CS412 Mid

Qi Long

qilong2

1

1.1 (a)

Concluded from the given Bayesian Belief Network, $P(S = T, M = F, G = T, V = T, A = F)$
$= P(S = T)P(M = F)P(G = T|S = T, M = F)P(V = T|G = T)P(A = F|G = T)$
$= 0.3 \times (1 - 0.4) \times 0.6 \times 0.75 \times (1 - 0.8) = 0.0162$

1.2 (b)

$P(G = T|S = T) = \frac{P(G=T,S=T)}{P(S=T)}$
$= \frac{P(G=T,S=T,M=T)+P(G=T,S=T,M=F)}{P(S=T)}$
$= \frac{P(S=T)P(M=T)P(G=T|S=T,M=T)+P(S=T)P(M=F)P(G=T|S=T,M=F)}{P(S=T)}$
$= \frac{0.3\times0.4\times0.8+0.3\times(1-0.4)\times0.6}{0.3} = 0.68$

1.3 (c)

$P(G = T|S = F) = \frac{P(G=T,S=F)}{P(S=F)}$
$= \frac{P(G=T,S=F,M=T)+P(G=T,S=F,M=F)}{P(S=F)}$
$= \frac{P(S=F)P(M=T)P(G=T|S=F,M=T)+P(S=F)P(M=F)P(G=T|S=F,M=F)}{P(S=F)}$

$= \frac{(1-0.3)\times0.4\times0.5+(1-0.3)\times(1-0.4)\times0.25}{1-0.3} = 0.35$

1.4 (d)

By equation:

$$LogLikelihood = ln(\prod_{i=1}^{N} P_w(x_i)) \tag{1}$$

$LogLikelihood = ln(\prod_{d=1}^{|D|} P(x^d))$
$= \Sigma_{d=1}^{|D|} lnP(x^d)$
$= \Sigma_{d=1}^{|D|} lnP(S = s^i, M = m^i, G = g^i, V = v^i, A = a^i)$
$= \Sigma_{d=1}^{|D|} ln(P(S = s^i)P(M = m^i)P(G = g^i|S = s^i, m^i)P(V = v^i|G = g^i)P(A = a^i|G = g^i))$
$= \Sigma_{d=1}^{|D|} [lnP(S = s^i) + lnP(M = m^i) + lnP(G = g^i|S = s^i, M = m^i) + lnP(V = v^i|G = g^i) + lnP(A = a^i|G = g^i)]$

2

2.1 (a)

By equation:

$$Sensitivity = \frac{TP}{P} \tag{2}$$

$Sen = \frac{a}{a+b}$
$Sen_{M1} = \frac{300}{300+20} = 93.750\%$
$Sen_{M2} = \frac{320}{320+1} = 99.688\%$

2.2 (b)

By equation:

$$Specificity = \frac{TN}{N} \tag{3}$$

$Spe = \frac{d}{c+d}$
$Spe_{M1} = \frac{11670}{10+11670} = 99.914\%$

$Spe_{M2} = \frac{11677}{2+11677} = 99.983\%$

2.3 (c)

By equation:

$$Accuracy = \frac{TP + TN}{ALL} \tag{4}$$

$Acc = \frac{a+d}{a+b+c+d}$
$Acc_{M1} = \frac{300+11670}{12000} = 99.750\%$
$Acc_{M2} = \frac{320+11677}{12000} = 99.975\%$

2.4 (d)

By equation:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$Pre = \frac{a}{a+c}$
$Pre_{M1} = \frac{300}{300+10} = 96.774\%$
$Pre_{M2} = \frac{320}{320+2} = 99.379\%$

2.5 (e)

By equation:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$Rec = \frac{a}{a+b}$
$Rec_{M1} = \frac{300}{300+20} = 93.750\%$
$Rec_{M2} = \frac{320}{320+1} = 99.688\%$

2.6 (f)

By equation:

$$F1 = \frac{2Pre \times Rec}{Pre + Rec} \tag{7}$$

$F1 = \frac{2\times(\frac{a}{a+c})\times(\frac{a}{a+b})}{\frac{a}{a+c}+\frac{a}{a+b}}$
$F1_{M1} = \frac{2\times0.968\times0.938}{0.968+0.938} = 95.238\%$
$F1_{M2} = \frac{2\times0.994\times0.997}{0.994+0.997} = 99.533\%$

3

3.1 (a)

Since $k = 10$, degree of freedom $= k - 1 = 9$.
A brief explanation is that since 10 partitions of dataset are mutually exclusive, when nine partitions are selected independently and fixed, the last partition is fixed.

