PRACTICAL APPLICATION 2

Breast Cancer Probabilistic Graphical Models

Author:
Serena Alderisi
s.alderisi@alumnos.upm.es

Course:
Machine Learning
Master Degree in Data Science
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

June, 2020

Contents

Introduction

The aim of this work is focused on learning and inferencing the Bayesian Network built for the breast cancer dataset. Bayesian Networks are a probabilistic graphical model based on a firm mathematical foundation: the Bayes theorem .

This tecnique allows to adquire knowledge about a certain domain through its graphical representation, using the informations about how the features are relate to each others. The strengh of this model is the capability of allowing for an explicit representation, in an intuitive graphical form. It makes the relations between the data trasparent to the researcher, that in this way can also query the model to gain further knwoledge. Hence, a Bayesian Network structure is expressed as a graphical representation of the dependence of the features, where nodes symbolize the variables and the arcs represent the dependence relationship between nodes.

This paper will explore the different methods to learn the strcuture and the parameters of a Bayesian Network. Furthermore, it will follow a comparison of all the built candidates and it will be show how to select the network structure that fits better the breast cancer data.

Once obtained the Bayesian network, exact and approximate inference is performed. It's possible, for example, having an idea about the probability of having breast cancer before to do the biopsy to diagnostic the nature of the tumor. The inference is based on the question "What is it the probability of having breast cancer, given the results of a mammography?"

Problem description

The dataset used for this work is "Original Wisconsin Breast Cancer Database", collected by Dr. WIlliam H. Wolberg from University of Wisconsin Hospitals Madison, Wisconsin, USA. This dataset was made public in 1992 and collects the samples reported by Dr. Wolberg about his clinical cases. It consists of 699 observations and 11 variables:

- *Sample code number*: id number
- *Clump Thickness*: Assesses if cells are mono- or multi-layered.
- *Uniformity of Cell Size*: Evaluates the consistency in size of the cells in the sample.
- *Uniformity of Cell Shape*: Estimates the equality of cell shapes and identifies marginal variances.
- *Marginal Adhesion*: Quantifies how much cells on the outdside of the epithelial tend to stick together.
- *Single Epithelial Cell Size*: Relates to cell uniformity, determines if epithelial cells are significantly enlarged.
- *Bare Nuclei*: Calculates the proportion of the number of cells not surrounded by cytoplasm to those that are.

- *Bland Chromatin*: Rates the uniform ”texture” of the nucleus in a range from fine to coarse.
- *Normal Nucleoli*: Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.
- *Mitoses*: Describes the level of mitotic (cell reproduction) activity.
- *Class*: the nature of the cancer 2 for benign, 4 for malignant.

Except Sample code number and Class variables, all the other features are described by a range of numbers between 0 and 10. All the variables are integers except the Bare Nuclei that is a character. The data type will be a point to work on since bnlearn, the main R package used in this work, since both of them are not supported.

Methodology

This work has been carried out using RStudio.
The aim of this work consists of learning a DAG (directed acyclic graph) structure and the set of corresponding parameters, on which performe Bayesian inference at the end.

Before to do that, a pre-processing analysis is performed.
Firstly, the id number variable has been removed from the dataset and the variable names changed is something more readable. As mentioned in section 2, integer and character data types are not supported in bnlearn package. Hence all the 10 variables have been converted into factors. The first 9 variables of 10 levels and the last one of 2 levels. The breast cancer dataset contains some missing data, exactly 16, that have been dropped out being a little portion compared to the 683 remaining observartions. After the pre-processing, the dataset to work on is composed by 10 variables and 683 observations, of which 65% are benign tumors and 35% are malignant tumors.

