EEL709: Major Test

May 3, 2015

Maximum Marks: 25

Instructions: Please be careful and consistent with your notation. The examiner must be able to clearly understand what you have written. If you are introducing any notation not already defined in the question, you must explicitly define it. You should clearly distinguish between scalars and vectors, for instance by denoting the latter with an underbar. In the below questions, vectors are indicated via bold font.

1. The course EEL888: *Tough Machine Learning* is attended by some students who have done EEL709, and some who have not. In Minor I of EEL888, only 50% of the EEL709 students pass; and only 30% of the non-EEL709 students pass. Given that 60% of the entire EEL888 class are non-EEL709 students, what is the fraction of EEL709 students amongst those that pass Minor I?
(Please introduce appropriate random variables and explicitly use Bayes' theorem in your solution.) **[1.5]**

2. Let K be a function defined on pairs of English words, such that $K(x, y)$ counts the number of positions in which x and y have the same letter. For example, $K(\text{“}hello\text{”}, \text{“}bell\text{”}) = 3$, and $K(\text{“}time\text{”}, \text{“}signify\text{”}) = 1$. Is K a Mercer kernel? Prove either way. **[3]**

3. Normally, when using a neural network for classification, the cross-entropy error function is used. Here, we consider replacing the cross-entropy error with the mean-squared error $E(\mathbf{w}) = (1/2N)\sum_{n=1}^{N}(t_n - y_n)^2$.

   (a) In this case, derive the partial derivatives of $E(\mathbf{w})$ with respect to the final-layer network weights. (For simplicity you may assume a single training data point and a single output neuron with a sigmoidal activation function.) **[1.5]**

   (b) Compare with the cross-entropy error case. What is the difference? Does this tell you something about which error function is better? Give a specific example of a situation where one error function would outperform the other. **[2]**

4. Consider the probability distribution

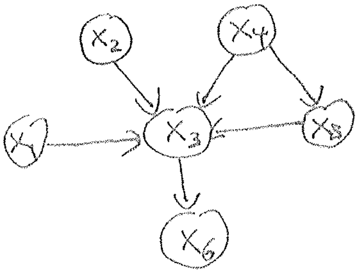
$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{1}{K} g_1(x_1, x_2) g_2(x_2, x_3) g_3(x_1, x_3) g_4(x_3, x_4, x_5) g_5(x_3, x_6) \qquad (1)$$

   where the g_i are non-negative functions and K is a normalising constant.

   (a) Draw the appropriate graphical model for this distribution. **[1]**

   (b) For each of the following marginal and conditional independence relations, state whether it is true or false, with reasons:
   (i) $x_1 \perp\!\!\!\perp x_5 | x_3$
   (ii) $x_2 \perp\!\!\!\perp x_6$
   (iii) $x_4 \perp\!\!\!\perp x_2 | \{x_3, x_6\}$
   (iv) $\{x_1, x_2, x_6\} \perp\!\!\!\perp \{x_4, x_5\} | x_3$ **[2]**

(c) Now consider this directed graphical model:



List (with justification) three conditional or marginal independencies in this directed graph which are *not* present in the distribution of Equation (1). **[1.5]**

5. This question seeks to model the distribution of waiting times at a Metro station. For this purpose, we will make the following assumptions:
(I) A Metro station has K different kinds of trains which stop there. (For instance, at Rajiv Chowk $K = 4$: Blue line eastbound, Blue line westbound, Yellow line northbound, and Yellow line southbound.)
(II) For each type of train, the waiting time distribution (with time rounded to a whole number of minutes) is a Poisson 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

$$p(x|\lambda_k) = \frac{\lambda_k^x e^{-\lambda_k}}{x!}. \quad (x \in \{0, 1, 2, 3, ...\}), \tag{2}$$

where the parameter λ_k gives the mean waiting time for train type k.
(III) The fraction of passengers arriving at a Metro station who intend to board the k^{th} type of train is a constant: call it π_k.

Now, supposing I ask N random passengers at my chosen Metro station 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 $\mathbf{X} = \{x_1, x_2, ..., x_N\}$.

Model this situation using an appropriate latent variable model. Make sure to clearly specify all your notation. Write down the joint distribution over the observed and latent variables, and hence the complete-data log likelihood. Derive the E and M step updates for the EM algorithm in order to estimate the model parameters $\{\lambda_k\}$ and $\{\pi_k\}$. **[6]**

6. Consider a setting where, over 3 successive days, when I get back home in the evening I observe the grass on my lawn to be either *wet* or *dry*. Because I work far away from home, I could not observe what the daytime weather was like on those 3 days, but I know that each day it was either *sunny* or *rainy*. Suppose also that I know the following: if the weather was rainy, the probability of the grass being wet in the evening is 0.9; if it was sunny, this probability is 0.2 (there is a sprinkler which the gardener switches on sometimes); if it is rainy one day, then the probability of rain the next day is 0.3; if it is sunny, then the probability of rain the next day is 0.1; and finally, the probability of it being rainy to start with is 0.1.

(a) Draw an appropriate Hidden Markov Model to represent this situation. Specify clearly your notation for random variables, and the corresponding initialisation, emission, and transition probabilities. **[1.5]**

(b) Suppose my actual observations over the 3 days are $\{dry, dry, wet\}$. Based on this and my specified model, I wish to estimate the probability that the weather on the 2nd day was sunny. Use the forward-backward algorithm to compute this. Referring to the notation used in class, which α or β value(s) do you need to evaluate for this purpose? Show the steps of the recursion involved in doing so. **[3]**

(c) Use your model to predict the probability of my finding the grass wet on day 4. **[2]**