# CSE/ISYE 6740 Mid-term Exam

Le Song

## 1 Probability and Bayes' Rule [14 pts]

**(a)** A probability density function (pdf) is defined by

$$f(x,y) = \begin{cases} C(x+2y) & \text{if } 0 < y < 1 \text{ and } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

**(i)** Find the value of $C$ [3 pts].
Answer:

$$\int_0^1 \int_0^2 C(x+2y)dxdy = 4C = 1$$

Thus, $C = \frac{1}{4}$.

**(ii)** Find the marginal distribution of $X$ [2 pts].
Answer:

$$f_X(x) = \begin{cases} \int_0^1 \frac{1}{4}(x+2y)dy = \frac{1}{4}(x+1) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

**(iii)** Find the joint cumulative density function (cdf) of $X$ and $Y$ [2 pts].
Answer: The complete definition of $F_{XY}$ is:

$$F_{XY}(x,y) = \begin{cases} 0 & x \le 0 \text{ or } y \le 0 \\ x^2y/8 + y^2x/4 & 0 < x < 2 \text{ and } 0 < y < 1 \\ y/2 + y^2/2 & 2 \le x \text{ and } 0 < y < 1 \\ x^2/8 + x/4 & 0 < x < 2 \text{ and } 1 \le y \\ 1 & 2 \le x \text{ and } 1 \le y \end{cases}$$

**(b)** When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses "dots" and "dashes", which are known to occur in the proportion of 3:4. This means that for any given symbol,

$$P(\text{dot set}) = \frac{3}{7} \text{ and } P(\text{dash sent}) = \frac{4}{7}.$$

Suppose there is interference on the transmission line, and with probability $\frac{1}{8}$ a dot is mistakenly received as a dash, and vice versa. If we receive a dot, what is the probability that a dot was sent? That is, compute $P(\text{dot sent}|\text{dot received})$. [7 pts]

Answer: Using Bayes' Rule, we write

$$\begin{aligned} P(\text{dot sent}|\text{dot received}) &= P(\text{dot received}|\text{dot sent})P(\text{dot sent})/P(\text{dot received}) \\ &= \frac{(7/8) \times (3/7)}{(7/8) \times (3/7) + (1/8) \times (4/7)} = \frac{21}{25}. \end{aligned}$$

## 2 Maximum Likelihood [12 pts]

**(a)** The independent random variables $X_1, X_2, ..., X_n$ have the common distribution

$$P(X_i \leq x|\alpha,\beta) = \begin{cases} 0, & x < 0 \\ (\frac{x}{\beta})^\alpha, & 0 \leq x \leq \beta \\ 1, & x > \beta \end{cases}$$

where the parameters $\alpha$ and $\beta$ are positive. Find the MLEs of $\alpha$ and $\beta$. [7 pts]

After differentiating, we can get the P.D.F. as

$$P(X_i = x|\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{(\alpha-1)}, & 0 \leq x \leq \beta \\ 0, & otherwise \end{cases}$$

Let $x_{(n)}$ be the maximum value. For any fixed $\alpha$, the likelihood, $L(\alpha,\beta|x) = 0$ if $\beta < x_{(n)}$, and $L(\alpha,\beta|x)$ is a decreasing function of $\beta$ if $\beta \geq x_{(n)}$. Thus $x_{(n)}$ is the MLE of $\beta$

For the MLE of $\alpha$ calculate $\frac{\partial}{\partial\alpha}\Big[nlog\alpha - n\alpha log\beta + (\alpha-1)log\prod_i x_i\Big] = 0$ From this we get,

$$\alpha = \frac{n}{nlogx_{(n)} - log\prod_i x_i}$$

**(b)** Suppose that a particular gene occurs as one of the two alleles (A and a), where allele A has frequency $\theta$ in the population. That is a random copy of the gene is $A$ with probability $\theta$ and $a$ with probability $1-\theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

| genotype | $AA$ | $Aa$ | $aa$ |
|---|---|---|---|
| probability | $\theta^2$ | $2\theta(1-\theta)$ | $(1-\theta)^2$ |