Once the pre-processing phase is finished, The Learning Process can start.
Firstly, the structural learning is performed using the Score+Search approach: in this step, four structures are built using different metrics, either with the Hill-Climbing algorithm and the Tabu search algorithm. The structural learning is performed also using the Constraint-Based approach: the network structures are built either with Grow-Shrink algorithm and Incremental Association Markov Blanket algorithm. The best bayesian network structure to represent the breast cancer data is selected between the built ones, following the criteria of the interpretability and comparing the Loss functions obtained after the application of the K-fold Cross Validation. The selected bayesian structure is built with the BIC metric applying the HC algorithm.

The next phase is the Parameter Learning process: finding the best probability distributions for the nodes in the structure conditionally dependent on the parents. The first attempt is performed using the Maximum Likelihood approach that generated some conditional probabilities set to zero. For this reason the learning parameter process, is subsequently performed with the Bayesian estimation method.

Finally, the last phase of this work is dedicated to Bayesian inference: exact and approximate. The query to inference is the same for both approaches: given specific values of the cell shape and the cell size, what is it the probability of having a malignant tumor?

Results and Discussion

4.1 Learning Process

The learning process is the first step to build a bayesian network where the goal is deriving the structure from the data and it is usually performed as a two-stage method: Structure Learning (section 4.1.1) and Parameter Learning (section 4.1.2).

4.1.1 Strucure Learning

In this first phase the aim is to learn the graph structure from the data.
Firstly, the **Score + Search** approach is followed. It's a method based on the applications of various heuristic search algorithms to find the highest-scoring directed acyclic graph (DAG). The score function measures the appropriateness of the DAG for the data. Indeed, score-based methods search over the space of possible graphs maximizing the score that reflects how well the graph fits the data. Different score functions can be used to find the best network structure that fits the data.

The learning process has been carried out by the application of the Hill-Climbing (HC) algorithm that generated the following four network structures. Respectively Figure 4.1, Figure , Figure 4.3 and Figure 4.4 are learnt applying different metrics: penalize log-likelihood metrics (BIC, AIC and Log-Likelihood) that define a score to penalize the complexity of the structure and Bayesian metrics (K2) that measure the probability of the structure given the data.

