Bayesian network analysis
for
Probabilistic Modeling

by

Marticola No: 942091

Amanpreet singh

Università degli Studi di Milano

TABLE OF CONTENTS

Chapter 1

Bayesian Network Analysis

1.1 Data description-Risks of Cardiovascular disease

The Framingham Heart Study (FHS) is an ongoing cohort study dedicated to identifying common factors or characteristics that contribute to cardiovascular disease (CHD). The aim of the project is to finding conditional joint probabilities for the factors that can lead to CHD among different type of people and habits

Variables

Demographic Risks

Male: 0 = Female; 1 = Male
Age at exam time.:Range 30-72
education:1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4

Behavioral Risks:

CurrentSmoker - Current smoker or not
CigsPerDay - Average number of cigarettes smoked per day

Medical experiments:

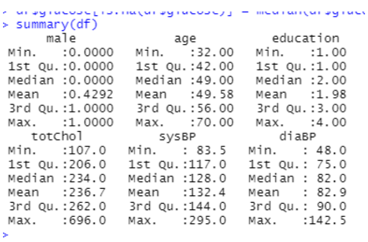
BPMeds - Patient is under blood pressure medication
PrevalentStroke - Previously had a stroke or not
PrevalentHyp - Prevalent Hypertension or not
Diabetes - Patient has diabetes or not
TotChol - Total Cholesterol
Glucose - Glucose level

Physical examination:

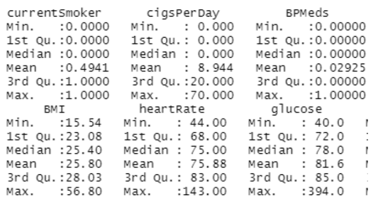
DiaBP - Diastolic blood pressure
BMI - Body mass index
Heart - Rate Heart Rate
SysBP - Symbolic blood pressure

Prediction Label

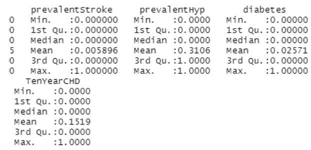
: TenYearCHD- Predicting if someone will have 10 year risk of coronary heart disease CHD



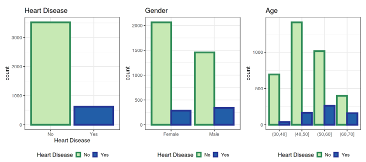
```
> summary(df)
      male             age          education
 Min.   :0.0000   Min.   :32.00   Min.   :1.00
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.00
 Median :0.0000   Median :49.00   Median :2.00
 Mean   :0.4292   Mean   :49.58   Mean   :1.98
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.00
 Max.   :1.0000   Max.   :70.00   Max.   :4.00
    totChol          sysBP           diaBP
 Min.   :107.0   Min.   : 83.5   Min.   : 48.0
 1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
 Median :234.0   Median :128.0   Median : 82.0
 Mean   :236.7   Mean   :132.4   Mean   : 82.9
 3rd Qu.:262.0   3rd Qu.:144.0   3rd Qu.: 90.0
 Max.   :696.0   Max.   :295.0   Max.   :142.5
>
```

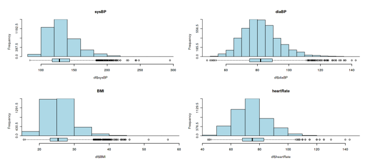


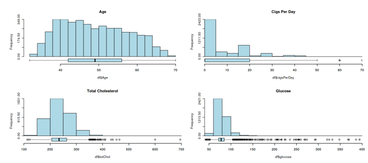
```
 currentSmoker     cigsPerDay        BPMeds
 Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000
 Median :0.0000   Median : 0.000   Median :0.00000
 Mean   :0.4941   Mean   : 8.944   Mean   :0.02925
 3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :70.000   Max.   :1.00000
      BMI          heartRate         glucose
 Min.   :15.54   Min.   : 44.00   Min.   : 40.0
 1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 72.0
 Median :25.40   Median : 75.00   Median : 78.0
 Mean   :25.80   Mean   : 75.88   Mean   : 81.6
 3rd Qu.:28.03   3rd Qu.: 83.00   3rd Qu.: 85.0
 Max.   :56.80   Max.   :143.00   Max.   :394.0
```

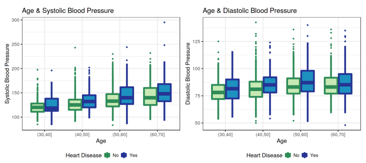


```
    prevalentStroke     prevalentHyp        diabetes
0   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000
0   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000
0   Median :0.000000   Median :0.0000   Median :0.00000
5   Mean   :0.005896   Mean   :0.3106   Mean   :0.02571
0   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000
0   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000
   TenYearCHD
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1519
 3rd Qu.:0.0000
 Max.   :1.0000
```









