## 2 Questions

1. **Bayesian Decision:** Given a new point $\boldsymbol{x}$, you are asked to decide whether it is of class 1, class 2, or to reject. For simplicity of derivation, assume "*reject*" is of class 3.

   Let $\alpha_i$ is the action of deciding the input $\boldsymbol{x}$ is of class $i, i = 1, 2, 3$. Let $\lambda_{i,j}$ be the loss if you take action $\alpha_i$ but $\boldsymbol{x}$ is of class $j$, where $1 \le i, j \le 3$. We further assume that (a) $\lambda_{i,i} = 0$ for $i = 1, 2$. (b) $\lambda_{1,2} = 2$, (c) $\lambda_{2,1} = 3$, (d) $\lambda_{3,1} = 1$ and $\lambda_{3,2} = 2.5$.

   Answer the following questions:

   - Derive the expected loss $R(\alpha_i|\boldsymbol{x})$, for $i = 1, 2, 3$, in terms of the posterior probability $P(C_1|\boldsymbol{x})$.
   - For *all possible values* of $P(C_1|\boldsymbol{x})$, illustrate (or draw in a figure) your decision (e.g., when and how to do classification or and when to reject). Under what condition you will choose to "*reject*"?
   - If you are given the posterior $P(C_1|\boldsymbol{x}) = 0.735$, what is the proper action you need to make?
   - If you are given the posterior $P(C_2|\boldsymbol{x}) = 0.8$, what is the proper action you need to make?

2. **Statistical Sampling:** Assume there are 300,000 vacant apartments in Hong Kong and they have an average price of $10 million dollars with a standard deviation of $2.5 million dollars. Prof. John C.S. Lui hates the office politics in the university so he wants to retire and he needs to buy an apartment. He has seen 20 available apartments so far. What is the probability that the apartment John has seen so far is between $8.5 million dollars to $11 million dollars?

   Let $\Phi(x) = \text{Prob}(Z \le x)$, where $Z$ is a random variable which has a *standard normal distribution*. In other words, $\Phi$ is the cumulative distribution function of $Z$.

   Express your answer in terms of $\Phi$.

3. **Vapnik-Chervonenkis dimension:** We are given 4 points and each input point $\boldsymbol{x}$ has five features. All these five features are not correlated. Each input $\boldsymbol{x}$ can be classified as class 0 ($C_0$) or class 1 ($C_1$). You are given that all four inputs are in "*general positions*". Alice said that the VC dimension of a linear classifier on these points is 4 (e.g., # of points) while Bob said it is 5 (e.g., # of features). Who is correct? State your reason.

4. **Parametric Methods** You are given $N$ labeled points of training data and they belong to either class $C_1$ or $C_2$. Each of these points has $d$ features. Let say the first $d_1$ features are modeled as Gaussian distribution, the next $d_2$ features are modeled as Bernoulli distribution and the last $d_3$ features are modeled as exponential distribution, where $d = d_1 + d_2 + d_3$. In addition, you know the last $d_2$ and $d_3$ features are independent of all other $d - 1$ features.

   I will provide the probability density (or mass) functions of all related random variables.

   If $X$ is a $d-$dimension Gaussian random variable, its probability density function is

   $$P(X = \boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \text{exp}\left[-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}\right)\right]$$

   where $\boldsymbol{\Sigma}$ is the covariance matrix.

   If $X$ is a random variable with a Bernoulli distribution, it has a probability mass function of

   $$\text{P}[X = 1] = p \quad \text{and} \quad \text{P}[X = 0] = 1 - p.$$

   If $Y$ is a random variable with an exponential distribution, it has a probability density function of

   $$P(Y = y) = \lambda e^{-\lambda y} \qquad y \ge 0.$$

   Explain clearly the following questions:

(a) Express the likelihood (or probability) of having a point $\boldsymbol{x}$ having $d = d_1 + d_2 + d_3$ features with the proper $d$ feature values.

(b) Develop a procedure to estimate the maximum likelihood estimators of all parameters of these $d$ features so that you can find the unnormalized posterior probability and prior probability.

(c) Given a new point $\boldsymbol{x}$, explain how you can do the classification.

5. **Dimensionality reduction:** In class, we discussed the PCA method to reduce the dimension of the data points. That is, if the data point has $d$ dimension, we can reduce it to $k$ dimension (with $k < d$). The idea is to find $k$ vectors, $\boldsymbol{w}_i, i = 1, 2, ...k$ such that each $\boldsymbol{w}_i$ is of $d$ dimensional vector and we project all our training data points to these $k$ $\boldsymbol{w}_i$ so that each data point will be reduced to $k$ dimensions. One assumption we made during the derivation of PCA is that all $\boldsymbol{w}_i$ have to be orthogonal. Explain why we need such a requirement.

6. **Feature embedding:** In class, we also discussed an alternative dimensionality reduction method known as "*feature embedding*". Explain in detail:

(a) Under what condition we should use the PCA method and under what condition we should use feature embedding method?

(b) Assume you want to use feature embedding method to reduce the dimension of the training data points. Now you have a new input point $\boldsymbol{x}$ with $d$ features. Explain the procedure you need to apply to $\boldsymbol{x}$ so that you can reduce the dimension to $k$, where $k < d$, and at the same time, do the classification of $\boldsymbol{x}$. (For this question, which classification method you will use is irrelevant.)

7. **Expectation Maximization:** All visitors to a casino have to play the following game. Each visitor will randomly pick one of the three available dice, say $D_a$, $D_b$ and $D_c$, from a bag. The visitor does not know which dice he selected. For the selected dice, if the outcome of the toss is an odd number, then the casino wins, else the visitor wins. For an honest dice, you know that the probability of having an outcome from any element in the set $\{1, 2, 3, 4, 5, 6\}$ is $\frac{1}{6}$. However, you suspect that all these three dices are loaded (or dishonest). Now you have access to a winning/losing log of the past $N$ visitors to this casino (assume $N$ is a very large number). Answer the following:

(a) Design a procedure to provide an unbiased estimator of $P_k(i)$, where $k \in \{a, b, c\}$ and $i \in \{1, 2, 3, 4, 5, 6\}$, and $P_k(i)$ is the probability of flipping dice $D_k$ ($k = a, b, c$) and it will generate an outcome $i$, for $i \in \{1, 2, 3, 4, 5, 6\}$.

(b) The above setting assumes each visitor can select any of $D_a$, $D_b$ and $D_c$ dice with probability of $\frac{1}{3}$ each. Will your estimation procedure be the same if the probability of selecting $D_a$ is $\frac{1}{2}$, while the probability of selecting $D_b$ or $D_c$ is $\frac{1}{4}$. Explain your answer.

8. **Bayesian Learning:** You are given a Bayesian network as shown below. Assume all these seven events are Bernoulli random variables which can be either 0 or 1. For event $x$, if it is written as $x$, we assume it takes on value of 1, if it is written as $\tilde{x}$, we assume it takes on value of 0. Express the joint probability of

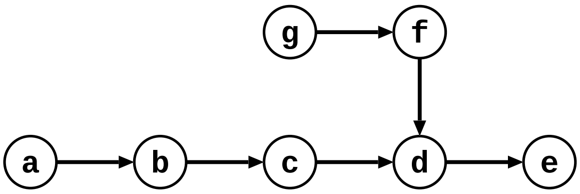

Figure 1: Bayesian network

$$P(a, b, c, d, e, f, g)$$