STA 571 HW1

Ryan Tang

September 9th 2020

1 Excerise 2.6

(a) Let $H \in \{1, ..., K\}$ be a discrete random variable, and let e1 and e2 be the observed values of two other random variables E1 and E2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = [P(H = 1|e_1, e_2), ..., P(H = K|e_1, e_2)]^T$$

Which of the following sets of numbers are sufficient for the calculation?

1. $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$
2. $P(e_1, e_2), P(H), P(e_1, e_2|H)$
3. $P(e_1|H), P(e_2|H), P(H)$

Answer Without any more underlying assumptions, only point (2) is sufficient.
The conditional probability of H can be expanded based off the Bayes rule.

$$P(H|e_1, e_2) = \frac{P(e_1, e_2|H)P(H)}{P(e_1, e_2)}$$

(b) Now suppose we now assume $E_1 \perp E_2|H$

Answer Now all 3 are sufficient.
For (1), given the conditional independence, $P(e_1, e_2|H) = P(e_1|H)P(e_2|H)$.
For (2),

$$\begin{aligned}
P(e_1, e_2) &= \int P(e_1|H)P(e_2|H)P(H)\, dH \\
&= \int P(e_1, e_2|H)P(H)\, dH \\
&= \int P(e_1, e_2, H)\, dH \\
&= P(e_1, e_2)
\end{aligned}$$

2 Excerise 2.7

Pairwise independence does not imply mutual independence

Answer Suppose a simple example $X_1, X_2, X_3 \in 0, 1$ three random binary variables.
Mutual independence $\rightarrow P(X_1, X_2, X_3) = P(X_1)P(X_2)P(X_3)$.
Pairwise independence $\rightarrow P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2)P(X_3)$ instead.
If we suppose *mutual independence* $\rightarrow$ *pairwise independence*, then we have

$$P(X_1) = P(X_1|X_2, X_3)$$

However, the statement is violated given the following example. Say X_2 and X_3 are two fair binary random variables $\in \{0, 1\}$ and X_1 is, instead, generated in the following way where $P(X_1 = 0|X_2 = X_3) = 1$ and vice versa.

X2	X3	X1
0	0	0
1	1	0
1	0	1
0	1	1

3 Excerise 2.8

Conditional independence iff joint factorizes $X \perp Y|Z$ iff $p(x, y|z) = p(x|z)p(y|z)$
Proofs $X \perp Y|Z$ iff $p(x, y|z) = g(x, z)h(y, z)$

Answer Let's assume the above statement is true, then we can write

Proof.

$$\begin{aligned}
\iint p(x, y|z)\, dxdy &= \iint g(x, z)h(y, z)\, dxdy \\
p(x|z) &= \int_Y g(x, z)h(y, z)\, dy \propto g(x, z) \\
p(y|z) &\propto h(y, z) \\
\therefore p(x|z)p(y|z) &= c(z)g(x, z)h(y, z)
\end{aligned}$$

Because both side are proper probability and the intergral of the entire space should sum up to 1. Hence, $c(z)$, the constant has to be 1 which implies

$$g(x, z)h(y, z) \Rightarrow p(x|z)p(y|z)$$

$\square$

4 Excerise 2.12

Proof.

$$
\begin{aligned}
\mathbb{MI}(X,Y) &= KL[p(x,y)||p(x)p(y)] \\
&= \sum_X \sum_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_X \sum_Y p(x,y) \log \frac{p(x,y)}{p(x)} - \sum_X \sum_Y p(x,y) \log p(y) \\
&= \sum_X p(x) \sum_Y p(y|x) \log p(y|x) - \sum_Y p(y) \log p(y) \sum_X p(x|y) \\
&= -\sum_X p(x) H(Y|X=x) - \sum_Y p(y) \log p(y) \\
&= H(Y) - H(Y|X) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

□

5 Excerise 2.15

Proof.

$$
\begin{aligned}
\hat{\theta} &= \arg\max_\theta q(X|\theta) && X \in R^{N\times D} \\
&= \arg\max_\theta \prod_{i\in X} q(x_i|\theta) \\
&= \arg\max_\theta \frac{1}{N} \sum_{i\in X} \log q(x_i|\theta) \\
&= \arg\min_\theta \mathop{\mathbb{E}}_X[-\log q(x|\theta)] && \text{MLE result}
\end{aligned}
$$

$$
\begin{aligned}
KL(p||q(X|\theta)) &= \sum_X p(x) \log \frac{p(x)}{q(x|\theta)} \\
&= \mathop{\mathbb{E}}_X[\log p(x) - \log q(x|\theta)] \\
&= \arg\min_\theta \mathop{\mathbb{E}}_X[-\log q(x|\theta)] && \text{remove constant} \\
&= \text{MLE } \hat{\theta}
\end{aligned}
$$

□

6 Excerise 3.20

(a) For joint distribution without factorization on conditional independence, a look-up table is the best approach to represent the joint distribution. In our binary random variable case with C classes and D features, the look-up table has the size of $[C \times (2^D - 1)]$ or $O(C2^D)$.

