Assignment-1, CMPSCI 688

Ashish Jain

February 18, 2015

Question 1. Factorization

Joint distribution for any Bayesian Network is obtained as a product of these conditional distribution:

$$P(X) = \prod_{i=1}^{N} P(X_i|Pa_{X_i}^{G})$$

Using the above equation, we can write the factorization for the given graph as follows:

$$\begin{aligned} P(A, G, BP, CH, HD, CP, EIA, ECG, HR) = {} & P(G).P(A).P(BP|G).P(CH|G, A).P(HD|BP, CH).P(HR|A, HD) \\ & P(CP|HD).P(EIA|HD).P(ECG|HD) \end{aligned}$$

Question 2. Likelihood Function

Log likelihood function as an empirical average over the data set is given by following expression:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N} logP_\theta(x_n) \tag{1}$$

$$P_\theta(X = x) = \prod_{d=1}^{D} P_\theta(X_d = x_d|X_{Pa(X_d)} = x_{Pa(X_d)}) = \prod_{d=1}^{D}\prod_{v=1}^{V}(\theta^{X_d}_{v|x_{Pa(X_d)}})^{[x_d=v]} \tag{2}$$

Using the above two equations we will write the probability expression for the given graph and than take log of it.

$$\begin{aligned} P_\theta(A, G, BP, CH, HD, CP, EIA, ECG, HR) = {} & P_\theta(A = a)P_\theta(G = g)P_\theta(BP = bp|G = g) \\ & P_\theta(CH = ch|G = g, A = a)P_\theta(HD = hd|BP = bp, CH = ch) \\ & P_\theta(HR = hr|A = a, HD = hd)P_\theta(CP = cp|HD = hd) \\ & P_\theta(EIA = eia|HD = hd)P_\theta(ECG = ecg|HD = hd) \end{aligned}$$

$$
\begin{aligned}
\mathcal{L}(\theta) &= \frac{1}{N}\sum_{n=1}^{N} logP_\theta(x_n) \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{g}[g_n = g]logP(G = g) + \frac{1}{N}\sum_{n=1}^{N}\sum_{a}[a_n = a]logP(A = a) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{bp,g}[bp_n = bp][g_n = g]logP(BP = bp|G = g) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{ch,a,g}[ch_n = ch][g_n = g][a_n = a]logP(CH = ch|G = g, A = a) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{hd,bp,ch}[hd_n = hd][bp_n = bp][ch_n = ch]logP(HD = hd|BP = bp, CH = ch) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{hr,a,hd}[hr_n = hr][a_n = a][hd_n = hd]logP(HR = hr|A = a, HD = hd) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{cp,hd}[cp_n = cp][hd_n = hd]logP(CP = cp|HD = hd) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{eia,hd}[eia_n = eia][hd_n = hd]logP(EIA = eia|HD = hd) \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{ecg,hd}[ecg_n = ecg][hd_n = hd]logP(ECG = ecg|HD = hd)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}(\theta) &= \frac{1}{N}\sum_{n=1}^{N}\sum_{g}[g_n = g]log\theta^{A}_{a} + \frac{1}{N}\sum_{n=1}^{N}\sum_{a}[a_n = a]log\theta^{G}_{g} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{bp,g}[bp_n = bp][g_n = g]log\theta^{BP}_{bp|g} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{ch,a,g}[ch_n = ch][g_n = g][a_n = a]log\theta^{CH}_{ch|g,a} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{hd,bp,ch}[hd_n = hd][bp_n = bp][ch_n = ch]log\theta^{HD}_{hd|ch,bp} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{hr,a,hd}[hr_n = hr][a_n = a][hd_n = hd]log\theta^{HR}_{hr|a,hd} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{cp,hd}[cp_n = cp][hd_n = hd]log\theta^{CP}_{cp|hd} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{eia,hd}[eia_n = eia][hd_n = hd]log\theta^{EIA}_{eia|hd} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{ecg,hd}[ecg_n = ecg][hd_n = hd]log\theta^{ECG}_{ecg|hd}
\end{aligned}
$$

