ECE M146 Homework 5
Introduction to Machine Learning Thursday, May 2, 2019
Instructor: Lara Dolecek Due: Wednesday, May 13, 2019
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 $P(A, B|C = 0)$.
   (b) $P(A|C = 1)$, $P(B|C = 1)$ and $P(A, B|C = 1)$.
   (c) Is A conditional independent of B given C?
   (d) Compute $P(A)$, $P(B)$ and $P(A, B)$.
   (e) Is A independent of B?

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}.

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.

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

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 $P(X|G), X \in \{O, B, C, A\}$.

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

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

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

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:

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

   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.

   (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.