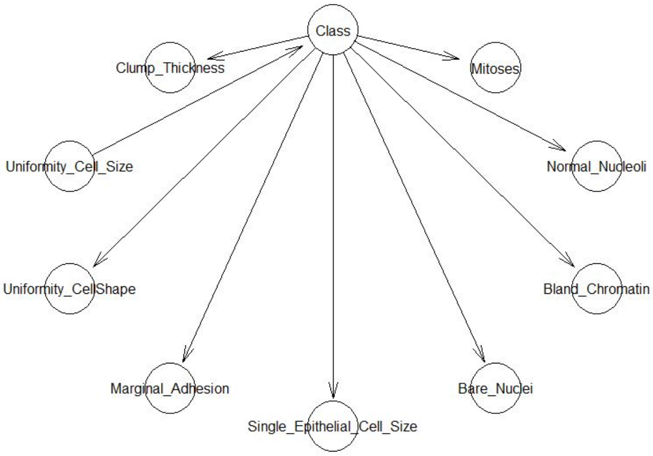


Figure 4.1: BIC Bayesian Network Structure built with HC algorithm

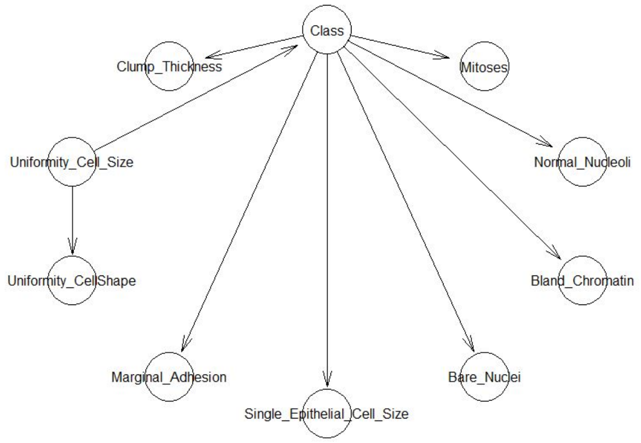


Figure 4.2: AIC Bayesian Network Structure built with HC algorithm

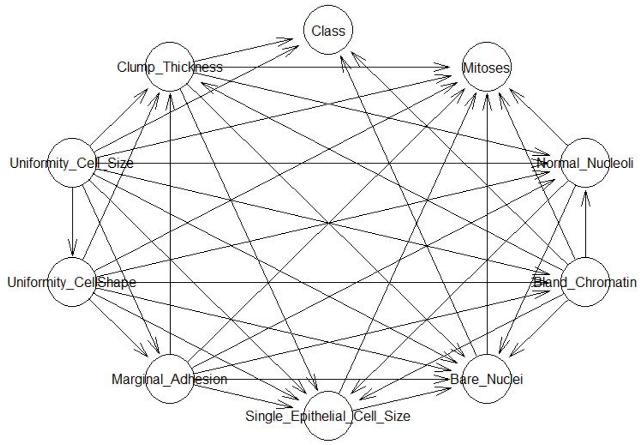


Figure 4.3: Log-Likelihood Bayesian Network Structure built with HC algorithm

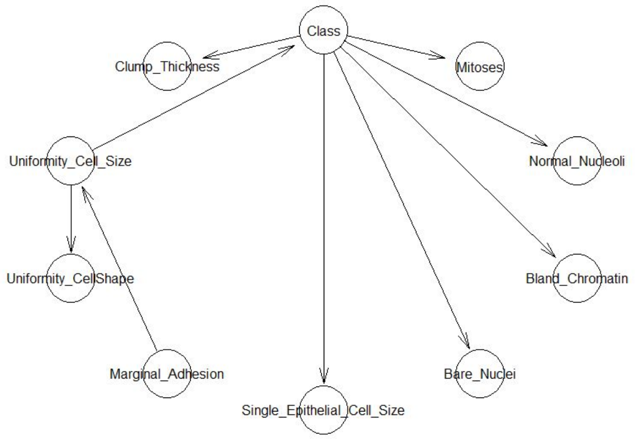
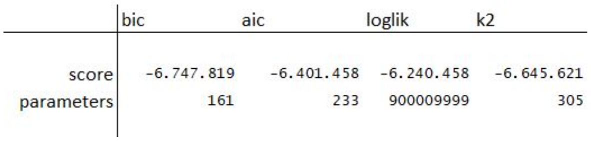


Figure 4.4: K2 Bayesian Network Structure built with HC algorithm

Looking at the networks obtained, the first thing you notice that the Log-link score gives the most complex and difficult structure to analyze between all of them. Indeed, since the likelihood score always increases with the complexity of the model, this approach is not appropriate for model selection. To evaluate which model represents better the relationships between the variable it should be necessary the knwoledge of an expert in the field. Without further medical insights it can be assumed that all the other three obtained models could make sense, even if, AIC and BIC metrics generated simplest models. The only difference between the BIC structure and the AIC structure is that in the latter the Cell Shape depends only on the Cell Size. Since be focused on medical details is not the aim of this work, in this context the simplest model has been prefered. The BIC structure states that the class feature only depends on the Cell Size and the Class variable is the only dependency for all the remaining variables.

