Gary Chen
804361027

Homework 5

EE M146
May 15, 2019

Problem 1

a)

$$\begin{aligned} P(A|C=0) &= 0.224 + 0.056 = 0.28 \\ P(B|C=0) &= 0.024 + 0.056 = 0.08 \\ P(A,B|C=0) &= 0.056 \end{aligned}$$

b)

$$\begin{aligned} P(A|C=1) &= 0.27 + 0.03 = 0.3 \\ P(B|C=1) &= 0.03 + 0.03 = 0.06 \\ P(A,B|C=1) &= 0.03 \end{aligned}$$

c) **Is A conditional independent of B given C?**

Conditional independence is given by:

$$\begin{aligned} (A \perp\!\!\!\perp B)|C \implies P(A,B|C) &= P(A|C)P(B|C) \\ 0.03 &= (0.06)(0.3) \\ 0.03 &\neq 0.0018 \end{aligned}$$

Thus, A conditional is not independent of B given C.

d) Total probability is defined as:

$$P(X) = \sum_{Y} P(X|Y)$$

$$\begin{aligned} P(A) &= P(A|C=0) + P(A|C=1) \\ &= 0.28 + 0.3 = 0.58 \\ P(B) &= P(B|C=0) + P(B|C=1) \\ &= 0.08 + 0.06 = 0.14 \\ P(A,B) &= P(A,B|C=0) + P(A,B|C=1) \\ &= 0.056 + 0.03 = 0.086 \end{aligned}$$

e) **Is A independent of B?**

Independence is defined as:

$$\begin{aligned} P(A,B) &= P(A)P(B) \\ 0.086 &= (0.58)(0.14) \\ 0.086 &\neq 0.0812 \end{aligned}$$

Thus, A is not independent of B

Problem 2

The pdf for two jointly Gaussian random variables X and Y is of the following form parameterized by the scalars m_1, m_2, σ_1, σ_2 and ρ_{XY}:

$$f_{XY}(z) = \frac{\exp\left\{-\frac{1}{2(1-\rho_{XY}^2)}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{XY}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}} \tag{1}$$

The pdf for multivariate jointly Gaussian random variable $Z \in \mathbb{R}^k$ is of the following form parameterized by $\mu \in \mathbb{R}^k$ and $\Sigma \in \mathbb{R}^{k\times k}$.

$$f_z(z) = \frac{\exp\left\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\}}{\sqrt{(2\pi)^k|\Sigma|}} \tag{2}$$

Suppose $Z = [X,Y]^T$, i.e., $z = [x,y]^T$, find μ, Σ^{-1} and Σ in terms of m_1, m_2, σ_1, σ_2 and ρ_{XY}.

We have $k = 2$, where we let $\Sigma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and $\mu = \begin{bmatrix} u \\ v \end{bmatrix}$. We equate equations 1 and 2 and start by looking at the denominators:

$$\begin{aligned} 2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2} &= \sqrt{(2\pi)^2|\Sigma|} \\ \sigma_1\sigma_2\sqrt{1-\rho_{XY}^2} &= \sqrt{|\Sigma|} \\ \sigma_1^2\sigma_2^2(1-\rho_{XY}^2) &= ad-bc \end{aligned} \tag{3}$$

Comparing the numerators:

$$\frac{1}{(1-\rho_{XY}^2)}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{XY}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right] = (z-\mu)^T\Sigma^{-1}(z-\mu)$$

We rewrite the right side to be:

$$\begin{aligned}(z-\mu)^T\Sigma^{-1}(z-\mu) &= \begin{bmatrix}x-u & y-v\end{bmatrix}\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}\begin{bmatrix}x-u\\ y-v\end{bmatrix}\\ &= \frac{1}{ad-bc}[(x-u)(d(x-u)-b(y-v))+(y-v)(-c(x-u)+a(y-v))]\\ &= \frac{1}{ad-bc}[d(x-u)^2-(b+c)(x-u)(y-v)+a(y-v)^2] \qquad (4)\end{aligned}$$

