ELL784: Introduction to Machine Learning

Major Test, Maximum marks: 24

(Answer all questions on the answer script. Read all section-specific instructions carefully.)

Section 1.

Instructions: Please write out your working as clearly and fully as possible, showing all steps used in obtaining the answers to each part of each question. Use clear and consistent notation, including denoting vector variables explicitly (*e.g.*, using an underbar as done in class). In the below questions bold font is used to denote vector variables.

1. Suppose you have been given a regression data set which contains noise ϵ drawn from the following distribution (density function):

$$p(\epsilon) = \begin{cases} \alpha & \text{if } -2 \leq \epsilon \leq 0 \\ \beta & \text{if } 0 < \epsilon \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Obtain an expression for the expected value of ϵ, as a function of α and β. Your expression should be simplified as much as possible. **[1.5]**

(b) What must be the relation between α and β if the noise is to be zero-mean? **[0.5]**

(c) Obtain the exact values of α and β, for the zero-mean case. **[1]**

(d) Suppose the given data set has been generated via the following model:

$$t = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M + \epsilon,$$

where the values of α and β are as given in part (c). I try fitting an order-M polynomial function $y(x)$ to the data, using least-squares regression. What is the value of $y(x)$ which will minimise the *expected loss*, in this case? **[1]**

2. Consider a simple example (adapted from Judea Pearl), where a burglar alarm at my house (A) can be set off by a burglary (B), or an earthquake (E), or a hurricane (H). I have two neighbours, John (J) and Mary (M), either of whom could call me in case the alarm goes off.

(a) Draw a Bayesian network to represent the causal relationships between these six binary random variables. **[2]**

(b) Write down the factorisation of the full joint distribution represented by your network. **[1]**

(c) Specify any three of the conditional independencies implied by this factorisation. **[1.5]**

(d) Give an instance in this network of the *explaining away* property, i.e., when a particular variable is observed then another pair of variables which were previously independent, become conditionally dependent. **[1]**

(e) Prove that if our model is such that the alarm always (deterministically) goes off whenever there is an earthquake:

$$P(A=1|B=1,E=1,H=1) = P(A=1|B=0,E=1,H=1)$$
$$= P(A=1|B=1,E=1,H=0) = P(A=1|B=0,E=1,H=0) = 1,$$

then $P(B=1|A=1,E=1) = P(B=1)$ and $P(H=1|A=1,E=1) = P(H=1)$, i.e., observing an earthquake provides a full explanation for the alarm. **[3]**

3. We wish to model the distribution of train waiting times at Rajiv Chowk Metro station. This station has 4 kinds of trains which stop at it: Blue line eastbound, Blue line westbound, Yellow line northbound, and Yellow line southbound.
Let us assume that for each type of train, the waiting time distribution is an exponential distribution. This means that if we denote the waiting time by x, then for the k^{th} type of train, the waiting time distribution is

$$p(x|\lambda_k) = \lambda_k e^{-\lambda_k x} \quad (x \geq 0),$$

where the parameter λ_k gives the inverse of the mean waiting time for train type k.
Further, let π_k denote the probability that a random passenger travelling from Rajiv Chowk intends to board the k^{th} type of train.

Now, supposing I visit Rajiv Chowk station and ask N random passengers how long they had to wait to get their desired train. This gives me a data set of observed waiting times; let us denote it $\mathbf{X} = \{x_1, x_2, ..., x_N\}$.

(a) I wish to model this situation using a mixture model, just like the GMMs discussed in class. So, corresponding to each x_n, I can have a latent variable vector $\mathbf{w}_n = (w_{n1}, w_{n2}, ..., w_{nK})^{\mathrm{T}}$. What should the dimension K of $\mathbf{w}_n$ be here? What will $\mathbf{w}_n$ denote? **[1]**

(b) Give an expression for $p(\mathbf{w}_n)$. **[0.5]**

(c) Give an expression for $p(x_n|\mathbf{w}_n)$. **[0.5]**

(d) Draw a graphical model corresponding to the mixture model we have just set up. **[1]**

(e) Let $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N\}$. Obtain an expression for the complete-data log likelihood, $\log p(\mathbf{X}, \mathbf{W})$. **[1.5]**

(f) Now I wish to estimate the the model parameters $\{\lambda_k\}$ and $\{\pi_k\}$ using the EM algorithm. Derive an expression for the E-step updates that will be involved in this, using appropriate notation. **[2]**

(g) Derive expressions for the M-step updates, again using appropriate notation consistent with the previous part. **[3]**

4. Give (depict visually) a simple example of a data set where a natural clustering exists, but it will be lost on doing dimensionality reduction via PCA. **[2]**