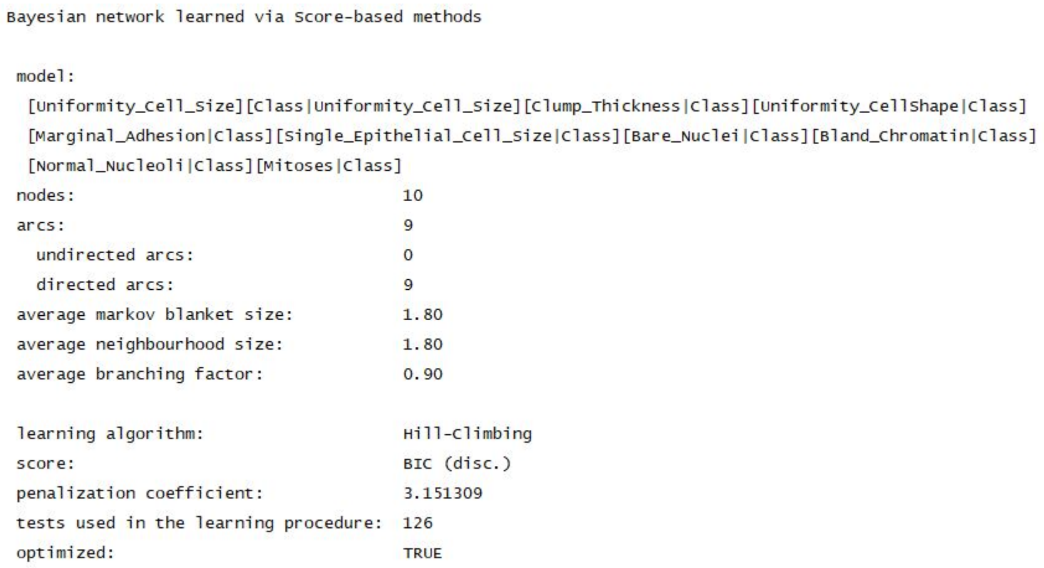
The analysis and the evaluation of the four network structures obtained with the HC algortihm, are shown in Figure 4.5: a comparison between the scores and the numbers of free parameters for each graph it's proposed.



	bic	aic	loglik	k2
score	-6.747.819	-6.401.458	-6.240.458	-6.645.621
parameters	161	233	900009999	305

Figure 4.5: Scores Comparison

The selection of the best network structure is resumed to a maximization problem: the better the network, the greater the score. This implies pick the network with the highest likelihood and the lowest penalty between the considered networks. In this case the network with the highest score is the one built using the Log-Likelihood score, the same that presented the most complex structure. Rather than the structure with the best causal or probabilistic dependency between variables, in this work the interest is directed to find the best predictive structure: the goal is to classify a new potential case of breast cancer as malignant or benignant. Hence, the choise of the network in this case is based on the interpretability: the BIC network is the easiest to interpret.The output is shown is Figure 4.6.



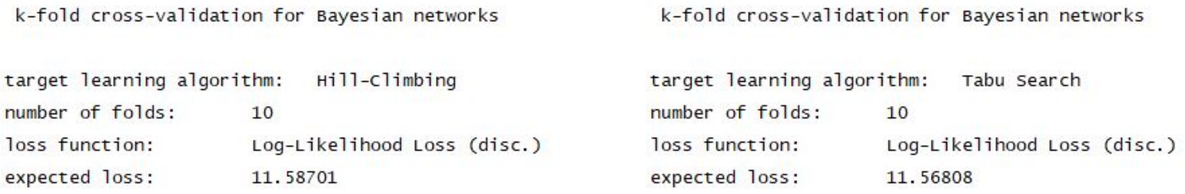
```
Bayesian network learned via Score-based methods

  model:
   [Uniformity_Cell_Size][Class|Uniformity_Cell_Size][Clump_Thickness|Class][Uniformity_CellShape|Class]
   [Marginal_Adhesion|Class][Single_Epithelial_Cell_Size|Class][Bare_Nuclei|Class][Bland_Chromatin|Class]
   [Normal_Nucleoli|Class][Mitoses|Class]
  nodes:                                 10
  arcs:                                  9
    undirected arcs:                     0
    directed arcs:                       9
  average markov blanket size:           1.80
  average neighbourhood size:            1.80
  average branching factor:              0.90

  learning algorithm:                    Hill-Climbing
  score:                                 BIC (disc.)
  penalization coefficient:              3.151309
  tests used in the learning procedure:  126
  optimized:                             TRUE
```

Figure 4.6: BIC Bayesian Network Structure Output

Here is possible to see in detail the number of nodes (variables) and arcs (dependency relationship) that we have in the network, and how many of this arcs are directed or undirected. In this network there are nine 9 directed arcs, that means the dependency relationship has a specific direction. ”The Markov blanket size gives us information about the set of neighboring nodes: its parents, children and the other parents of all its children” - (Wikipedia).