Then the left side:

$$\implies \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{XY}^2)}\left[\sigma_2^2(x-m_1)^2-2\rho_{XY}\sigma_1\sigma_2(x-m_1)(y-m_2)+\sigma_1^2(y-m_2)^2\right] \qquad (5)$$

By comparing equation 4 with 5, we can determine:

$$u=m_1, \qquad v=m_2, \qquad a=\sigma_1^2, \qquad d=\sigma_2^2$$

$$\sigma_1^2\sigma_2^2(1-\rho_{XY}^2) = ad-bc \qquad (6)$$
$$2\rho_{XY}\sigma_1\sigma_2 = b+c \qquad (7)$$

From equation 6 with the definitions of a and d:

$$\begin{aligned}\sigma_1^2\sigma_2^2(1-\rho_{XY}^2) &= \sigma_1^2\sigma_2^2-bc\\ bc &= \sigma_1^2\sigma_2^2(1-(1-\rho_{XY}^2))\\ bc &= \sigma_1^2\sigma_2^2\rho_{XY}^2\end{aligned}$$

Combining this result with equation 7, we find characteristic equations for b and c, yielding:

$$\begin{aligned}0 &= b^2-2\rho_{XY}\sigma_1\sigma_2 b+\sigma_1^2\sigma_2^2\rho_{XY}^2\\ b &= \frac{1}{2}\left(2\rho_{XY}\sigma_1\sigma_2\pm\sqrt{4\rho_{XY}^2\sigma_1^2\sigma_2^2-4\rho_{XY}^2\sigma_1^2\sigma_2^2}\right)\\ b=c &= \sigma_1\sigma_2\rho_{XY}\end{aligned}$$

Thus:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{XY} \\ \sigma_1\sigma_2\rho_{XY} & \sigma_2^2 \end{bmatrix}$$
$$\Sigma^{-1} = \frac{1}{1-\rho_{XY}^2}\begin{bmatrix} \sigma_2^2 & -\sigma_1\sigma_2\rho_{XY} \\ -\sigma_1\sigma_2\rho_{XY} & \sigma_1^2 \end{bmatrix}$$
$$\mu = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

Problem 3

Consider the jointly Gaussian random variables X and Y that have the following joint PDF:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)\right]$$

a) **Prove that Y is a Gaussian random variable by deriving its marginal PDF, $f_Y(y)$. Find the mean and variance of Y.**

The marginal PDF for y is found by integrating over x.

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)\right]dx \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right)^2-\frac{y^2}{2\sigma_Y^2}\right]dx \end{aligned}$$

Where $\int_{-\infty}^{\infty} e^{-a(x+b)^2}dx = \sqrt{\frac{\pi}{a}}$

$$\begin{aligned} &= \frac{\sqrt{2\pi\sigma_X^2(1-\rho^2)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right] \end{aligned}$$

Where we see this is a Gaussian with $\mathrm{E}[y] = 0$, $\mathrm{Var}[y] = \sigma_Y^2$.

b) **Prove that $f_{X|Y}(x|y)$ corresponds to another Gaussian random variable, then find its mean and variance.**

We may find the conditional PDF as:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{\sqrt{2\pi}\sigma_Y \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right)^2 - \frac{y^2}{2\sigma_Y^2}\right]}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right]} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_X}\exp\left[-\frac{(x-\rho y\frac{\sigma_X}{\sigma_Y})^2}{2\sigma_X^2(1-\rho^2)}\right] \end{aligned}$$

Where we see that we have a Gaussian with $\mathrm{E}[x|y] = \rho y\frac{\sigma_X}{\sigma_Y}$, and $\mathrm{Var}[x|y] = \sigma_X^2(1-\rho^2)$.

Problem 4

a) **Train the Naive Bayes classifier by calculating the maximum likelihood estimate of class priors and class conditional distributions. Namely, calculate the maximum likelihood estimate of the following: $P(G)$, and $P(X|G), X \in \{O, B, C, A\}$.**

From the chart, we find: $P(G=1) = 5/8$, $P(G=0) = 3/8$.

