CS 7641 CSE/ISYE 6740 Mid-term Exam Solutions

Le Song

10/13/2016

1 Expectation, Co-variance and Independence [14 pts]

(a) Suppose X, Y and Z are discrete random variables. Show that $\mathbb{E}_{(X,Y)\sim P(X,Y)}[XY] = \mathbb{E}_{Z\sim P(Z)}[\mathbb{E}_{(X,Y)\sim P(X,Y|Z)}[XY]]$. [4 pts]

Answer:

$$\begin{aligned}
RHS &= \mathbb{E}_{Z\sim P(Z)}[\mathbb{E}_{(X,Y)\sim P(X,Y|Z)}XY] \\
&= \sum_Z \{\sum_{(X,Y)} XYP(X,Y|Z)\}P(Z) \\
&= \sum_Z \{\sum_{(X,Y)} XYP(X,Y,Z)\} \\
&= \sum_{(X,Y)} XY \sum_Z P(X,Y,Z) \\
&= \sum_{(X,Y)} XYP(X,Y) \\
&= LHS
\end{aligned}$$

(b-c) Let $X \sim \mathcal{N}(0,1)$, and $Z \in \{\pm 1\}$ where $p(Z=-1) = p(Z=1) = 0.5$. X and Z are independent. Let $Y = ZX$. Note that X and Y are dependent as Y is function of X.

(b) Show that $Y \sim \mathcal{N}(0,1)$. [4 pts]

Answer:

$$\begin{aligned}
p(Y) &= p(Z=1)p(Y|Z=1) + p(Z=-1)p(Y|Z=-1) \\
&= 0.5p(X) + 0.5p(-X) \\
&= p(X) \qquad [X \sim \mathcal{N}(0,1), \text{ thus } p(X) = p(-X)]
\end{aligned}$$

Since the probability density functions of Y and X are the same, $Y \sim \mathcal{N}(0,1)$.

(c) Show that $cov[X, Y] = 0$. [6 pts]

Hint: Use definition of covariance and the statement that you proved in **(a)**. Statement in **(a)** holds analogously for continuous random variables and hence can be used here.

Answer:

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] \\ &= \mathbb{E}[X * XZ] \\ &= \mathbb{E}[X^2]\mathbb{E}[Z] \\ &= 1 * 0 \\ &= 0 \end{aligned}$$

2 Maximum Likelihood [12 pts]

(a) You are in a casino in Las Vegas, playing slot machine games. You can win $20 with machine A with probability of θ. Machine B has 4 times higher probability of winning, with just one forth of dividend (fair enough!). Suppose you played 10 times with either of machine A or B, and the result was as follows. What is the maximum likelihood estimation for θ? [6 pts]

Machine	Result	Machine	Result
A	Win	B	Win
A	Win	B	Lose
B	Lose	B	Lose
B	Lose	B	Win
B	Win	B	Lose

Answer: 1/8

(b) Uniform distribution [6 pts]

A uniform distribution in the range of $[\theta, \theta + 1]$ is given by

$$p(x|\theta) = \begin{cases} 1 & \theta \leq x \leq \theta + 1 \\ 0 & \text{otherwise} \end{cases}.$$

Suppose we observed 10 data points (as shown below), What is the maximum likelihood estimator of θ?

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.19	0.5	-0.35	0.44	-0.23	-0.5	0.45	0.23	-0.49	0.33

Answer:

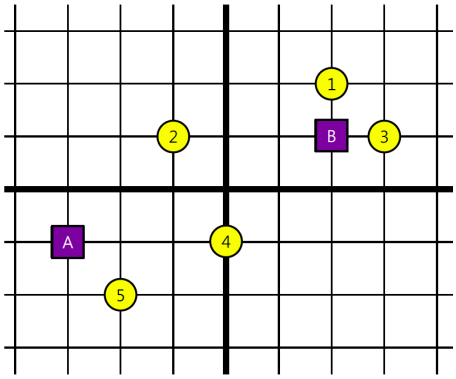
Condition: θ satisfying $\theta \leq x^1, x^2, ..., x^n \leq \theta + 1$

Solution: $\theta = -0.5$

3 Clustering [20 pts]

3.1 K-means [12 pts]

The following figure shows an intermediate state of running K-means. Circles are data points, and squares are centroid. We assume two-dimensional Euclidean space, where the point that two thick lines are crossing is the origin. For example, the centroid A is in (-3, 1), and the data point 2 is in (-1, 1).



(a) Which data points are belonging to cluster A and B, respectively? [4 pts]

- Cluster A: 2,5
- Cluster B: 1,3,4

(b) Suppose we proceed centroid recalculation step from the above figure. Where are those two centroids located after this step? [4 pts]

- Centroid A:(-1.5,-0.5)
- Centroid B:(5/3,2/3) or (1.667,0.667)

(c) After the step in (b), is the K-means iteration done? Mark one of the following. [4 pts]

- Yes / No: Answer is No.