3.2 (b)

By formula of t-test:

$$t = \frac{e\bar{r}r(M_1) - e\bar{r}r(M_2)}{\sqrt{var(M_1 - M_2)/k}} \tag{8}$$

$$var(M_1 - M_2) = \frac{1}{k}\Sigma_{i=1}^{k}[err(M_1)_i - err(M_2)_i - (e\bar{r}r(M_1) - e\bar{r}r(M_2))]^2 \tag{9}$$

$$ErrorRate = 1 - Accuracy \tag{10}$$

$e\bar{r}r(A_1) = \frac{(1-0.908)+(1-0.962)+\cdots+(1-0.949)}{10}$
$= \frac{0.695}{10} = 0.0695$
$e\bar{r}r(B_1) = \frac{(1-0.449)+(1-0.585)+\cdots+(1-0.443)}{10}$
$= \frac{5.129}{10} = 0.5129$
so $e\bar{r}r(A_1) - e\bar{r}r(B_1) = 0.0695 - 0.5129 = -0.4434$
$var(A_1 - B_1) = \frac{1}{10}[(1-0.908-1+0.449+0.4434)^2 + (1-0.962-1+0.585+0.4434)^2 + \cdots + (1-0.949-1+0.443+0.4434)^2] = 0.00515$
$t = \frac{0.0695-0.5129}{\sqrt{0.000515}} = -19.533$
By running the provided code, $p - value = 1.118 \times 10^{-8}$
Since $p - value < 0.025 = \alpha/2$, the null hypothesis should be rejected and so the mean accuracy of the two algorithms A1 and B1 are not the same.

3.3 (c)

By formula (8), (9), (10),
$e\bar{r}r(A_1) = \frac{(1-0.908)+(1-0.962)+\cdots+(1-0.949)}{10}$
$= \frac{0.695}{10} = 0.0695$
$e\bar{r}r(B_2) = \frac{(1-0.968)+(1-1)+\cdots+(1-0.966)}{10}$
$= \frac{0.184}{10} = 0.0184$
so $e\bar{r}r(A_1) - e\bar{r}r(B_1) = 0.0695 - 0.0184 = 0.0511$
$var(A_1 - B_2) = \frac{1}{10}[(1-0.908-1+0.968+0.0511)^2 + (1-0.962-1+1+0.0.0511)^2 + \cdots + (1-0.949-1+0.966+0.0511)^2] = 0.000359$
$t = \frac{0.0695-0.0184}{\sqrt{0.0000359}} = 8.529$
By running the provided code, $p - value = 1.323 \times 10^{-5}$
Since $p - value < 0.025 = \alpha/2$, the null hypothesis should be rejected and so the mean accuracy of the two algorithms A1 and B2 are not the same.

4

4.1 (a)

By the Cross Entropy Loss and sigmoid equation:

$$l(y_i, \hat{y_i}) = -y_i log(\hat{y_i}) - (1-y_i) log(1-\hat{y_i}) \tag{11}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1} \tag{12}$$

$$L(\theta) = \frac{1}{N}\Sigma_{i=1}^{N} l(y^i, \sigma(\theta^T x^i)) = \frac{1}{N}\Sigma_{i=1}^{N}[-y^i log(\sigma(\theta^T x^i)) - (1-y^i) log(1-\sigma(\theta^T x^i))]$$

$$= \frac{1}{N}\Sigma_{i=1}^{N}[-y^i log(\frac{1}{1+exp(-\theta^T x^i)}) - (1-y^i) log(1 - \frac{1}{1+exp(-\theta^T x^i)})]$$

4.2 (b)

Since $\hat{y^i}$ is within range [0,1], it can be interpreted as predicting the probability of x^i having true label $y^i = 1$.
Therefore, for prediction:

- When x^i given has true label $y^i = 1$, the predicted probability of x^i having true label is $\hat{y^i}$.
- When x^i given has false label $y^i = 0$, the predicted probability of x^i having true label is $\hat{y^i}$, so that having false label is $1 - \hat{y^i}$.

Therefore, a conditional Bernoulli model can be generated:

$$P(\hat{y^i}|x^i) = \begin{cases} \hat{y^i}, & y^i = 1 \\ 1 - \hat{y^i}, & y^i = 0 \end{cases}$$

By equation (1),
$LogLikelihood = log(\prod_{i=1}^{N} P(\hat{y^i}|x^i)) = log(\prod_{i=1}^{N}((\hat{y^i})^{y^i} \cdot (1 - \hat{y^i})^{1-y^i})$
$= \Sigma_{i=1}^{N}(log((\hat{y^i})^{y^i} \cdot (1 - \hat{y^i})^{1-y^i})) = \Sigma_{i=1}^{N}(y^i log((\hat{y^i})) + (1 - y^i)log((1 - \hat{y^i})))$
$-LogLikelihood = \Sigma_{i=1}^{N}(-y^i log((\hat{y^i})) - (1 - y^i)log((1 - \hat{y^i})))$
In conclusion, Maximize Log Likelihood results in the same loss formula as Cross Entropy Loss in (a).