X	$P(X=1\|G=1)$	$P(X=0\|G=1)$	$P(X=1\|G=0)$	$P(X=0\|G=0)$
O	2/5	3/5	2/3	1/3
B	1/5	4/5	1/3	2/3
C	4/5	1/5	0	1
A	4/5	1/5	0	1

b) **For Sample #9 and #10, make the decision using**

$$\hat{G}_i = \underset{G_i\in\{0,1\}}{\operatorname{argmax}}\ P(G_i)P(O_i, B_i, C_i, A_i|G_i)$$

where O_i, B_i, C_i, and A_i are the feature values for the i-th sample.

Given the Naive Bayes assumption that all the variables are conditionally independent, we find the probability of a set of features to be:

$$P(G_i)P(O_i, B_i, C_i, A_i|G_i) = P(G_i)P(O_i|G_i)P(B_i|G_i)P(C_i|G_i)P(A_i|G_i)$$

We have the features of samples 9 and 10 to be:

Sample	O	B	C	A
9	0	1	0	1
10	1	1	1	1

For restaurant 9, we have the probabilities:

$$\begin{aligned}
P(G=0) &= P(G=0)P(O=0|G=0)P(B=1|G=0)P(C=0|G=0)P(A=1|G=0) \\
&= (3/8)(1/3)(1/3)(1)(0) \\
&= 0 \\
P(G=1) &= P(G=1)P(O=0|G=1)P(B=1|G=1)P(C=0|G=1)P(A=1|G=1) \\
&= (5/8)(3/5)(1/5)(1/5)(4/5) \\
&= 0.012
\end{aligned}$$

For restaurant 10, we have the probabilities:

$$\begin{aligned}
P(G=0) &= P(G=0)P(O=1|G=0)P(B=1|G=0)P(C=1|G=0)P(A=1|G=0) \\
&= (3/8)(2/3)(1/3)(0)(0) \\
&= 0 \\
P(G=1) &= P(G=1)P(O=1|G=1)P(B=1|G=1)P(C=1|G=1)P(A=1|G=1) \\
&= (5/8)(2/5)(1/5)(4/5)(4/5) \\
&= 0.032
\end{aligned}$$

We then determine that both restaurants 9 and 10 are good restaurants, as their probabilities of being good were higher than the probabilities of being bad.

Problem 5

Extend the Naive Bayes classifier for binary features i.e. $x_j \in \{0,1\}$ **to cases that are non-binary. Given a training set** $\{(x^{(i)}, y^{(i)}); i = \{1,\ldots,m\}$**, where** $x^{(i)} \in \{1,2,\ldots,s\}^n$ **and** $y^{(i)} \in \{0,1\}$**. Again, we model the label as a biased coin with** $\theta_0 = P(y^{(i)} = 0)$ **and** $1 - \theta_0 = P(y^{(i)} = 1)$**. We model each non-binary feature value** $x_j^{(i)}$ **as a biased dice for each class.**

$$
\begin{aligned}
&P(x_j = k|y=0) = \theta_{j,k|y=0} && k = 1, \ldots, s-1 \\
&P(x_j = s|y=0) = \theta_{j,s|y=0} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0} && \\
&P(x_j = k|y=1) = \theta_{j,k|y=1} && k = 1, \ldots, s-1 \\
&P(x_j = s|y=1) = \theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1} &&
\end{aligned}
$$

Notice that we do not model $P(x_j = s|y=0)$ and $P(x_j = s|y=1)$ directly. Instead we use the above equations to guarantee all probabilities for each class sum to 1.

a) **Using the Naive Bayes assumption, write down the joint probability of the data:**

$$P(x^{(1)}, \ldots, x^{(m)}, y^{(1)}, \ldots, y^{(m)})$$

in terms of the parameters θ_0, $\theta_{j,k|y=0}$, and $\theta + j, k|y = 1$.