3.2 Spectral Clustering [8 pts]

A graph on n nodes is defined by its affinity matrix $A \in \mathbb{R}^{n \times n}$, where $a_{ij} \geq 0$. Recall that the graph Laplacian is defined as $L = D - A$, where D is a diagonal matrix, with diagonal elements $d_{ii} = \sum_{j=1}^{n} a_{ij}$.

(a) Show that for any $x \in \mathbb{R}^n$, $x^\top L x = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(x_i - x_j)^2$. [8 pts]

Answer:

$$
\begin{aligned}
x^\top L x &= x^\top D x - x^\top A x \\
&= \sum_i d_{ii} x_i^2 - \sum_{i,j} a_{ij} x_i x_j \\
&= \frac{1}{2} \left[\sum_i d_{ii} x_i^2 + \sum_j d_{jj} x_j^2 - \sum_{i,j} 2 a_{ij} x_i x_j \right] \\
&= \frac{1}{2} \left[\sum_i \sum_j a_{ij} x_i^2 + \sum_j \sum_i a_{ij} x_j^2 - \sum_{i,j} 2 a_{ij} x_i x_j \right] \\
&= \frac{1}{2} \sum_i \sum_j a_{ij} \left(x_i^2 + x_j^2 - 2 x_i x_j \right) \\
&= \frac{1}{2} \sum_i \sum_j a_{ij} \left(x_i - x_j \right)^2
\end{aligned}
$$

4 Principal Component Analysis [14 pts]

Suppose we have four points in 3-dimensional Euclidean space, namely $(2, 0, 2)$, $(3, -1, 3)$, $(4, -2, 4)$, and $(5, -3, 5)$.

(a) Find the first principal component and provide brief reasoning of how you find it (*hint: visualize the points*). [7 pts]

Answer: $\frac{1}{\sqrt{3}}[1, -1, 1]^\top$. Note that to get full credits for this question, you need to mention that visually all the points fall in the straight line. And the direction of the straight line is the first principal component (3 points). Then you have to give the detailed calculation showing how to get the first principal component direction from the data points (4 points).

(b) When we reduce the dimensionality from 3 to 1 based on the principal component you found in (a), what is the reconstruction error? Provide brief reasoning. [7 pts] (*The reconstruction error for centered data points $x_1, \cdots, x_n$, i.e. $\sum_i x_i = 0$, using the first principal component u, with $||u|| = 1$, is $\frac{1}{n}\sum_{i=1}^{n} ||x_i - uu^\top x_i||^2$*)

Answer: 0 (3 points). Note that for the reasoning, you have to mention that all the points fall in the straight line. And the direction of this straight line is exactly the first principal component. Therefore, if you reduce the dimensionality from 3 to 1 based on the principal component, there will be no "information loss" (most important in your reasoning), and the reconstruction error is hence 0 (4 points).

5 Expectation Maximization [22 pts]

You have learned the Gaussian Mixture Model in class. For some discrete valued problems, like binary images, Bernoulli Mixture Model (BMM) is a good choice. In this model, we also have K components and parameters $\{\pi, \theta\}$. Each component i with prior π_i has a D-dimensional Bernoulli probability function parameterized by $\theta_i = [\theta_{i1}, \theta_{i2}, \ldots, \theta_{iD}]^T \in [0,1]^D$. Suppose we have N i.i.d samples $x_1, x_2, \ldots, x_N$. Given the component i, the likelihood of observing an instance $x \in \{0,1\}^D$ is

$$P(x|i) = \prod_{d=1}^{D} \theta_{id}^{x_d} (1-\theta_{id})^{(1-x_d)}$$

(a) Write down the log-likelihood $L(\pi, \theta)$ for N observations using BMM. If we use EM algorithm to find MLE, what are the latent variables? [4 pts]

Answer:

$$L(\pi, \theta) = \sum_{n=1}^{N} \log \left(\sum_{i=1}^{K} \pi_i \prod_{d=1}^{D} \theta_{id}^{x_{nd}} (1-\theta_{id})^{(1-x_{nd})} \right)$$

Latent variable: $z_n \in \{1, 2, \ldots, K\}$ which corresponds to the component the x_n belongs to. $n = 1, 2, \ldots, N$.