The research of the best network has been conducted also using the Tabu algorithm, with the same metrics seen above. The networks structures are the same of the ones obtained with the Hill-Climbing algorithm.

Cross-validation is a standard way to obtain unbiased estimates of a model’s goodness of fit. By comparing such estimates for different learning strategies it’s possible to choose the optimal one for the data at hand in a principled way. k-fold cross-validation is performed on the algorithms seen so far, Figure 4.7. Looking at the Loss function, the lowest value is the one from the HC algorithm, making this the best one to predict the potential breast cancer.



```
 k-fold cross-validation for Bayesian networks

target learning algorithm:   Hill-Climbing
number of folds:       10
loss function:         Log-Likelihood Loss (disc.)
expected loss:         11.58701
```

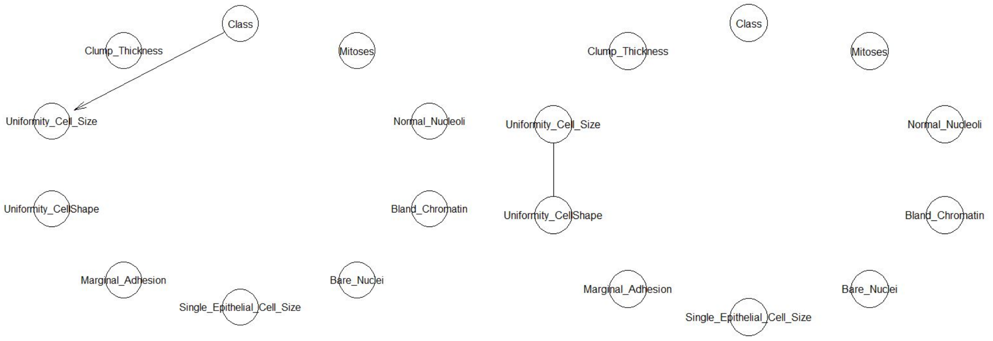
```
 k-fold cross-validation for Bayesian networks

target learning algorithm:   Tabu Search
number of folds:       10
loss function:         Log-Likelihood Loss (disc.)
expected loss:         11.56808
```

Figure 4.7: K-fold Cross Validation Comparison for ’Search + Score’ algorithms

The learning structure process is performed also following the **Constraint-Based** approach. These methods detect conditional independences with appropriate statistical test means and find the structure that represents most of the relationships between the variables. Constraint-based algorithms try to eliminate all the graphs that are inconsistent with the observed constraints, and ultimately return only the statistically equivalent graphs consistent with all the tests.

Firstly, to learn the structure has been used the Grow-Shrink (GS) algorithm and as it’s pos-

sible to see from Figure (a) 4.8 this model is pretty different to the BIC model built with the Hill-Climbing algorithm: the structure built with GS algorithm seems not to make sense.
An other Constraint-Based Learning Algorithms is applied: Incremental Association Markov Blanket (iamb), Figure (b) 4.8. "It's based on the Markov blanket detection algorithm of the same name, which is based on a two-phase selection scheme: a forward selection followed by an attempt to remove false positives" - (bnlearn documentation). The model is even more complex to interpret: it has only one undirected arc.

