ECE M146 Homework 5 Solution
Introduction to Machine Learning Monday, May 13, 2019
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

1. Consider 3 random variables A,B and C with joint probabilities $P(A, B, C)$ listed in the following table.

	C=0		C=1	
	B=0	B=1	B=0	B=1
A=0	0.096	0.024	0.27	0.03
A=1	0.224	0.056	0.27	0.03

Calculate the distributions in (a) and (b). Answer the questions in (c), (d), and (e):

(a) $P(A|C = 0)$, $P(B|C = 0)$ and $P(A, B|C = 0)$.
Solution:

$$P(A|C=0) = \begin{cases} 0.3, A = 0 \\ 0.7, A = 1 \end{cases} \qquad P(B|C=0) = \begin{cases} 0.8, B = 0 \\ 0.2, B = 1 \end{cases}$$

$$P(A, B|C=0) = \begin{cases} 0.24, A = 0, B = 0 \\ 0.06, A = 0, B = 1 \\ 0.56, A = 1, B = 0 \\ 0.14, A = 1, B = 1 \end{cases}$$

(b) $P(A|C = 1)$, $P(B|C = 1)$ and $P(A, B|C = 1)$.
Solution:

$$P(A|C=1) = \begin{cases} 0.5, A = 0 \\ 0.5, A = 1 \end{cases} \qquad P(B|C=1) = \begin{cases} 0.9, B = 0 \\ 0.1, B = 1 \end{cases}$$

$$P(A, B|C=1) = \begin{cases} 0.45, A = 0, B = 0 \\ 0.05, A = 0, B = 1 \\ 0.45, A = 1, B = 0 \\ 0.05, A = 1, B = 1 \end{cases}$$

(c) Is A conditional independent of B given C?
Solution: Yes. From above, we can verify $P(A|C = 1)P(B|C = 1) = P(A, B|C = 1)$ and $P(A|C = 0)P(B|C = 0) = P(A, B|C = 0)$.

(d) $P(A)$, $P(B)$ and $P(A, B)$.
Solution:

$$P(A) = \begin{cases} 0.42, A = 0 \\ 0.58, A = 1 \end{cases} \qquad P(B) = \begin{cases} 0.86, B = 0 \\ 0.14, B = 1 \end{cases}$$

$$P(A,B) = \begin{cases} 0.366, A=0, B=0 \\ 0.034, A=0, B=1 \\ 0.494, A=1, B=0 \\ 0.086, A=1, B=1 \end{cases}$$

(e) Is A independent of B?
Solution: No. It is easy to verify that $P(A|C=1)P(B|C=1) \neq P(A,B|C=1)$ and $P(A|C=0)P(B|C=0) \neq P(A,B|C=0)$.

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}:

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

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 $\Sigma \in \mathbb{R}^{k\times k}$.

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

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}.

Solution: We find the following result by directly comparing (1) and (2):

$$\mu = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix},$$

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{XY}^1)}\begin{bmatrix} \sigma_1^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{XY}\sigma_1\sigma_2 \\ \rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

One can verify that by plugging the above expressions into (2), we get (1) back.

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.

Solution:

The marginal PDF of Y, $f_Y(y)$ is derived as follows:

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

To perform this integral, we need to complete a square inside the argument of the exponential.

$$\begin{aligned} \text{Exp\_arg} &= -\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - \frac{2\rho xy}{\sigma_X\sigma_Y}\right) \\ &= -\frac{1}{2(1-\rho^2)}\left(\left[\frac{x}{\sigma_X} - \frac{\rho y}{\sigma_Y}\right]^2 - \frac{\rho^2 y^2}{\sigma_Y^2} + \frac{y^2}{\sigma_Y^2}\right) \\ &= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X} - \frac{\rho y}{\sigma_Y}\right]^2 - \frac{1}{2(1-\rho^2)}\frac{(1-\rho^2)y^2}{\sigma_Y^2} \\ &= -\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x - \frac{\rho\sigma_X y}{\sigma_Y}\right]^2 - \frac{y^2}{2\sigma_Y^2}. \end{aligned}$$

Substituting this exponential argument in the integral of $f_Y(y)$ gives us:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right]\int_{x=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}}\exp\left[-\frac{\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2}{2\sigma_X^2(1-\rho^2)}\right]dx$$

The value of this integral is 1. Thus,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right], \quad -\infty < y < -\infty,$$

which proves that Y is a Gaussian random variable with mean 0 and variance σ_Y^2.

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

Solution:

The conditional PDF $f_{X|Y}(x|y)$ is derived as follows:

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

One more time, we operate on the exponential argument:

$$\begin{aligned}
\text{Exp_arg} &= -\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2+\frac{1}{2(1-\rho^2)}\left[\frac{\rho^2y^2}{\sigma_Y^2}-\frac{y^2}{\sigma_Y^2}\right]+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2-\frac{y^2}{2\sigma_Y^2}+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2\sigma_X^2(1-\rho^2)}\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2.
\end{aligned}$$

Consequently, we conclude that:

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

where $-\infty<x<\infty$. This proves that $f_{X|Y}(x|y)$ corresponds to another Gaussian random variable with mean $\rho\sigma_X y/\sigma_Y$, and variance $\sigma_X^2(1-\rho^2)$.

4. Let us revisit the restaurant selection problem in HW3. You are trying to choose between two restaurants (sample 9 and sample 10) to eat at. To do this, you will train a classifier based on your past experiences (sample 1-8). The features for each restaurants and your judgment on the goodness of sample 1-8 are summarized by the following chart.