(b) Given π and θ, derive the lower bound of your log-likelihood, and write down the update rule for E-step. (*Hint: use Jensens inequality*) [8 pts]

Answer:

$$\begin{aligned} L(\pi, \theta) &= \sum_{i=1}^{N} \log \left(\sum_{z_n=1}^{K} P(x_n, z_n; \pi, \theta) \right) \\ &= \sum_{i=1}^{N} \log \left(\sum_{z_n=1}^{K} q(z_n) \frac{P(x_n, z_n; \pi, \theta)}{q(z_n)} \right) \\ &\geq \sum_{i=1}^{N} \sum_{z_n=1}^{K} q(z_n) \log \frac{P(x_n, z_n; \pi, \theta)}{q(z_n)} \end{aligned}$$

E-step:

$$z_{ni} = P(z_n = i) = P(i|x_n) = \frac{\pi_i \prod_{d=1}^{D} \theta_{id}^{x_{nd}} (1-\theta_{id})^{(1-x_{nd})}}{\sum_{i=1}^{K} \pi_i \prod_{d=1}^{D} \theta_{id}^{x_{nd}} (1-\theta_{id})^{(1-x_{nd})}}$$

(c) Write down the M-step which maximizes your lower bound written above. To receive full credits, you should provide the answer step by step. [10 pts]

Answer:

M-step:

$$\pi_i = \frac{\sum_{n=1}^{N} z_{ni}}{N}$$

$$\theta_i = \frac{\sum_{n=1}^{N} z_{ni} x_n}{\sum_{n=1}^{N} z_{ni}}$$

6 Information Theory [18 pts]

For a pair of discrete random variables X and Y (**that are identically distributed but not necessarily independent**) with the joint distribution $p(x, y)$, the *joint entropy* $H(X, Y)$ is defined as

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \tag{1}$$

which can also be expressed as

$$H(X,Y) = -\mathbb{E}[\log p(X,Y)] \tag{2}$$

Let X and Y take on values $x_1, x_2, ..., x_r$ and $y_1, y_2, ..., y_s$ respectively. Let Z also be a discrete random variable and $Z = X + Y$.

Hint: You are allowed to use any of the three statements that you proved in Homework 2 without proving them again.

(a) Show that $H(Z|X) = H(Y|X)$. Argue that when X, Y are independent, then $H(X) \leq H(Z)$ and $H(Y) \leq H(Z)$. Therefore, the addition of *independent* random variables add uncertainty. [5 pts]

Answer:

$$\begin{aligned} H(Z|X) &= \sum_{i=1}^{r}\sum_{j=1}^{s} p(x_i + y_j|x_i)p(x_i)log\frac{1}{p(x_i + y_j|x_i)} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{s} p(y_j|x_i)p(x_i)log\frac{1}{p(y_j|x_i)} \\ &= H(Y|X) \end{aligned}$$

As we know,

$$\begin{aligned} H(Z|X) &= H(Z,X) - H(X) \\ H(Z,X) - H(X) &= H(Y,X) - H(X) \iff H(Z,X) = H(Y,X) \end{aligned}$$

Also, $H(Y, X) \leq H(Y) + H(X)$. This holds when X and Y are independent.

$H(Z, X) = H(Y, X) = H(Y) + H(X)$

Since $H(Z, X) \leq H(Z) + H(X)$, we get
$H(Y) + H(X) \leq H(Z) + H(X)$

This implies $H(Y) \leq H(Z)$.
Similarly, one can prove that $H(X) \leq H(Z)$.

(b-d) Define a new quantity s as:

$$s = 1 - \frac{H(Y|X)}{H(X)} \tag{3}$$

(b) Show $s = \frac{I(X,Y)}{H(X)}$ [5 pts]

Answer:

$$\begin{aligned} I(X,Y) &= H(X) - H(Y|X) \\ r &= \frac{H(X)}{H(X)} - \frac{H(Y|X)}{H(X)} \\ &= \frac{H(X)}{H(X)} - \frac{H(X|Y)}{H(X)} \qquad \text{(because H(X) = H(Y) due to i.i.d.)} \\ &= \frac{I(X,Y)}{H(X)} \end{aligned}$$

(c) Show $0 \leq s \leq 1$ [4 pts]

Answer:

$$\begin{aligned} 0 \leq H(X|Y) &\leq H(X) \\ 0 \leq \frac{H(X|Y)}{H(X)} &\leq 1 \\ 0 \leq \frac{H(Y|X)}{H(X)} &\leq 1 \\ 0 \leq 1 - \frac{H(Y|X)}{H(X)} &\leq 1 \\ 0 \leq r &\leq 1 \end{aligned}$$

(d) When does the value of s=0 and when is the value of s=1? [4 pts]

Answer:

$$r = \frac{I(X,Y)}{H(X)}$$

so $r = 0$ when $I(X,Y) = 0$. This happens when X and Y are independent.

$$r = 1 - \frac{H(X|Y)}{H(X)}$$

so $r = 1$ when $H(X|Y) = 0$. This happens when X and Y are perfectly correlated.