4.3 (c)

By formula:

$$D_{KL}(p(x)||q(x)) = \Sigma p(x) ln \frac{p(x)}{q(x)} \tag{13}$$

$KL(Bern(p^i), Bern(\hat{p^i})) = \Sigma Bern(p^i) ln \frac{Bern(p^i)}{Bern(\hat{p^i})}$
$= \Sigma[Bern(p^i)(0) ln \frac{Bern(p^i)(0)}{Bern(\hat{p^i})(0)} + Bern(p^i)(1) ln \frac{Bern(p^i)(1)}{Bern(\hat{p^i})(1)}]$
$= (1 - y^i) ln \frac{1-y^i}{1-\hat{y^i}} + y^i ln \frac{y^i}{\hat{y^i}}$
$= (1 - y^i) ln(1 - y^i) - (1 - y^i) ln(1 - \hat{y^i} + y^i ln y^i - y^i ln \hat{y^i}$
$l(y^i, \hat{y^i}) = -y^i ln((\hat{y^i})) - (1 - y^i) ln((1 - \hat{y^i}))$
So $KL(Bern(p^i), Bern(\hat{p^i})) - l(y^i, \hat{y^i})$
$= (1 - y^i) ln(1 - y^i) + y_i lnyi$

- When $y^i = 0$, $KL(Bern(p^i), Bern(\hat{p^i})) - l(y^i, \hat{y^i}) = ln1 = 0$.
- When $y^i = 1$, $KL(Bern(p^i), Bern(\hat{p^i})) - l(y^i, \hat{y^i}) = ln1 = 0$.

Therefore, $KL(Bern(p^i), Bern(\hat{p^i})) - l(y^i, \hat{y^i}) = 0$.

4.4 (d)

I agree. Still using (11), (12) from (a), the loss funtion can be expressed as:

$$L(\theta^*) = \frac{1}{N}\Sigma_{i=1}^{N}[-y^i log(\sigma(\theta^* x^i)) - (1-y^i)log(1-\sigma(\theta^* x^i))] \tag{14}$$

Then the gradient of (14) is that:

$$\nabla L(\theta^*) = \frac{\partial L}{\partial \theta^*} = \frac{1}{N}\Sigma_{i=1}^{N}[-\frac{y^i}{\sigma(\theta^* x^i)} \cdot \frac{\partial \sigma(\theta^* x^i)}{\partial \theta^*} + \frac{1-y^i}{1-\sigma(\theta^* x^i)} \cdot \frac{\partial \sigma(\theta^* x^i)}{\partial \theta^*}] \tag{15}$$

Calculate the derivative first,

$$\frac{\partial \sigma(\theta^* x^i)}{\partial \theta^*} = \frac{exp(-\theta^* x^i) \cdot x^i}{(1+exp(-\theta^* x^i))^2} = (1-\sigma(\theta^* x^i))\sigma(\theta^* x^i)x^i \tag{16}$$

Therefore,

$$\nabla L(\theta^*) = \frac{1}{N}\Sigma_{i=1}^{N}[x^i \sigma(\theta^* x^i)(1-\sigma(\theta^* x^i)) \cdot \frac{\sigma(\theta^* x^i) - y^i\sigma(\theta^* x^i) - y^i + y^i\sigma(\theta^* x^i)}{\sigma(\theta^* x^i)(1-\sigma(\theta^* x^i))}]$$

$$= \frac{1}{N}\Sigma_{i=1}^{N}[x^i(\sigma(\theta^* x^i) - y^i)]$$

Considering the given formula $y^i = \frac{1}{2}(1 + sign(\theta^* x^i))$, and linearly separable property, $sign(\theta^* x^i)$ can only have value -1 or 1, which result in $y^i = 0$ or 1 respectively.

Considering the sigmoid function (12), its value converges to 1 when $\theta^* x^i \to \infty$, converges to 0 when $\theta^* x^i \to -\infty$.

Therefore, when $|\theta^*|$ keeps increasing to infinity, $\sigma(\theta^* x^i)$ will approaches 0 or 1, which is closer to label 0 or 1, resulting in smaller loss.