Question 4. Learning

P(A)	(A)
0.1769	<45
0.3086	45-55
0.5144	>=55

P(BP|G)	P(BP)	P(G)
0.3658	Low	Female
0.6341	High	Female
0.472	Low	Male
0.5279	High	Male

P(HD|BP, CH)	HD	BP	CH
0.5263	N	Low	Low
0.4736	Y	Low	Low
0.5909	N	High	Low
0.409	Y	High	Low
0.5862	N	Low	High
0.4137	Y	Low	High
0.513	N	High	High
0.4869	Y	High	High

P(HR|A, HD)	HR	A	HD
0.0606	Low	<45	N
0.9393	High	<45	N
0.173	Low	45-55	N
0.8269	High	45-55	N
0.3333	Low	>=55	N
0.6666	High	>=55	N
0.6	Low	<45	Y
0.4	High	<45	Y
0.5217	Low	45-55	Y
0.4782	High	45-55	Y
0.5714	Low	>=55	Y
0.4285	High	>=55	Y

5. Probability Queries

We will use following joint probability expression for the given two queries:

$$P(A,B) = P(A|B).P(B)$$

(a)

Random variable CH (cholestrol) can take following two values: Low and High. Let us solve the query using $CH = L$ using the above joint probability equation.

$$P(CH = L|A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No) = \frac{P(CH = L, A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)}{P(A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)} =$$

$$\frac{P(CH = L, A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)}{\sum_{ch \in (L,H)} P(CH = ch, A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)} =$$

(marginalizing over CH)

$$\frac{P(CH = L|A = 2, G = M)P(HD = L|CH = L, BP = L)}{\sum_{ch \in (L,H)} P(CH = ch|A = 2, G = M)P(HD = L|CH = ch, BP = L)} =$$

(Using factorization and conditional independence property, terms indepedent of CH will get cancelled out.)

By using learned CPT tables in Part 4 over training file 1, we get following answer for this Query:

$P(CH = L|A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)$ =0.1522
$P(CH = H|A = 2, G = M, CP = None, BP = L, ECG = Normal, HR = L, EIA = No, HD = No)$ =0.8477

(b)

BP can take two values: Low and High. Let us solve the expression for $BP = L$. We have unobserved variable G in this Query.

$$P(BP = L|A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No) = \frac{P(BP = L, A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No)}{\sum_{bp} P(BP = bp, A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No)} =$$

(marginalizing over BP in denominator)

$$\frac{\sum_{g} P(BP = L, A = 2, G = g, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No)}{\sum_{bp}\sum_{g} P(BP = bp, A = 2, G = g, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No)}$$

(marginalizing over unobserved variable G in denominator and numerator.)

Finally we get after applying factorization and canceling out the terms in numerator and denominator :

$$\boxed{\frac{\sum_{g} P(G = g)P(CH = H|G = g, A = 2)P(BP = L|G = g)P(HR = H|A = 2, BP = L, HD = No)P(HD = No|BP = L, CH = H)}{\sum_{bp}\sum_{g} P(G = g)P(CH = H|G = g, A = 2)P(BP = bp|G = g)P(HR = H|A = 2, BP = bp, HD = No)P(HD = No|BP = bp, CH = H)}}$$

Using the CPT tables learned on training file 1 in Question4, we get

$P(BP = L|A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No) =$ 0.4685
$P(BP = H|A = 2, CP = Typical, CH = H, ECG = Normal, HR = H, EIA = Yes, HD = No) =$ 0.5314

6. Classification

(b)

To simplify: $P(HD = hd|A = a, G = g, BP = bp, CH = ch, CP = cp, EIA = eia, ECG = ecg, HR = hr)$
$hd \in No, Yes$

$$\begin{aligned} &P(HD = hd|A = a, G = g, BP = bp, CH = ch, CP = cp, EIA = eia, ECG = ecg, HR = hr) \\ &= \frac{P(A = a, G = g, BP = bp, CH = ch, HD = hd, CP = cp, EIA = eia, ECG = ecg, HR = hr)}{\sum_{hd} P(A = a, G = g, BP = bp, CH = ch, CP = cp, EIA = eia, ECG = ecg, HR = hr)} \end{aligned}$$