Suppose we test a random sample of people and we find that $k_1$ are $AA$, $k_2$ are $Aa$, and $k_3$ are $aa$. Find the MLE of $\theta$ [5 pts]

likelihood = $(\theta)^{2k_1}(2\theta(1-\theta))^{k_2}(1-\theta)^{2k_3}$

log likelihood = constant $+2k_1ln(\theta) + k_2ln(\theta) + k_2ln(1-\theta) + 2k_3ln(1-\theta)$

MLE of $\theta = \frac{2k_1+k_2}{2k_1+2k_2+2k_3}$

# 3 Clustering [22 pts]

**(a)** Consider the below two datasets for the following questions:

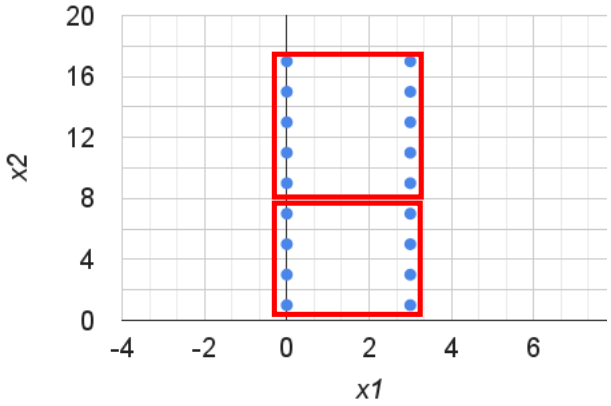


Figure 3a

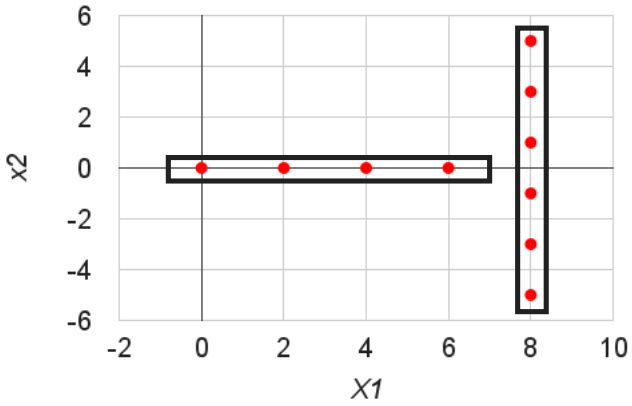


Figure 3b

**(i)** For Figure 3a, if k-means clustering ($k = 2$) is initialised with the two points whose coordinates are (3, 7) and (3,9), sketch the final clusters obtained (after the algorithm converges). [4 pts]

**(ii)** For Figure 3b, we will use spectral clustering. Our spectral clustering algorithm uses a neighbourhood graph obtained by connecting each point to its two nearest neighbours, and by weighting the resulting edges between points $x_i$ and $x_j$ by $W_{ij} = exp(-\|x_i - x_j\|)$. Sketch the clusters you will obtain using spectral clustering for Figure 3b. Provide justification for your answer. [4 pts]

**Answer:** The random walk induced by the weights can switch between the clusters in the figure in only two places, (8,-1) and (6,0). Since the weights decay with distance, the weights corresponding to transitions within clusters are higher than those going across in both places. The random walk would therefore tend to remain within the clusters indicated in the figure.

**(iii)** For figure 3b, can the solution obtained in the previous part using spectral clustering also be obtained by k-means clustering ($k = 2$) ? Justify your answer. [4 pts]

**Answer:** No. In the k-means algorithm points are assigned to the closest mean (cluster centroid). The centroids of the left and right clusters in the figure are (3,0) and (8,0), respectively. Point (6,0), for example, is closer to the right cluster centroid (8,0) and wouldnt be assigned to the left cluster. The two clusters in the figure therefore cannot be fixed points of the k-means algorithm.

**(b)** Consider a variant of *K-means* (called *L-means*) clustering algorithm:

**Input:** Integer $L$ and real-numbered threshold value $h$.

**Step 1:** Select $L$ instances (call them heads) and assign each of training instances to the cluster of the closest head.

