart grid in Sweden, containing 1657 observations.

data_artificial.xlsx

The attributes in this dataset are: Number_of_Customers, Time, Day, Season, Weather, Demand_Factor, Overload, and Outage_Duration, with values listed in the following. Part of this dataset is shown in Figure 2.

- **Season**: Spring, Summer, Autumn, Winter

- **Outage_Duration**: Less_than_1H, More_than_1H

- **Number_of_Customers**: Low, High

- **Overload**: Yes, No

- **Weather**: Cold, Warm

- **Time**: Morning, Afternoon, Evening, Night

- **Demand_Factor**: Low, Medium, High

- **Day**: Weekdays, Weekend

Season	Outage_Duration	Number_of_Customers	Overload	Weather	Time	Demand_Factor	Day
Autumn	Less_than_1H	Low	Yes	Cold	Morning	Low	Weekdays
Winter	Less_than_1H	Low	No	Cold	Evening	Low	Weekdays
Spring	More_than_1H	Low	No	Cold	Evening	Low	Weekdays
Winter	Less_than_1H	High	No	Warm	Morning	Low	Weekdays
Spring	More_than_1H	Low	No	Cold	Morning	Low	Weekend
Winter	More_than_1H	Low	No	Cold	Morning	Medium	Weekdays
Autumn	More_than_1H	Low	No	Warm	Evening	Low	Weekdays
Spring	More_than_1H	Low	No	Cold	Evening	High	Weekend
Summer	Less_than_1H	Low	Yes	Warm	Evening	High	Weekend
Winter	More_than_1H	High	No	Cold	Night	Low	Weekdays
Autumn	Less_than_1H	Low	Yes	Cold	Night	Low	Weekend
Spring	Less_than_1H	High	Yes	Cold	Afternoon	High	Weekend
Summer	Less_than_1H	High	Yes	Warm	Afternoon	Medium	Weekend
Autumn	Less_than_1H	High	Yes	Warm	Evening	High	Weekend
Autumn	More_than_1H	Low	No	Cold	Afternoon	Medium	Weekdays
Autumn	More_than_1H	Low	No	Cold	Morning	Low	Weekend
Summer	Less_than_1H	Low	No	Cold	Morning	Medium	Weekend

Figure 2: An example of dataset data_artificial.xlsx

data_real.xlsx

The attributes in this dataset are: Year, Season, Hour, Outage duration, Switchgear, Volt.level, Cause, Facility part, with values listed in the following. Part of this dataset is shown in Figure 3.

- **Year**: before_2011, after_2011

- **Season**: Spring, Summer, Autumn, Winter

- **Hour**: Morning, Afternoon, Evening, Night

- **Outage_Duration**: $Otg \leq 1$, $1 < Otg \leq 2$, $Otg > 2$

- **Switchgear**: H2, H3, H4, H7, H8, H10, LINEHED EON

- **Volt.level**: $0.2 \leq u \leq 1.0$, $10 \leq u \leq 12$

- **Cause**: Digging, Fabrication fault, and ...

- **Facility part**: Ground cable pillar, Concr.sec.substation indoor man., and ...

Year	Month	Hour	Outage duration (h)	Switchgear	Volt.level	Cause	Facility part
after_2011	Spring	Morning	1<Otg<=2	H7	0.2 <= u <= 1.0	Digging	Concr.sec.substation indoor man.
before_2011	Summer	Morning	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Concr.sec.substation indoor man.
before_2011	Summer	Evening	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Concr.sec.substation indoor man.
before_2011	Autumn	Afternoon	1<Otg<=2	H10	0.2 <= u <= 1.0	Fuse break	Concr.sec.substation indoor man.
before_2011	Spring	Morning	1<Otg<=2	H3	0.2 <= u <= 1.0	Fuse break	Concr.sec.substation indoor man.
after_2011	Summer	Evening	1<Otg<=2	H3	0.2 <= u <= 1.0	Fuse break	Concr.sec.substation indoor man.
before_2011	Spring	Night	1<Otg<=2	H4	0.2 <= u <= 1.0	Fuse break	Concr.sec.substation indoor man.
after_2011	Winter	Evening	1<Otg<=2	H4	0.2 <= u <= 1.0	Fuse break	Concr.sec.substation indoor man.
after_2011	Summer	Afternoon	1<Otg<=2	H7	0.2 <= u <= 1.0	Incorrect method/instruction	Concr.sec.substation indoor man.
before_2011	Autumn	Evening	1<Otg<=2	H7	0.2 <= u <= 1.0	Incorrect operating	Concr.sec.substation indoor man.
before_2011	Spring	Evening	1<Otg<=2	H10	0.2 <= u <= 1.0	Unknown	Concr.sec.substation indoor man.
before_2011	Winter	Morning	1<Otg<=2	H3	0.2 <= u <= 1.0	Overload	Concr.sec.substation outdoor man.
after_2011	Autumn	Afternoon	1<Otg<=2	H3	0.2 <= u <= 1.0	Digging	Ground cable fuse-/apparatus box
before_2011	Autumn	Morning	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Ground cable fuse-/apparatus box
before_2011	Summer	Afternoon	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Ground cable fuse-/apparatus box
after_2011	Spring	Morning	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Ground cable fuse-/apparatus box
after_2011	Autumn	Morning	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Ground cable fuse-/apparatus box
after_2011	Spring	Evening	1<Otg<=2	H10	0.2 <= u <= 1.0	Fabrication fault	Ground cable fuse-/apparatus box

Figure 3: An example of dataset data_real.xlsx

Task 1: data_artificial.xlsx

a. Load the nodes from **data_artificial.xlsx** dataset by using the feature "Nodes from data". Then, add some connections between the nodes by using the feature "Link". Now you have the structure of your BN. Use the feature "Parameter learning" to learn the probability distributions of your BN based on the connections you have created. Finally, click on **All** at the bottom of the page that you have created your network to see the parameters.