(Simplifying the expression for $hd = Y$.)

$$= \boxed{\frac{P(HD = Y|BP = bp, CH = ch).P(HR = hr|A = a, HD = Y).P(CP = cp|HD = Y).P(EIA = eia|HD = Y).P(ECG = ecg|HD = Y)}{\sum_{hd} P(HD = hd|BP = bp, CH = ch).P(HR = hr|A = a, HD = hd).P(CP = cp|HD = hd).P(EIA = eia|HD = hd).P(ECG = ecg|HD = hd)}}$$

(Simplifying the expression using properties of conditional Independence.)
(Probability terms independent of HD in numerator and denominator will get cancelled out.)

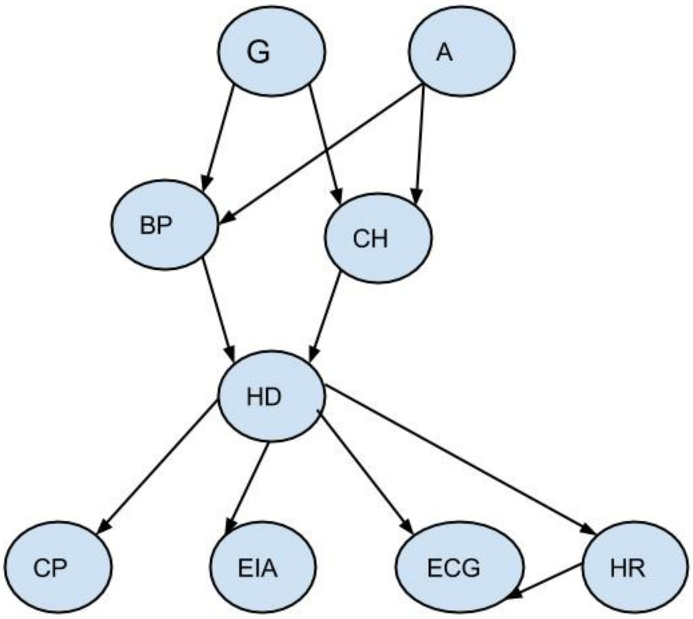
(c)

Fold	Correct	Total	$Accuracy(\%)$
1	44	60	73.33
2	48	60	80
3	40	60	66.66
4	48	60	80
5	47	60	78.33
Mean	45.4	60	75.66

Mean Prediction accuracy over the five test files = 75.66.
Standard deviation of the prediction accuracy over the five test files = 5.12.

7. Modeling

(a)



(b)

Factorization for above Bayes Net can be written as follows:

$$P(A, G, BP, CH, HD, CP, EIA, ECG, HR) = P(A)P(G)P(BP|G, A)P(CH|G, A)P(HD|BP, CH)P(HR|HD) \\ P(CP|HD)P(EIA|HD)P(ECG|HD, HR)$$

(c)

Some of the choices that went into desiging network structure:

- I removed some irrelevant factors like dependence relationship between Age (A) and HeartRate (HR).
- Adding new factor or relationship between Age (A) and Blood Pressure (BP), Electrocardiograph (ECG) and Heart Rate (HR). It affect causal relationships between variables.

Hence overall the network is simplified as compared to the original given network.

(d)

Fold	Correct	Total	$Accuracy(\%)$
1	44	60	73.33
2	49	60	81.66
3	40	60	66.66
4	48	60	80
5	48	60	80
Mean	45.8	60	76.33

Mean Prediction accuracy over the five test files $= 76.33$.
Standard deviation of the prediction accuracy over the five test files $= 5.61$.

Accuracy of above designed network is better than the original network. We removed irrelevant relationships between variables and introduced new causal relationship.