**Step 2:** If the distance of training instance to its closest head is greater than the input threshold $h$, then this training instance becomes a new head. During the same assignment step, remaining points can be assigned to these new heads.

**Step 3:** After all the training instances have been assigned to cluster, new heads are calculated by taking mean of all instances for each cluster.

**Step 4:** Repeat the process until the cluster assignments do not change.

**(i)** Which of the two methods, *K-means* or *L-means*, will be better at dealing with outliers? Justify your answer. [4 pts]

**Answer:** *L*-means will be better, with an appropriate value for t. *K*-means will always try to put instances into the nearest cluster, which may cause clusters to shift "artificially" (that is, the outlier does not belong in them). By including a threshold $h$ there is a cutoff for reasonable values in a cluster and outliers are thus explicitly defined and singled out.

**(ii)** Given a dataset and a value $L$, let $h$ vary from 0 to a very large value. When does *L-means* produce more, the same number, or fewer clusters than *K-means*, assuming that the initial assignments are the same for both? When will the clusterings produced by both algorithms be identical? [6 pts]

**Answer:** As the initial assignments are assumed to be same, $K = L$ at the starting point (i.e. before any iteration of either algorithm is run.)

**Fewer clusters:** As we start with same $K$ and $L$ and the $L$-means algorithm does not have any step that performs reduction of clusters, it can never produce fewer number of clusters than $K$-means.

**More clusters:** As $h$ approaches 0, there will be more and more smaller clusters produced, reaching its maximum when $h = 0$ and every instance is its own cluster. Thus, for smaller $h$ values, $L$-means will produce more clusters than $K$-means algorithm.

**Same clusters:** As $h$ approaches $\infty$ or $h$ is big enough (larger than the largest distance between two instances in the dataset), $L$-means essentially becomes $K$-means clustering as the initial number of clusters will be retained.

# 4 Principal Component Analysis [16 pts]

**(a)** Suppose we have four points in 3-dimensional Euclidean space, namely $(1,1,1)$, $(-1,1,-1)$, $(1,-1,1)$, and $(-1,-1,-1)$.

**(i)** Find the first two principal components and provide brief reasoning of how you find them. [6 pts]

**Answer:** The four points are in the same plane, forming a rectangular (not a square!). The first principal component that preserves the most variance is $\frac{\sqrt{2}}{2}(1,0,1)$. The second principal component should be in the plane and orthogonal to the first one. Therefore, it's $(0,1,0)$.

**(ii)** What are the Cartesian coordinates of the given points in the two-dimensional principal subspace? [4 pts]

**Answer:** $(\sqrt{2},1),(-\sqrt{2},1),(\sqrt{2},-1),(-\sqrt{2},-1)$ if you don't scale the axes by the eigenvalues, as instructed in PRML.
$(1,1),(-1,1),(1,-1),(-1,-1)$ if you scale them, as instructed in lecture slides.

**(b)** Suppose $\mathbf{X} \in \mathbb{R}^{N\times D}$ is centered data with $N$ data points, whose $n^{th}$ row is given by $(\mathbf{x_n} - \bar{\mathbf{x}})^T$. Given $k$ eigenvectors, $\mathbf{u}_1,...,\mathbf{u}_k$, corresponding to the $k$ largest eigenvalues of the covariance matrix and let $\mathbf{U}_k = [\mathbf{u}_1,...,\mathbf{u}_k]$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{U}_K\mathbf{U}_K^T\mathbf{x}_i$. We wish to find the basis vector $\mathbf{u_{k+1}}$to maximize $\frac{1}{n}\sum_i(\mathbf{u}_{k+1}^T\tilde{\mathbf{x}}_i)^2$. Show that $\mathbf{u}_{k+1}$ is the eigenvector corresponding to the $(k+1)^{th}$ largest eigenvalue of the covariance matrix. [6 pts]