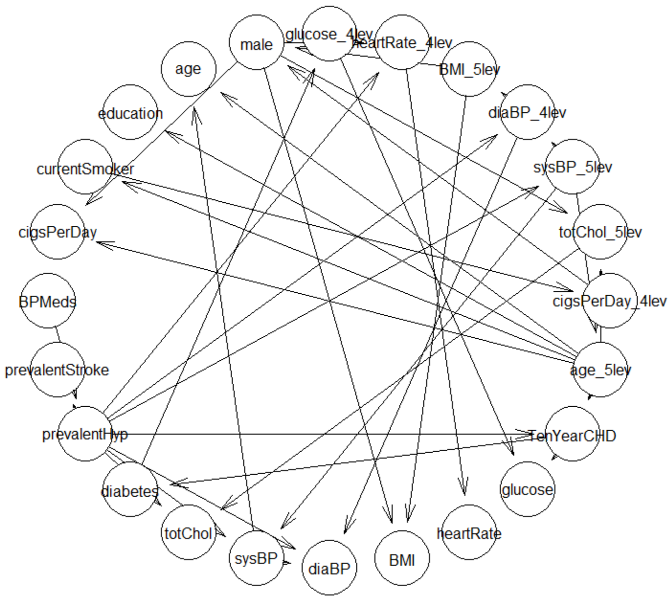
1.2 Methodology

1.2.1 Understanding Connections through Baysean Networks

To understand ,what is the probability and measuring the risk of Coronory Disease, we can link various variables with all possible combinations,of age,smoking habits,gender,BP,glucose levels etc,and create Baysean Networks and joint probability tables and figure out if a person have 10 year CHD or not. Baysean Networks are based on Baysean law,which create various netowrks,where a variable is pointing towards another variable,called nodes, and with each node and each variable linked is independent of its non descendants given its immediate predecessors.

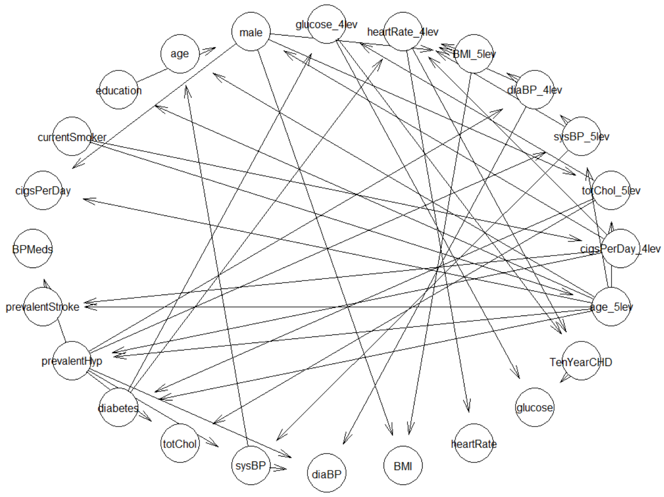
1.2.2 The directed acyclic graph

The DAG represents a factorization of the joint probability distribution into a joint probability distribution and also estimate the parameters of the conditional probability distribution using Bayesian estimation



1.2.3 Whitelist and Blacklist

The first step in learning a Bayesian network is structure learning, that is, using the data to determine which arcs are present in the graph that underlies the model. Often, we would like for that to be a purely automated process—for the purpose of exploring data, or just because we do not know much about the heart disease.Though if we have prior knowledge on what the structure of the network should look like and we can incorporate such knowledge in the structure learning Arcs blacklisted in one direction only (i.e. A $\rightarrow$ B is blacklisted but B $\rightarrow$ A is not) are never present in that particular direction, but may be present in the other direction. Arcs blacklisted in both directions (i.e. both A $\rightarrow$ B and B $\rightarrow$ A are blacklisted) are never present in the graph Arcs whitelisted in one direction only (i.e. A $\rightarrow$ B is whitelisted but B $\rightarrow$ A is not) have the respective reverse arcs blacklisted, and are always present in the graph. Arcs whitelisted in both directions (i.e. both A $\rightarrow$ B and B $\rightarrow$ A are whitelisted) are present in the graph, but their direction is set by the learning algorithm.



