CS 725 Foundations of Machine Learning
Quiz-2; **6:30pm-9pm**, 09-Apr-2015

Roll No. ______________________

Important Instructors

- Please fill in your **roll number** in the blank above.
- Please write your answers in the space provided. There is **no separate answer sheet**. Answers written beyond the space provided may NOT be evaluated.
- It is recommended that you first write the solutions and answers in rough sheets. This will also help you to understand whether your answer will fit the space provided or you will need to shorten it by skipping less important details etc. Once you are convinced that the answer cannot be further improved, then only make a fair version of it in the space provided for the answer. You may take as many rough sheets as you need.
- Note that for most of the questions the markings will be binary based on the exactness of your final answer written in the space provided. So make sure your answers are legible, precise, and do not commit silly mistakes/typos in writing your final answers.
- Make sure you always simplify your answers to the extent possible. Technically correct answers, which can be simplified further will get you NO credit.
- Needless to say, a good way of verifying your answer is to check if it intuitively, geometrically, syntactically makes sense.
- This is a closed book exam. Electronic gadgets, including a calculator, are strictly NOT allowed.

Section 1. Analytical Questions

1. Nagarjunsagar-Srisailam Tiger Reserve is the largest Tiger reserve in India with a core area of 1,200 sq.km., located across two states: Andhra Pradesh and Telangana. Consider the problem of estimating the number of tigers in this reserve, denoted by m_t. Systematic search of every inch is impractical as the area is not only vast but also has extremely dense jungles. Besides, one may not want to disturb the tigers and the other fauna as much as possible. Hence a common sampling strategy happens to be that of capture-release-recapture survey:

   Hidden traps are laid at random points in the reserve. The traps are furnished with transmitters that signal a catch and each captured tiger is retrieved immediately. When m_p tigers have been caught, the traps are removed. Each of these m_p tigers are carefully sedated and marked with an ear tag. Then, all are released together back to the positions they were originally caught. Some time later[1], hidden traps are laid again, but at different random points on the island until m_s tigers have been caught and the number of tagged tigers is recorded, say as m_{sp}. Note that ALL the captured tigers are held in captivity until the m_s^{th} tiger has been caught.

   Given the numbers m_p, m_s, m_{sp}, your job is to estimate the total number of tigers m_t. In particular, attempt the following. which will help you in the estimation:

   (a) Write down an expression for the likelihood function (pmf in this case) for the Random variable, M_{sp}, representing the number of tagged tigers caught in the second round of traps in terms of the unknown parameter m_t (which is to be estimated). (2marks)

   (b) Consider now MLE of m_t. Write down this formal optimization problem, assuming valid values of m_t are numbers between $m_p +$

[1] By when it is assumed that the tagged tigers "mix completely" in the population.

$m_s - m_{sp}$ and $m_{\max}$. (0.5marks)

(c) Let the number of values between $m_p + m_s - m_{sp}$ and $m_{\max}$ be m. Provide an $O(\log_2(m))$ algorithm for solving the above MLE optimization problem. You will get no credit unless you justify the correctness of your algorithm. (3marks)

(d) If we now want to obtain a Bayesian estimate, then what will be a suitable conjugate prior? (0.5marks)

(e) Write down the posterior likelihood function for m_t. (1mark)

Typically, after obtaining the posterior of m_t, one repeats the whole experiment, now by using this posterior as the new prior. And this process is repeated until the final posterior is peaked enough[2].

Now consider an alternate survey procedure, where after tagging the tigers as above, a trap is placed at a random location and it is noted whether the tiger caught is a tagged one or not. After sometime of releasing this tiger, another trap is put at another random location and the same is noted. This is repeated for say m_s times. If the total

[2] so that one is confident that a MAP estimate with the final posterior is accurate enough.

no. of tagged tigers caught in these m_s independent traps is m_{sp}, then the MLE for m_t is going to be (1mark)

Ofcourse for obvious pragmatic reasons[3], nobody employs this alternate survey procedure.

2. Consider a density estimation problem for the dataset

$$\mathcal{D} = \{x_1, \ldots, x_m\} = \left\{ \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ ? \end{bmatrix}, \begin{bmatrix} ? \\ x_{32} \end{bmatrix}, \begin{bmatrix} ? \\ x_{42} \end{bmatrix}, \ldots, \begin{bmatrix} x_{m1} \\ x_{m2} \end{bmatrix} \right\},$$

where '?' represents a missing datum. Let the model be 2-d Gaussian. Your job is to perform an MLE for the parameters of the Gaussian, $\theta = (\mu_{2\times 1}, \Sigma_{2\times 2})$, using the EM algorithm. Make use of the following notation and answer the following questions. For the i^{th} example x_i, let x_{io} and x_{ih} denote the observed and missing feature values. For e.g., $x_{2o} = x_{21}$ and $x_{2h} = x_{22}$; whereas $x_{1o} = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$ and x_{2h} is null. Let us denote the Gaussian pdf at x_i and parameter θ by $f_\theta(x_{io}, x_{ih})$. Also let $\mathcal{X}_{ih}$ denote $\mathbb{R}$ if one of the values for x_i is missing, $\mathbb{R}^2$ if both are missing and null set $\{\}$ if both values are observed.