Sample #	HasOutdoorSeating	HasBar	IsClean	HasGoodAtmosphere	IsGoodRestaurant
1	0	0	1	1	1
2	1	0	0	0	0
3	0	1	1	1	1
4	0	0	0	0	0
5	1	1	0	0	0
6	1	0	1	0	1
7	1	0	0	1	1
8	0	0	1	1	1
9	0	1	0	1	?
10	1	1	1	1	?

In this exercise, instead of a decision tree, you will use the Naive Bayes classifier to decide whether restaurant 9 and 10 are good or not. For clarity, we abbreviate the names of the features and label as follows: HasOutdoorSeating $\rightarrow O$, HasBar $\rightarrow B$, IsClean $\rightarrow C$, HasGoodAtmosphere $\rightarrow A$, and IsGoodRestaurant $\rightarrow G$.

(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 $P(X|G), X \in \{O, B, C, A\}$.
Solution: The maximum likelihood of class priors are just the relative frequency of each class. We therefore have:

$$P(G=0) = \frac{3}{8}, P(G=1)\frac{5}{8}.$$

The class conditional distribution can be estimated similarly by calculating the relative frequency of the features conditional on the class. We get:

$$\begin{aligned}
P(O=0|G=0) &= \frac{1}{3}, P(O=0|G=1) = \frac{3}{5};\\
P(B=0|G=0) &= \frac{2}{3}, P(B=0|G=1) = \frac{4}{5};\\
P(C=0|G=0) &= 1, P(C=0|G=1) = \frac{1}{5};\\
P(A=0|G=0) &= 1, P(A=0|G=1) = \frac{1}{5}.
\end{aligned}$$

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

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

where O_i, B_i, C_i, and A_i are the feature values for the i-th sample.
Solution: Using previous results, for $i = 9$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{1}{3} \times \frac{1}{3} \times 1 \times 0 = 0,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{3}{5}\times\frac{1}{5}\times\frac{1}{5}\times\frac{4}{5} > P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0).$$

We then decide $\hat{G}_9 = 1$.
For $i = 10$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{2}{3} \times \frac{1}{3} \times 0 \times 0 = 0,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{2}{5}\times\frac{1}{5}\times\frac{4}{5}\times\frac{4}{5} > P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0).$$

We then decide $\hat{G}_{10} = 1$.

5. In class, we learned a Naive Bayes classifier for binary feature values, i.e., $x_j \in 0, 1$ where we model the class conditional distribution to be Bernoulli. In this exercise, you are going to extend the result to the case where features that are non-binary.

   We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \cdots, m\}\}$, where $x^{(i)} \in \{1, 2, \cdots, s\}^n$ 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 $x_j^{(i)}$ (an element of $x^{(i)}$) as a biased dice for each class. This is parameterized by:

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

   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 (NB) assumption**, write down the joint probability of the data:

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

   in terms of the parameters θ_0, $\theta_{j,k|y=0}$ and $\theta_{j,k|y=1}$. You may find the indicator function $\mathbf{1}(\cdot)$ useful.
   Solution:

$$
\begin{aligned}
&P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)})\\
=&\prod_{i=1}^{m} P(x^{(i)}, y^{(i)})\\
=&\prod_{i=1}^{m} \theta_0^{\mathbf{1}(y^{(i)}=0)} (1-\theta_0)^{\mathbf{1}(y^{(i)}=1)} \prod_{j=1}^{n}\prod_{k=1}^{s} \theta_{j,k|y=0}^{\mathbf{1}(x_j^{(i)}=k \wedge y^{(i)}=0)} \theta_{j,k|y=1}^{\mathbf{1}(x_j^{(i)}=k \wedge y^{(i)}=1)}.
\end{aligned}
\tag{3}
$$

   (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. Use words to explain the meaning of your results.
   Solution: Take the negative log of Equation (1) and we get:

$$
\begin{aligned}
J(\theta_0,\theta_{j,k|y=0}, \theta_{j,k|y=1}) = -\sum_{i=1}^{m} \Bigg\{ &\mathbf{1}(y^{(i)} = 0) \log(\theta_0) + \mathbf{1}(y^{(i)} = 1) \log(1 - \theta_0)\\
&+ \sum_{j=1}^{n}\sum_{k=1}^{s} \Big[\mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0) \log(\theta_{j,k|y=0}) + \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 1) \log(\theta_{j,k|y=1})\Big] \Bigg\}.
\end{aligned}
$$

We first find θ_0 that minimize J.

$$\frac{\partial J}{\partial \theta_0} = -\frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0)}{\theta_0} + \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 1)}{1 - \theta_0}.$$

Setting the derivative to 0 we get

$$\theta_0 = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0)}{m}.$$

Next we find $\theta_{j,k|y=0}$ for a particular j and $k \neq s$. We first take the derivative with respect to $\theta_{j,k|y=0}$. Notice that in J, we also have $\theta_{j,s|y=0}$ that also depends on $\theta_{j,k|y=0}$.

$$\frac{\partial J}{\partial \theta_{j,k|y=0}} = -\frac{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\theta_{j,k|y=0}} + \frac{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = s \wedge y^{(i)} = 0)}{\theta_{j,s|y=0}}.$$

Setting the derivative to 0 we get

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = s \wedge y^{(i)} = 0)} \theta_{j,s|y=0}.$$

Using the above equation for all $k \neq s$ and $\theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1}$ we get:

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0)}.$$

Similarly, we have:

$$\theta_{j,k|y=1} = \frac{\sum_{i=1}^{m} \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 1)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 1)}.$$