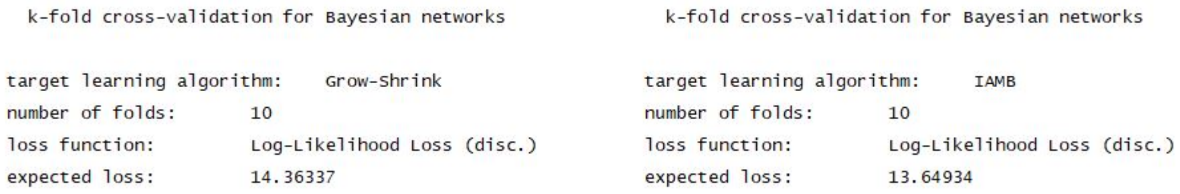


(a) GS algorithm

(b) IAMB algorithm

Figure 4.8: Bayesian Network Structure - Constraint Based approach

K-fold cross-validation comparison is performed also on the two Constraint-based algorithms, Figure 4.9: the one with the lowest Loss function is the iamb algorithm. To end up with the Learning Structure phase, the comparison is done also with the results obtained in Figure 4.7: the algorithm with the lowest Loss function is the Hill-Climbing. It means that, between all the algorithms seen so far, Hill-climbing generalizes better with respect to log-likelihood, HC will predict the new or unseen data better than the others.



```
  k-fold cross-validation for Bayesian networks

target learning algorithm:    Grow-Shrink
number of folds:        10
loss function:          Log-Likelihood Loss (disc.)
expected loss:          14.36337
```

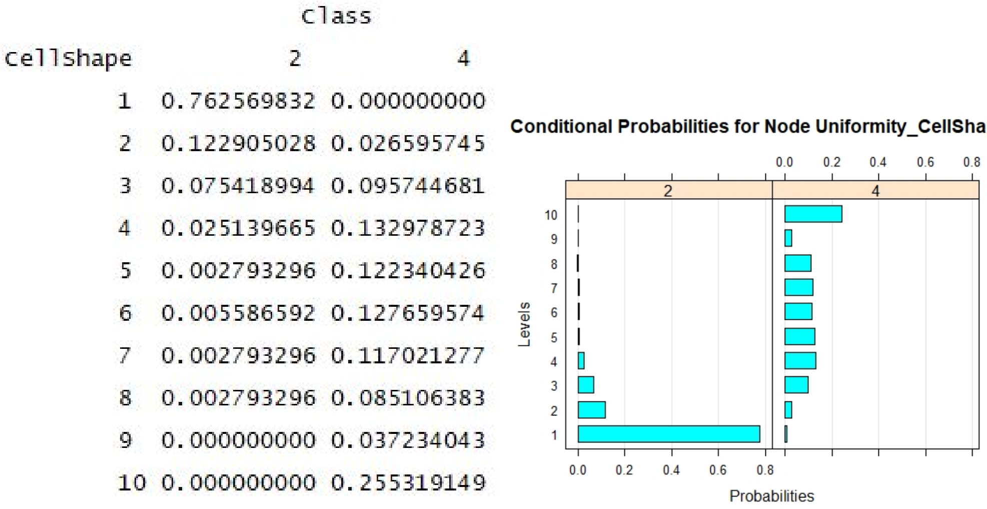
```
  k-fold cross-validation for Bayesian networks

target learning algorithm:     IAMB
number of folds:        10
loss function:          Log-Likelihood Loss (disc.)
expected loss:          13.64934
```

Figure 4.9: K-fold Cross Validation Comparison for Constraint-based algorithms

4.1.2 Parameters Learning

In this phase the aim is to learn the local distributions implied in the learned network structure. Now that the Bayesian network structure is completely directed, the next step is fitting the parameters of the local distributions, which take the form of conditional probability tables.

To do that firstly, the **Maximum Likelihood (ML)** method is used. For instance, the conditional probability table of Cell Shape variable and the relative barchart are plotted in Figure 4.10.