Consider the joint probability of one pair in the training set:

$$
\begin{aligned}
P(x^{(i)}, y^{(i)}) =& P(y^{(i)}) \prod_{j=1}^{n} P(x_j^{(i)}|y^{(i)}) \\
=& \theta^{1[y^{(i)}=0]} (1-\theta_0)^{1[y^{(i)}=1]} \prod_{j=1}^{n} \left(\prod_{k=1}^{s-1} \theta_{j,k|y=0}^{1[x_j^{(i)}=k, y^{(i)}=0]} \theta_{j,k|y=1}^{1[x_j^{(i)}=k, y^{(i)}=1]} \right) \times \\
& \prod_{j=1}^{n} \left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0} \right)^{1[y^{(i)}=0] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0]} \times \\
& \prod_{j=1}^{n} \left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1} \right)^{1[y^{(i)}=1] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=1]}
\end{aligned}
$$

Extending this to all training sets:

$$
\begin{aligned}
P(x^{(1)},\ldots,x^{(m)},y^{(1)},\ldots,y^{(m)}) &= \prod_{i=1}^{m} P(x^{(i)},y^{(i)}) \\
&= \prod_{i=1}^{m} \left\{ \theta_0^{1[y^{(i)}=0]} (1-\theta_0)^{1[y^{(i)}=1]} \prod_{j=1}^{n} \left(\prod_{k=1}^{s-1} \theta_{j,k|y=0}^{1[x_j^{(i)}=k,y^{(i)}=0]} \theta_{j,k|y=1}^{1[x_j^{(i)}=k,y^{(i)}=1]} \right) \times \right. \\
&\quad \prod_{j=1}^{n} \left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0} \right)^{1[y^{(i)}=0] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k,y^{(i)}=0]} \times \\
&\quad \left. \prod_{j=1}^{n} \left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1} \right)^{1[y^{(i)}=1] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k,y^{(i)}=1]} \right\}
\end{aligned}
$$

b) **Maximizing the joint probability you get in (a) with respect to θ_0, $\theta_{j,k|y=0}$, and $\theta_{j,k|y=1}$. Write down your resulting θ_0, $\theta_{j,k|y=0}$, and $\theta_{j,k|y=1}$ and show intermediate steps. Comment on the meaning of your results.**

To find the maximum, we take the derivatives of the joint probabilities with respect to each value. First we define the log of the joint probability, $\mathcal{L} = \log(P(x^{(1)},\ldots,x^{(m)},y^{(1)},\ldots,y^{(m)}))$.

$$
\begin{aligned}
\mathcal{L} = \sum_{i=1}^{m} \Bigg\{ & 1[y^{(i)}=0]\log(\theta_0) + 1[y^{(i)}=1]\log(1-\theta_0) + \\
& \sum_{j=1}^{n}\sum_{k=1}^{s-1} \left[1[x_j^{(i)}=k,y^{(i)}=0]\log(\theta_{j,k|y=0}) + 1[x_j^{(i)}=k,y^{(i)}=1]\log(\theta_{j,k|y=1}) \right] + \\
& \sum_{j=1}^{n} \left[\left(1[y^{(i)}=0] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k,y^{(i)}=0] \right) \log\left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0} \right) \right] + \\
& \sum_{j=1}^{n} \left[\left(1[y^{(i)}=1] - \sum_{k=1}^{s-1} 1[x_j^{(i)}=k,y^{(i)}=1] \right) \log\left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1} \right) \right] \Bigg\}
\end{aligned}
$$

(a) θ_0:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_0} = 0 &= \sum_{i=1}^{m} \left[\frac{1[y^{(i)} = 0]}{\theta_0} - \frac{1[y^{(i)} = 1]}{(1-\theta_0)} \right] \\
&= \sum_{i=1}^{m} \left[1[y^{(i)} = 0](1-\theta_0) - 1[y^{(i)} = 1]\theta_0 \right] \\
&= \sum_{i=1}^{m} \left[1[y^{(i)} = 0] - \theta_0(1[y^{(i)} = 0] + 1[y^{(i)} = 1]) \right] \\
\theta_0 &= \frac{\sum_{i=1}^{m} 1[y^{(i)} = 0]}{\sum_{i=1}^{m} 1[y^{(i)} = 0] + 1[y^{(i)} = 1]} \\
&= \frac{1}{m} \sum_{i=1}^{m} 1[y^{(i)} = 0]
\end{aligned}$$

