CSC412 – Example Midterm Test

2004
Time: 60 minutes
Worth: 25%

Name: Student Number:

1 Short Answers

Complete the statements in the space given.

- When speaking of "i.i.d. data", "i.i.d." stands for

  ____.

- In a directed tree, each node (except the root) has exactly ______________________________.

- Maximum likelihood structure learning in fully observed tree models involves solving a ____________________

  ______________________________ problem, for example using ____________________

  algorithm.

- The key inequality used to lower bound the likelihood when deriving an EM algorithm is

  ____________________ inequality for convex functions.

- Consider a binary output y and some binary inputs $x_i, i = 1 \ldots P$.
  A "Noisy-OR" model for y with failure probabilities α_i corresponding to each input x_i is:

  $p(y = 1 | x_1 \ldots x_P, \alpha_1 \ldots \alpha_P) =$ ______________________________

- In factor analysis the covariance of the posterior distribution over the latent variable given an observation is

  ______________________________ of the observation.

2 Maximum Likelihood: Poisson

As a reminder, the *Poisson* distribution over a positive integer (count) random variable i is defined by the following mass function for a single positive parameter $\lambda > 0$.

$$p(i) = e^{-\lambda}\frac{\lambda^i}{i!} \quad i = 0, 1, \ldots, \infty$$

- Write the log likelihood of a dataset $i^1, \ldots, i^N$ in terms of λ.

- Find the maximum likelihood parameter λ in terms of $i^1, \ldots, i^N$.

- What are the sufficient statistics of a dataset $i^1, \ldots, i^N$ for the Poisson distribtion?

- Of course, the distribution must normalize properly, and thus $\sum_{i=0}^{\infty} p(i) = 1$.
  Use this to derive an expression for $\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$ in terms of λ.

- Calculate the mean and variance of i in terms of λ. (You may need the result: $\frac{\partial \lambda^i}{\partial \lambda} = \frac{i}{\lambda}\lambda^i$.)

3 Mixture of Gaussians Posteriors (20 points)

Consider a simple mixture of two one-dimensional Gaussians:

$$\begin{aligned} p(x) &= p(z=1)p(x|z=1) + p(z=2)p(x|z=2) \\ &= \alpha_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \alpha_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \end{aligned}$$

Where $\sigma_1, \sigma_2 > 0$; $0 < \alpha_1, \alpha_2 < 1$; $\alpha_1 + \alpha 2 = 1$.

- Using Bayes' rule, calculate the posterior $p(z=1|x)$.

- In general, there may be *more than one* value of x for which $p(z=1|x) = p(z=2|x)$ (i.e. the posteriors over the two components are equal).
Write a quadratic formula $ax^2 + bx + c = 0$ that must be satisfied when $p(z=1|x) = p(z=2|x)$, expressing a, b, c in terms of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha_1, \alpha_2$.

- If $\sigma_1^2 = \sigma_2^2$, and $\mu_1 \neq \mu_2$, what is the single value of x such that $p(z=1|x) = p(z=2|x)$?

- Under what conditions are there infinitely many values of x such that $p(z=1|x) = p(z=2|x)$?

- Under what conditions are there no values of x that make $p(z=1|x) = p(z=2|x)$?

- If $a \neq 0$ describe the conditions under which there is only only a single value of x at which the posteriors are equal.

4 EM Algorithm for Unobserved Naive Bayes

Consider the following "unobserved naive Bayes" model which has P *observed binary* variables $x_i \in \{0, 1\}$ $(i = 1 \dots P)$, and an *unobserved discrete* latent variable $z \in \{1, 2, \dots, K\}$.

$$p(z = k) = a_k$$
$$p(x_i = 1 | z = k) = b_{ik} \quad \forall i$$

Below you will derive the EM algorithm for maximum likelihood learning in this model.

- Write the complete data log likelihood for a dataset with N observations x_i^n and latent variables z^n, $i = 1 \dots P$, $n = 1 \dots N$.

- Calculate the marginal (incomplete) data log likelihood for some observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$.

- E-step: calculate the posterior $p(z = k | x_1, \dots, x_P)$ of the latent variable given the binary observations.

- Calculate the expected complete data log likelihood for the observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$ under a distribution $p(z^n = k | x_1^n \dots x_P^n) = q_k^n$.

- M-step: For a fixed q_k^n, and fixed observed data x_i^n, find the parameter settings a_k^* and b_{ik}^* which maximize the expected complete log likelihood. Be sure to enforce the normalization constraint $\sum_k a_k = 1$.

- Assume we have some observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$ and we want to fit this model using the EM algorithm. Using the results of the previous subquestions, write down the $E - step$ update for q_k^n and the $M - step$ updates for a_k and b_{ik}.
  Make sure the updates you write don't contain unspecified quantities.
  You should be able to turn your updates into code without any further derivations.