1.2.4 Structure Learning :Hill-Climbing Algorithm

Hc is Score-based algorithm,Each candidate DAG is assigned a network score reflecting its goodness of fit, which the algorithm then attempts to maximise.
Hill climbing tries to delete and to reverse each arc in the current candidate DAG and to add each possible arc that is not already present and that does not introduce any cycles.The result with with the highest score is compared with current candidate and if it has a better score then current candidate of DAG becomes the new max.

```
Bayesian network learned via Score-based methods

model:
 [currentSmoker][age_5lev|currentSmoker][cigsPerDay_4lev|currentSmoker][education|age_5lev][prevalentStroke|age_5lev:cigsPerDay_4lev]
 [prevalentHyp|age_5lev:cigsPerDay_4lev][male|education:cigsPerDay_4lev][BPMeds|prevalentHyp][sysBP_5lev|prevalentHyp:age_5lev][cigsPerDay|male:age_5lev]
 [sysBP|prevalentHyp:sysBP_5lev][totChol_5lev|male:age_5lev][diaBP_4lev|prevalentHyp:sysBP_5lev][age|sysBP:age_5lev][diabetes|age_5lev:totChol_5lev]
 [totChol|prevalentHyp:totChol_5lev][diaBP|prevalentHyp:sysBP:diaBP_4lev][BMI_5lev|male:diaBP_4lev][BMI|male:BMI_5lev]
 [heartRate_4lev|diabetes:cigsPerDay_4lev:sysBP_5lev:diaBP_4lev:BMI_5lev][glucose_4lev|diabetes:totChol_5lev][heartRate|heartRate_4lev]
 [TenYearCHD|heartRate_4lev:glucose_4lev][glucose|TenYearCHD:glucose_4lev]
nodes:                                24
arcs:                                 45
  undirected arcs:                    0
  directed arcs:                      45
average markov blanket size:          5.08
average neighbourhood size:           3.75
average branching factor:             1.88

learning algorithm:                   Hill-Climbing
score:                                BIC (cond. Gauss.)
penalization coefficient:             4.176159
tests used in the learning procedure: 1138
optimized:                            TRUE
```

1.2.5 Arc Strength

Model validation based on boot.strength(),which resample and and does model averaging, First it sample a new data set from the original data using learn the structure,then estimate the strength that each possible arc is present in the true DAG.

```
> bootstr[(bootstr$strength > 0.75) & (bootstr$direction >= 0.5), ]
                from               to strength direction
4               male        cigsPerDay    1.000 1.0000000
18              male      totChol_5lev    0.940 1.0000000
86      currentSmoker cigsPerDay_4lev    1.000 0.5450000
122           BPMeds     prevalentHyp    1.000 0.5000000
167     prevalentHyp           BPMeds    1.000 0.5000000
180     prevalentHyp       sysBP_5lev    1.000 0.5000000
181     prevalentHyp       diaBP_4lev    1.000 0.5000000
207         diabetes     glucose_4lev    1.000 0.7840000
232            sysBP              age    0.788 0.6624365
241            sysBP            diaBP    1.000 0.7220000
370         age_5lev              age    1.000 1.0000000
371         age_5lev        education    0.998 0.6332665
372         age_5lev    currentSmoker    0.994 0.5442656
373         age_5lev       cigsPerDay    0.890 1.0000000
384         age_5lev       TenYearCHD    0.888 0.9797297
386         age_5lev     totChol_5lev    1.000 0.9700000
387         age_5lev       sysBP_5lev    1.000 0.5000000
392  cigsPerDay_4lev             male    1.000 1.0000000
424     totChol_5lev          totChol    1.000 1.0000000
445       sysBP_5lev     prevalentHyp    1.000 0.5000000
448       sysBP_5lev            sysBP    1.000 1.0000000
454       sysBP_5lev         age_5lev    1.000 0.5000000
457       sysBP_5lev       diaBP_4lev    1.000 0.5000000
468       diaBP_4lev     prevalentHyp    1.000 0.5000000
472       diaBP_4lev            diaBP    1.000 1.0000000
480       diaBP_4lev       sysBP_5lev    1.000 0.5000000
481       diaBP_4lev         BMI_5lev    1.000 0.5510000
484         BMI_5lev             male    0.938 0.9424307
496         BMI_5lev              BMI    1.000 1.0000000
520    heartRate_4lev        heartRate    1.000 1.0000000
544     glucose_4lev          glucose    1.000 1.0000000
```

1.2.6 Cross Validation:k-fold cross-validation

Cross-validation is done to obtain unbiased estimates for model's goodness of fit. By comparing different combinations of learning algorithms, fitting techniques and the respective parameters.In K fold the data is randomly partitioned into k subsets. Each subset is used in turn to validate the model fitted on the remaining k - 1 subsets.