**Answer:**

$$
\begin{aligned}
J(\mathbf{u}_{k+1}) &= \frac{1}{n}\sum_i(\mathbf{u}_{k+1}^T\tilde{\mathbf{x}}_i)^2 && (1)\\
&= \frac{1}{n}\sum_i(\mathbf{u}_{k+1}^T\mathbf{x}_i - \mathbf{u}_{k+1}^T\mathbf{U}_K\mathbf{U}_K^T\mathbf{x}_i)^2 && (2)\\
&= \frac{1}{n}\sum_i(\mathbf{u}_{k+1}^T\mathbf{x}_i)^2 && (3)\\
&= \frac{1}{n}\sum_i\mathbf{u}_{k+1}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{u}_{k+1} && (4)\\
& && (5)
\end{aligned}
$$

subject to the constraints that $||\mathbf{u}_{k+1}|| = 1$ and $\mathbf{u}_{k+1} \perp \mathbf{u}_i, i = 1,...,k$. This is the $(k+1)^{th}$ eigenvector of the covariance matrix.

## 5 Expectation Maximization [20 pts]

The Student's distribution can be written as a Gaussian scale mixture, *i.e.*,

$$\mathcal{T}(x|\mu,\lambda,\nu) = \int \mathcal{N}\left(x|\mu,(\lambda\eta)^{-1}\right) Ga\left(\eta \,\middle|\, \frac{\nu}{2},\frac{\nu}{2}\right) d\eta,$$

marginal distribution

where $Ga\left(\eta \,\middle|\, \frac{\nu}{2},\frac{\nu}{2}\right) = \frac{1}{\Gamma(\frac{\nu}{2})}\frac{\nu}{2}^{\frac{\nu}{2}}\eta^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\eta\right)$ is the Gamma distribution, $\Gamma(\frac{\nu}{2})$ denotes the Gamma function. Given samples $\{x_i\}_{i=1}^N$ from the Student's distribution, we will use EM to estimate the parameters $\Theta = \{\mu,\lambda,\nu\}$.

**(a)** Write down the complete log-likelihood $\log p(x,\eta|\Theta)$ [4 pts].

joint distribution

**Answer:**

$$\sum_{i=1}^{N}\frac{1}{2}\log\frac{\lambda\eta_i}{2\pi} - \frac{\lambda\eta_i}{2}(x_i-\mu)^2 - \log\Gamma(\frac{\nu}{2}) + \frac{\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2}-1\right)\log\eta_i - \frac{\nu}{2}\eta_i$$

**(b)** Write down the $Q(\Theta,\Theta_{old})$ which the EM is optimizing [7 pts].
**Hint:** The posterior of latent variable is

写这些几把玩儿干啥

$$p(\eta|x,\Theta) = Ga(\eta|a,b)$$

where $a = \frac{\nu+1}{2}$ and $b = \frac{\nu+\lambda(x-\mu)^2}{2}$. We also compute $\mathbb{E}[\eta] = \frac{a}{b}$ and $\mathbb{E}\left[\log\eta\right] = \psi\left(a\right) - \log b$ as E-step for you.
**Answer:**

$$Q(\Theta,\Theta_{old}) = \sum_{i=1}^{N}\mathbb{E}_{p(\eta|x_i,\Theta_{old})}\left[\log p(x_i,\eta|\Theta)\right]$$

$$= \sum_{i=1}^{N}\frac{1}{2}\log\frac{\lambda\mathbb{E}[\eta_i]}{2\pi} - \frac{\lambda\mathbb{E}[\eta_i]}{2}(x_i-\mu)^2 - \log\Gamma(\frac{\nu}{2}) + \frac{\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2}-1\right)\mathbb{E}[\log\eta_i] - \frac{\nu}{2}\mathbb{E}[\eta_i]$$

**(c)** Derive the update rule for $\mu, \lambda$ [7 pts].
**Hint:** You are not required to derive the update rule for $\nu$.
**Answer:**

$$\frac{\partial Q}{\partial \mu} = 0 \implies \lambda \sum_{i=1}^{N} (x_i - \mu)\mathbb{E}[\eta_i] = 0 \implies \mu = \frac{\sum_{i=1}^{N} \mathbb{E}[\eta_i] x_i}{\sum_{i=1}^{N} \mathbb{E}[\eta_i]}$$

$$\frac{\partial Q}{\partial \lambda} = 0 \implies \frac{N}{2\lambda} - \frac{1}{2} \sum_{i=1}^{N} (x_i - u)^2 \mathbb{E}[\eta_i] = 0 \implies \lambda = \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)\mathbb{E}[\eta_i] \right)^{-1}$$