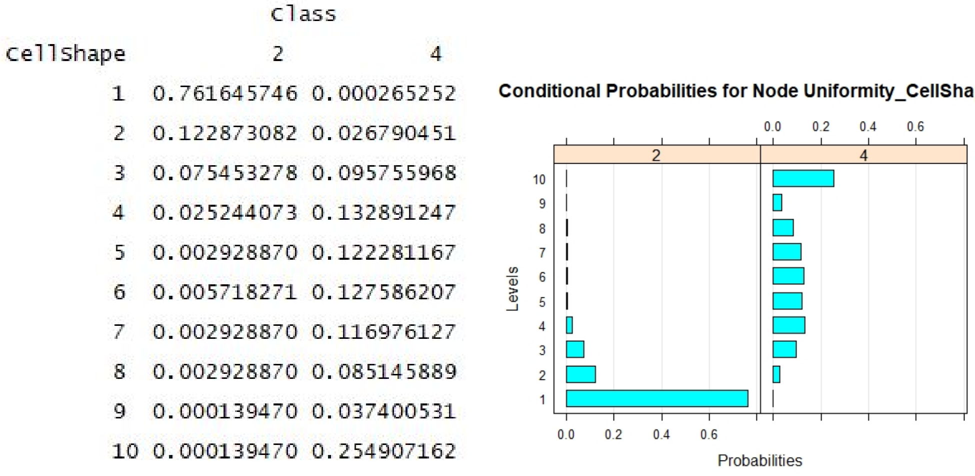


```
                 Class
CellShape            2           4
        1  0.762569832 0.000000000
        2  0.122905028 0.026595745
        3  0.075418994 0.095744681
        4  0.025139665 0.132978723
        5  0.002793296 0.122340426
        6  0.005586592 0.127659574
        7  0.002793296 0.117021277
        8  0.002793296 0.085106383
        9  0.000000000 0.037234043
       10  0.000000000 0.255319149
```

Figure 4.10: Cell Shape conditional probabilities - Maximum Likelihood estimation

Looking in detail at the conditional probabilities tables some probabilities are seto to zero. This behaviour is repeat also for each one of the other 8 nodes.
Zero probabilities can produce undesirable results. In the Pathfinder study, 10% percent of cases were incorrectly diagnosed due to 0 probabilities. The correct disease was ruled out by a finding that had been given 0 probability, "Koller and Friedman, 2006, pp. 67" - (Bojan Mihaljevic Hands on Slides).

The solution to avoid zero probabilities is using **Bayesian Parameter Estimation**. Let's compare the conditional probability table of Cell Shape variable and the relative barchart plotted in Figure 4.11 to notice the changes.



```
                 Class
CellShape            2           4
        1  0.761645746 0.000265252
        2  0.122873082 0.026790451
        3  0.075453278 0.095755968
        4  0.025244073 0.132891247
        5  0.002928870 0.122281167
        6  0.005718271 0.127586207
        7  0.002928870 0.116976127
        8  0.002928870 0.085145889
        9  0.000139470 0.037400531
       10  0.000139470 0.254907162
```

Figure 4.11: Cell Shape conditional probabilities - Bayes estimation

Medical literature teachs us that cancer cells tend to vary in shape. That is why this feature is very valuable in determining whether the cells are cancerous or not (benign 2 or malignant 4). Looking at the plot, without an expert knowledge, it seems that there is a true correlation between

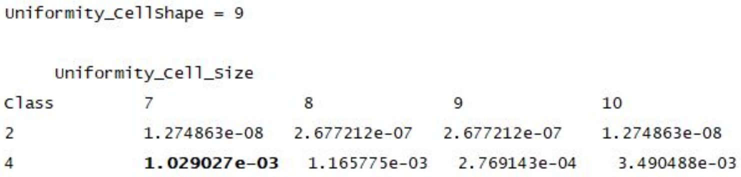
the feature of Cell Shape and the Class variable (the nature of the tumor): smaller values of cell shape suggest that the tumor is benignant and highest values, instead lead to the diagnosys of breast cancer.
The learnt parameters of the network structure chose above, are the conditional probabilities of each node, estimate with the Bayes criteria.