   (a) Write down the expression (using the above notation) for the likelihood of the training data. Your expression MUST involve the symbols $x_{ih}, \mathcal{X}_{ih}, f_\theta$. (2marks)

   (b) In the EM algorithm, at every iteration, we introduce an "auxiliary" distribution over the hidden variables at each example that facilitates us to obtain a lower bound that is a concave function of the parameters θ. At iteration t, let this auxiliary density function evaluated at $z \in \mathcal{X}_{ih}$ be given by $\lambda_i^t(z)$. Using this notation, write down the expression for the (concave) lower bound

[3] Obviously the experiment in the second survey will take a very long time compared to the first as now things are done 'sequentially' vs. 'parallelly' in the first.

for the likelihood of the training data[4]. (2marks)

(c) Let the parameter estimated at the t^{th} iteration of the EM algorithm be denoted by θ^t. Also, lets break down the parameters into the observed and hidden parts so that the likelihood at $\theta^t = (\mu_o^t, \mu_h^t, \Sigma_{oo}^t, \Sigma_{ov}^t, \Sigma_{vv}^t)$ can be written as

$$f_{\theta^t}(x_{io}, x_{ih}) = \frac{e^{-\frac{1}{2}\left(\left(x_{io}-\mu_o^t\right)^\top \Sigma_{oo}^t \left(x_{io}-\mu_o^t\right)+\left(x_{ih}-\mu_h^t\right)^\top \Sigma_{hh}^t \left(x_{ih}-\mu_h^t\right)+2\left(x_{io}-\mu_o^t\right)^\top \Sigma_{oh}^t \left(x_{ih}-\mu_h^t\right)\right)}}{2\pi\sqrt{det\left(\begin{bmatrix} \Sigma_{oo}^t & \Sigma_{oh}^t \\ (\Sigma_{oh}^t)^\top & \Sigma_{hh}^t \end{bmatrix}\right)}}$$

Write down an expression[5] for the "ideal" auxiliary distribution in terms of $(\mu_o^{t-1}, \mu_h^{t-1}, \Sigma_{oo}^{t-1}, \Sigma_{ov}^{t-1}, \Sigma_{vv}^{t-1})$:
$\lambda_i^t(z) =$ (4marks)

(d) In terms of $(\mu_o^{t-1}, \mu_h^{t-1}, \Sigma_{oo}^{t-1}, \Sigma_{ov}^{t-1}, \Sigma_{vv}^{t-1})$ and x_{io} write down the

[4] Recall that log is a concave function and hence satisfies the so called Jensen's inequality: $\log(\mathbb{E}[Z]) \geq \mathbb{E}[\log(Z)]$, where Z is *any* random variable; provided the involved expectations exist.

[5] Your expression must NOT involve the symbol f (used for likelihood function).

update equations for μ_h^t and μ_o^t. (4marks)

Realize that the EM algorithm does NOT simply replace a missing value at a feature with the sample mean of the same feature at observed positions.

3. It is proposed to evaluate the popularity of Indian celebrities by measuring the popularity of their YouTube channels. One way to measure popularity of a channel is by simply modeling the distribution of: the sum of number of "likes" and "dislikes" in its videos. Higher the mean of these sums, higher is the popularity of the celebrity. A naive way to do this is to model each celebrity/channel by a Gaussian distribution or by a Gaussian likelihood and a suitable conjugate prior. However, since all the channels belong to a particular community, Indians, there will be latent factors that connect/tie all of them[6]. Such factors should be taken into account especially if the number of videos in each channel are less. One way to connect/tie these multiple models is by using a common prior[7].

   In summary, here is the description of the model: for the j^{th} celebrity, the model is Gaussian with mean Θ_j and variance σ^2. The mean Θ_j is what finally has to be estimated to decide who is popular. Assume that the variance σ^2 is known. Now, the key modeling step is: we assume each Θ_j to come from a common Gaussian prior with mean μ and variance τ^2. Along with Θ_j, the hyper-parameters μ and τ^2 are to be estimated. Assume that we employ maximum marginal likelihood for hyper-parameter estimation and then using these estimates for the hyper-parameters, we perform a MAP estimate for the Θ_j. Derive the final simplified formula for the estimate of Θ_j in terms of training data

[6] For eg. perhaps all Indian channels get less viewership than say US channels etc.

[7] Multi-task learning via Hierarchical Bayes method.

and σ^2. (10marks)