# 6 Information Theory [16 pts]

**(a)** Consider an $M$-state discrete random variable $x$, and use Jensen's inequality to show that the entropy of its distribution $p(x)$ satisfies $H[x] \leq \log(M)$. [4 pts]

**Answer:**
The entropy of an $M$-state discrete variable $x$ can be written in the form

$$H(x) = -\sum_{i=1}^{M} p(x_i) ln p(x_i) = \sum_{i=1}^{M} p(x_i) ln \frac{1}{p(x_i)}$$

The function $ln(x)$ is concave, so we can apply Jensen's inequality as follows:

$$H(x) \leq ln(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)}) = lnM$$

**(b)** Consider two binary variables $x$ and $y$ having the joint distribution given in the following table. Evaluate the following quantities:
(a)$H[x]$ (b) $H[y]$ (c) $H[y|x]$ (d) $H[x|y]$ (e) $H[x, y]$ (f) $I[x, y]$
Draw a diagram to show the relationship between these various quantities. [12 pts]

| | | $y$ = 0 | $y$ = 1 |
|---|---|---|---|
| $x$ | 0 | 1/4 | 1/4 |
| | 1 | 0 | 1/2 |

**Answer:**

According to the table, we can obtain the marginal probabilities by summation and the conditional probabilities by normalization as follows:

$p(x)$:

| 0 | 1 |
|---|---|
| 1/2 | 1/2 |

$p(y)$:

| 0 | 1 |
|---|---|
| 1/4 | 3/4 |

$p(x|y)$:

| x \ y | 0 | 1 |
|---|---|---|
| 0 | 1 | 1/3 |
| 1 | 0 | 2/3 |

$p(y|x)$:

| x \ y | 0 | 1 |
|---|---|---|
| 0 | 1/2 | 1/2 |
| 1 | 0 | 1 |

From the tables, together with the definitions:

$$H(x) = -\sum_{i} p(x_i) log p(x_i)$$

$$H(x|y) = -\sum_{i} \sum_{j} p(x_i, y_j) log p(x_i|y_j)$$

$$H(y|x) = -\sum_{i} \sum_{j} p(x_i, y_j) log p(y_j|x_i)$$

$$H(x, y) = H(x) + H(y|x) = H(y) + H(x|y)$$

$$I(x; y) = H(x) - H(x|y) = H(x) + H(y) - H(x, y)$$

we obtain the following results:

$$
\begin{aligned}
&(a) H(x) = -1/2log(1/2) - 1/2log(1/2) = 1 \\
&(b) H(y) = -1/4log(1/4) - 3/4log(3/4) = 2 - 3/4log3 \quad (0.8113) \\
&(c) H(y|x) = -1/4log1/2 - 1/4log1/2 - 0log0 - 1/2log1 = 1/2 \\
&(d) H(x|y) = -1/4log1 - 1/4log1/3 - 0log0 - 1/2log2/3 = 3/4log3 - 1/2 \quad (0.6887) \\
&(e) H(x,y) = 3/2 \\
&(f) I(x;y) = 3/2 - 3/4log3 \quad (0.3113)
\end{aligned}
$$

The corresponding diagram is shown in the figure.

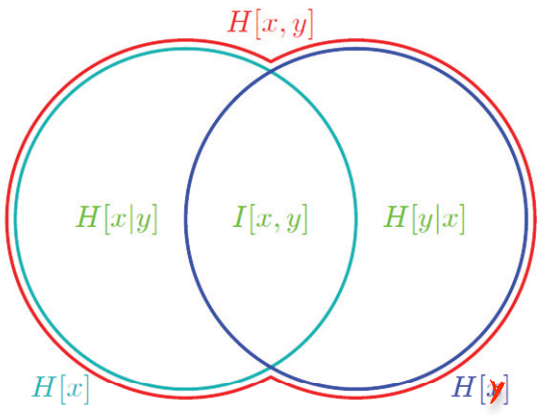


Figure 1: Set relationship among these quantities