(b) $\theta_{j,k|y=0}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_{j,k|y=0}} = 0 &= \sum_{i=1}^{m} \left[\frac{1[x_j^{(i)} = k, y^{(i)} = 0]}{\theta_{j,k|y=0}} - \frac{1[y^{(i)} = 0] - \sum_{l=1}^{s-1} 1[x_j^{(i)} = l, y^{(i)} = 0]}{1 - \sum_{l=1}^{s-1} \theta_{j,l|y=0}} \right] \\
&= \frac{\sum_{i=1}^{m} 1[x_j^{(i)} = k, y^{(i)} = 0]}{\theta_{j,k|y=0}} - \frac{\sum_{i=1}^{m} 1[y^{(i)} = 0] - \sum_{i=1}^{m} \sum_{l=1}^{s-1} 1[x_j^{(i)} = l, y^{(i)} = 0]}{1 - \sum_{l=1}^{s-1} \theta_{j,l|y=0}} \\
\implies & \left(1 - \sum_{l=1}^{s-1} \theta_{j,l|y=0} \right) \sum_{i=1}^{m} 1[x_j^{(i)} = k, y^{(i)} = 0] \\
& = \theta_{j,k|y=0} \left(\sum_{i=1}^{m} 1[y^{(i)} = 0] - \sum_{i=1}^{m} \sum_{l=1}^{s-1} 1[x_j^{(i)} = k, y^{(i)} = 0] \right)
\end{aligned}$$

Summing over k yields:

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{k=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0] - \sum_{i=1}^{m}\sum_{k=1}^{s-1}\sum_{l=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0]\theta_{j,l|y=0} \\
=&\sum_{k=1}^{s-1}\sum_{i=1}^{m} 1[y^{(i)}=0]\theta_{j,k|y=0} - \sum_{k=1}^{s-1}\sum_{i=1}^{m}\sum_{l=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0]\theta_{j,k|y=0} \\
\implies &\sum_{i=1}^{m}\sum_{k=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0] = \sum_{k=1}^{s-1}\sum_{i=1}^{m} 1[y^{(i)}=0]\theta_{j,k|y=0} \qquad (8)
\end{aligned}$$

Equating like indexed summations, we find:

$$\begin{aligned}
&\sum_{k=1}^{s} 1[x_j^{(i)}=k, y^{(i)}=0] = 1[y^{(i)}=0] \\
&\sum_{k=1}^{s-1} 1[x_j^{(i)}=k, y^{(i)}=0] = 1[y^{(i)}=0] - 1[x_j^{(i)}=k, y^{(i)}=0] \qquad (9)
\end{aligned}$$

$$\sum_{k=1}^{s-1} \theta_{j,k|y=0} = 1 - \theta_{j,s|y=0} \qquad (10)$$

Putting equations 9 and 10 into 8, we get:

$$\begin{aligned}
\sum_{i=1}^{m}\Big(1[y^{(i)}=0] - 1[x_j^{(i)}=k, y^{(i)}=0]\Big) &= \sum_{i=1}^{m} 1[y^{(i)}=0](1-\theta_{j,s|y=0}) \\
\theta_{j,s|y=0} &= 1 - \frac{\sum_{i=1}^{m} 1[y^{(i)}=0] - \sum_{i=1}^{m} 1[x_j^{(i)}=k, y^{(i)}=0]}{\sum_{i=1}^{m} 1[y^{(i)}=0]} \\
\theta_{j,s|y=0} &= \frac{\sum_{i=1}^{m} 1[x_j^{(i)}=k, y^{(i)}=0]}{\sum_{i=1}^{m} 1[y^{(i)}=0]}
\end{aligned}$$

Using the same method, we similarly find:

$$\theta_{j,s|y=1} = \frac{\sum_{i=1}^{m} 1[x_j^{(i)} = k, y^{(i)} = 1]}{\sum_{i=1}^{m} 1[y^{(i)} = 1]}$$

These forms are effectively identical to the ones for the binary case, where the weights of each combination is still the proportion of total items with the given features. This is what we expect, for we are still functioning with the Naive Bayes assumption, and simply generalizing from binary features to an arbitrary number of features.