In your report, you need to show your BN and the corresponding parameters. You can save your BN and the parameters using **File** >>> **Save as image** tab.

b. In the parameter learning, the performance of your BN based on the available data is represented by Log Likelihood, and BIC (see Figure 4).

Log Likelihood - the log likelihood of the data, given the candidate network.
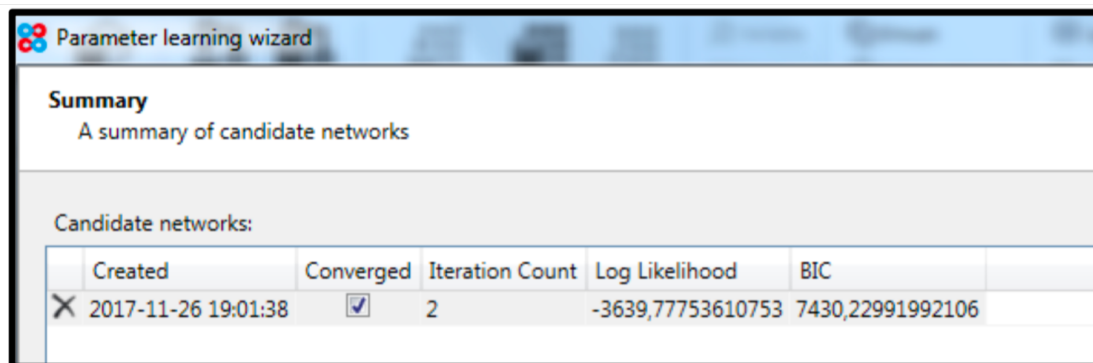
BIC - Bayesian Information Criterion.

In your report, explain what each of these measures represent.

c. Construct two more networks (the same way as task 1a). Compare the results of BIC and Log Likelihood of these three networks together.

In your report, you need to show the networks and corresponding parameters. Also you need to write the BIC and Log Likelihood of these networks. Furthermore, you need to explain your conclusions about the three networks considering the performance measures.

d. Start a new network and load the nodes from **data_artificial.xlsx** dataset by using the feature "Nodes from data". Then use the feature "Structure learning" to automatically learn the links between different attributes in the data file. Use the feature "Parameter learning" to learn the probability distributions of the BN.

In your report, you need to show the networks and corresponding parameters. Also, explain the results of the performance measures in comparison to other networks that you have created in the previous steps.

Figure 4: Example of the summary of candidate network

Task 2: data_real.xlsx

a. Repeat all the requirements of task 1 using the dataset **data_real.xlsx**.

In your report, you need to show all the four networks (for the first three networks you define the links, and for the fourth network the links are learnt using "Structure

learning") with the corresponding parameters and performance measures. You also need to explain the results.

b. Find the following conditional probabilities for the network which the links are learnt using the "Structure learning". **Write the results in your report**.

1. p(Cause=Animal | Season=Autumn)=?

2. p(Season=Autumn | Cause=Animal)=?

3. p(Season=Summer | Cause=Thunder)=?

4. p(Outage duration=$Otg \leq 1$ | Facility part = Ground cable pillar)=?

5. p(Facility part=Ground cable pillar | Switchgear=H7, Cause=Fuse break)=?

6. p(Facility part=Ground feeder cable in ground|Cause=(Digging, Fabrication fault), Switchgear=H7, Season=Summer)=?

7. p(Facility part=Ground feeder cable in ground | Cause=¬(Digging, Fabrication fault), Switchgear=H7, Season=Summer)=?

8. p(Cause=Digging | Facility part = OH line, Switchgear=H7)=?

9. p(Facility part = Ground cable pillar | Outrage duration=$Otg > 2$=?

10. $\frac{p(Cause=Unknown \mid Year=before2011)}{p(Cause=Unknown \mid Year=after2011)}$ =?

c. **In your report**, write 10 more of such examples with the corresponding probabilities.

d. **In your report**, explain why when we are selecting an attribute of a node (changing the probability of the node to 100%), not only the probability values of its child node are changing but also the probability values of the parent nodes are changing. Support your explanations by using some examples and their corresponding probability equations.

e. Suppose you want to do diagnostic on the smart grid base on the constructed network and probability values. **Explain in your report**, which attributes would maximize the probability of failure in "Ground feeder cable in ground"? Which attributes would maximize the probability of having "Fabrication fault" as the cause of failure?

f. Which attributes are highly correlated with each other compared to other attributes? **In your report**, provide at least 6 examples (3 with highly positive correlation and 3 with highly negative correlation) and interpret the results.

Example: "Cause=Overload" is highly correlated with "Season=Winter >>> p(Cause=Overload | Season=Winter)=18.38" is the highest among all the other seasons. **This means** if we know that "it is Winter", then the probability of having a failure caused by "Overload" is higher (more than two times higher) compared to other seasons.

Grading Criteria

- In order to pass this lab, you need to complete all the tasks.