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

  target network structure:
   [currentSmoker][age_5lev|currentSmoker][cigsPerDay_4lev|currentSmoker][education|age_5lev][prevalentStroke|age_5lev:cigsPerDay_4lev]
   [prevalentHyp|age_5lev:cigsPerDay_4lev][male|education:cigsPerDay_4lev][BPMeds|prevalentHyp][sysBP_5lev|prevalentHyp:age_5lev][cigsPerDay|male:age_5lev]
   [sysBP|prevalentHyp:sysBP_5lev][totChol_5lev|male:age_5lev][diaBP_4lev|prevalentHyp:sysBP_5lev][age|sysBP:age_5lev][diabetes|age_5lev:totChol_5lev]
   [totChol|prevalentHyp:totChol_5lev][diaBP|prevalentHyp:sysBP:diaBP_4lev][BMI_5lev|male:diaBP_4lev][BMI|male:BMI_5lev]
   [heartRate_4lev|diabetes:cigsPerDay_4lev:sysBP_5lev:diaBP_4lev:BMI_5lev][glucose_4lev|diabetes:totChol_5lev][heartRate|heartRate_4lev]
   [TenYearCHD|heartRate_4lev:glucose_4lev][glucose|TenYearCHD:glucose_4lev]
  number of folds:                       10
  loss function:                         Classification Error
  training node:                         TenYearCHD
  number of runs:                        10
  average loss over the runs:            0.1518868
  standard deviation of the loss:        0
```

1.3 Result

1.3.1 Comparing Variables for CVD and Diabetes

```
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Do not Smoke"))))

   No CHD       CHD
0.8469163 0.1530837
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20"))))

   No CHD       CHD
0.8468475 0.1531525
>
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("diabetes"), evidence = (cigsPerDay_4lev == "Do not Smoke"))))

        No        Yes
0.97097984 0.02902016
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("diabetes"), evidence = (cigsPerDay_4lev == "Smoke > 20"))))

        No        Yes
0.97700448 0.02299552
>
>
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Do not Smoke" & heartRate_4lev == "Heart Rate

   No CHD       CHD
0.8535762 0.1464238
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & heartRate_4lev == "Heart Rate 70

   No CHD       CHD
0.8568122 0.1431878
```

```
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Do not Smoke" & heartRate_4lev == "Heart Rate 70-80"))))

   No CHD       CHD
0.8535762 0.1464238
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & heartRate_4lev == "Heart Rate 70-80"))))

   No CHD       CHD
0.8568122 0.1431878
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & heartRate_4lev == "Heart Rate 90+"))))

  No CHD      CHD
0.816382 0.183618
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Do not Smoke" & male == "Female"))))

  No CHD      CHD
0.846783 0.153217
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Do not Smoke" & male == "Male"))))

   No CHD       CHD
0.8464446 0.1535554
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & male == "Female"))))

   No CHD       CHD
0.8470972 0.1529028
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & male == "Male"))))

   No CHD       CHD
0.8491731 0.1508269
>
>
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (diabetes == "No"))))

   No CHD       CHD
0.8482646 0.1517354
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (diabetes == "Yes"))))

   No CHD       CHD
0.7970643 0.2029357
>
>
>
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (as.numeric(totChol_5lev) == 5 & diabetes == "Yes"))))

   No CHD       CHD
0.8066887 0.1933113
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (as.numeric(totChol_5lev) == 1 & diabetes == "No"))))

   No CHD       CHD
0.8515005 0.1484995
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & heartRate_4lev == "Heart Rate 70-80"))))

   No CHD       CHD
0.8549373 0.1450627
> prop.table(table(cpdist(fit, n = 10^6, nodes = c("TenYearCHD"), evidence = (cigsPerDay_4lev == "Smoke > 20" & heartRate_4lev == "Heart Rate 90+"))))

   No CHD       CHD
0.8212845 0.1787155
> |
```

We can see quite intuitive patterns in the results and patterns are quite obvious as,on matching various variables like cigarettes per day ,diabetes etc,we get High probability of risk of Cardiovascular disease.Model also find that people with high cholesterol and diabetes and also people with low cholesterol and no Diabetes has different probabilities and also comparing with people who only has diabetes.

1.3.2 References

Probabilistic Modeling- Federica nicolussi
Understanding Bayesian- Networks Marco Scutari
BayesianNetworks- Francisco Iacobelli
Framingham Heart Study-Cohort (FHS-Cohort)