4.2 Inference

Given a Bayesian network, what questions could be asked? The goal of Bayesian Inference is investigating the effects of a new evidence using the knowledge encoded in the graph. It's possible to do that inferring the value of a set of variables by using other variables as evidence.
As said above, it's known that cancer cells tend to vary in shape and in size. For istance, let's imagine the following situation: a patient that suspects to suffer of breast cancer takes a mammography to know the size and the shape of the tumor. Without any further investigation, let's suppose that the results of the clinical exam are: Cell Shape equal to 9 and Cell Size equal to 7. An interisting query could be: "How likely it is for a patient to have a malignant tumor with a Cell Shape equal to 9 and Cell Size equal or greater than 7?". It's possible to answer to this question performing an exact inference or an approximate inference.

4.2.1 Exact

Exact inference implies the analytical computation of the conditional probability distribution over the variables of interest. The results are obtained with the query grain function that performs posterior inference via belief propagation. It's a message-passing algorithm that operates passing messages among the nodes of the network. Nodes act as processors that receive,calculate and send information, Figure 4.12.
The number highlighted in bold is the probability of having a malignant tumor, given the Cell Shape equal to 9 and Cell Size equal to 7 is 0.001029027. The other numbers are the probabilities, given Cell Size values greater tha 7 and Cell Shape egual to 9, of having a benign tumor (first row) or a malignant tumor (second row). Something weird in these results is that all the probabilities are very close to zero, regardless the nature of the tumor.



```
Uniformity_CellShape = 9

    Uniformity_Cell_Size
Class        7               8              9              10
2            1.274863e-08   2.677212e-07   2.677212e-07   1.274863e-08
4            1.029027e-03    1.165775e-03   2.769143e-04    3.490488e-03
```

Figure 4.12: Exact Inference Output

4.2.2 Approximate

Sometimes exact inference is too hard to perform, so it's possibile to use approximation techniques based on statistical sampling. Given the learnt Bayesian network, the Probabilistic Logistic Sampling estimation is performed. The function cpdist returns a data frame containing the random samples generated from the conditional distribution of the nodes conditional on evidence. In this case it haven't been conditioned on any evidence. Then, applying the cp query function the query the estimation is obtained: the probability of having a malignant tumor, given Cell shape equal to 9 and cell size equal to 7 is 0.0017.

Conclusion

The obtained results in the Inference, section 4.2, are quite sospicious. Even if it doesn't concern the final goal of this work, a further investigation could be performed with the partecipation of a medical expert to come back to the learning phase and choose a more realistic network.

This points out the difficulty of building a graphical model. In the learning structure process, in this work, has been selected the simplest network, using the bic metric and the hill-climbing algorithm. Maybe having a depeer knwoledge of the domain would has leaded the choice towards an other structure. An other interesting point for further works would be how the results change selecting an other structure. As discussed above the structure learning has been the trickest part of this work.

References

1. Approximate structure learning for large Bayesian networks - Mauro Scanagatta, Giorgio Corani, Cassio Polpo de Campos and Marco Zaffalon (2018).

2. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood - José A. Gámez, Juan L. Mateo and José M. Puerta - 2010.

3. Bielza, C., Li, G. and Larrañaga, P. (2011). Multi-Dimensional Classification with Bayesian Networks. International Journal of Approximate Reasoning, 52, 705-727.

4. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

5. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume

6. G. Cooper and E. Herskovitz, A Bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992), 330–347.

7. Russell SJ, Norvig P (2009). Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition.

8. Advanced State Space Methods for Neural and Clinical Data - Zhe Chen - Cambridge University Press, Oct 15, 2015.

9. A Bayesian Metric for Evaluating Machine Learning Algorithms - Lucas R. Hope and Kevin B. Korb - School of Computer Science and Software Engineering, Monash University Clayton, Australia.

10. A Brief Introduction to Graphical Models and Bayesian Networks - Kevin Murphy, 1998.

11. The Grow-Shrink Strategy for Learning Markov Network Structures Constrained by Context-Specific Independences - Alejandro Edera, Yanela Strappa, Facundo Bromberg 2014.

12. A Review of Parameter Learning Methods in Bayesian Network - Zhiwei Ji, Qibiao Xia, Guanmin Meng - 2015