(b) The Naive Bayes should perform better if we do not have much data.

(c) If we have infinite sample data, the full-distribution model should perform better.

(d) Both versions should have $O(ND)$. It needs at least one sweep of all available elements in the data set.

(e) Complexity on a single test sample takes $O(CD)$ for Naive Bayes. It needs to go through all the features for each class. Full Bayes takes $O(C)$; it only performs look-up for all classes.

(f) Missing feature data in the test set should not affect the performance in Naive Bayes because features are conditionally independent of each other given a class. Hence $O(C \cdot (v + h))$.

On the other hand, Full Bayes needs to impute the average for these missing averages which takes exponential time during a sweep of the look-up table. Hence $O(v \cdot 2^h)$.

7 Excerise 3.22

Label	Spam	Spam	Spam	Ham	Ham	Ham	Ham
secret		1	1		1		
offer	1	1					
low				1			1
price				1			1
valued				1			
customer				1			
today		1			1		
dollar	1						
million	1						
sports					1	1	
is			1			1	
for				1			
play					1		
healthy						1	
pizza							1

$$\theta_{spam} = \frac{3}{7}$$

$$\theta_{secret|spam} = \frac{2}{3}, \theta_{dollar|spam} = \frac{1}{3}$$

$$\theta_{secret|ham} = \frac{1}{4}, \theta_{sports|ham} = \frac{1}{2}$$

8 Excerise 4.21

(a) We are given the following: $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, $\sigma_2^2 = 10^6$. The decision boundary of QDA between two classes are given by

$$\begin{aligned} p(y=1|x,\theta_1) &= p(y=2|x,\theta_2) \\ p(x|y=1|x,\theta_1)p(y=1|\pi_1) &= p(y=2|x,\theta_2)p(y=2|\pi_2) \\ \log p(x|y=1|x,\theta_1) &= \log p(y=2|x,\theta_2) \qquad \text{Given } \pi_1 = \pi_2 = 0.5 \\ -\log\sigma_1 - \frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2 &= -\log\sigma_2 - \frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2 \end{aligned}$$

After plugging in the given μ and σ for both classes, the bound arrives at $-3.72 \le x \le 3.72$. It is a closed bound that anything in between is class 1 and everything outside is class 2.

(b) Now $\sigma_2^2 = 1$ instead of 10^6 this time. The bound is given by $x = 0.5$. Anything to the left is class 1 and vice versa. It makes sense because $x = 0.5$ is the average of both means.

9 Excerise 4.22

We first start with the posterior and its likelihood proportionality.

$$p(y=c|x,\theta_c) \propto p(x|y=c,\theta_c)p(y_c|\pi)$$

Given equal prior to all 3 classes, $p(y=1) = p(y=2) = p(y=3) = 1/3$. The classification task simplifies to whichever class has the highest multivariate normal density of observing the given data, $p(x|y=c,\theta_c)$, with respect to its parameters.
(a) Softmax $= [0.35, 0.31, 0.34] \rightarrow$ Class 1
(b) Softmax $= [0.347, 0.348, 0.305] \rightarrow$ Class 2, although really close

10 Excerise 4.23

We are given this data

index	x	label
1	67	m
2	79	m
3	71	m
4	68	f
5	67	f
6	60	f

(a) Fitting a Naive Bayes using MLE on a one-dimensional feature and a normal generative model, the resulting parameters are just the sample mean and variances. Hence,

$$\mu_m = 72.33, \sigma_m = 6.11, \pi_m = 0.5$$

$$\mu_f = 65.00, \sigma_m = 4.36, \pi_m = 0.5$$

with an equal likelihood prior. The general population has males and females distributed relatively evenly as well as the samples.

(b) To make a prediction on $p(y = m|x, \hat{\theta_m})$ with $x = 72$, we use the following equations.

$$p(y = m|x, \hat{\theta_m}) = \frac{p(x|y = m, \hat{\theta_m})\pi_m}{p(x|y = m, \hat{\theta_m})\pi_m + p(x|y = f, \hat{\theta_f})\pi_f}$$

The resulting probability is 0.721.

(c) Just use the QDA with a multivariate-normal generative model. So the likelihood becomes

$$p(\mathbf{x}|y = c, \theta) = \mathcal{N}(\mathbf{x}; \mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}})$$

Instead of estimating the variance, we can use the covariances matrix between two or more varialbes.