UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11073 MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Thursday 16 th May 2013

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: B. Franke
External Examiners: T. Attwood, R. Connor, R. Cooper, S. Denham, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **You MUST answer this question.**

(a) You receive one data item N_1 which is the sum of N Bernoulli random variables with Bernoulli parameter p. You know N and wish to infer p.

i. Give the equation for log likelihood in this model. [*3 marks*]

ii. Suppose you were using a Bayesian approach. What would be a reasonable prior on p? [*2 marks*]

iii. Give a specific equation for a function that, if maximized, would yield the MAP parameter estimate of p under your prior. [*2 marks*]

iv. What is the posterior distribution for p under your prior? [*2 marks*]

(b) Suppose that two people, Alfred and Biff, agree to play the following game. Alfred has two coins in a bag. The first is regular fair coin. The second is a weighted coin that comes up heads with some probability p. Alfred chooses one coin and flips it 10 times in front of Biff. Biff's goal is to figure out which coin Alfred has selected. Describe in detail a Bayesian procedure for this problem, including any necessary equations. [*7 marks*]

(c) Consider the following four distributions over the real line: [*4 marks*]

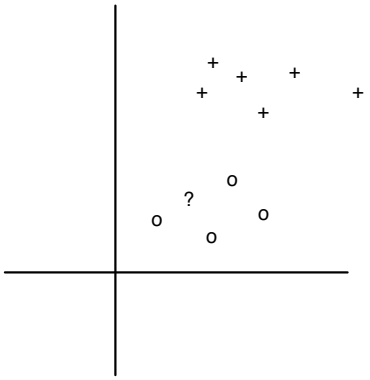
$$p_1(x) = \mathcal{N}(x; -2, 1)$$
$$p_2(x) = \mathcal{N}(x; -0.5, 1)$$
$$p_3(x) = \mathcal{N}(x; 0.5, 1)$$
$$p_4(x) = \mathcal{N}(x; 2.5, 1),$$

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes the pdf for the normal distribution with mean μ and variance σ^2. Which pair of distributions have the highest KL divergence? Which pair have the lowest KL divergence?

QUESTION CONTINUES ON NEXT PAGE

(d) Consider a binary classification problem with two features. The training data are shown in the following figure.

In this figure, the crosses mark the training points in class 0, and the circles the training points in class 1. In addition, there is one test point whose label is unknown. This is denoted by ?.



Suppose that you train a logistic regression classifier on this data set using maximum likelihood (ML).

i. Sketch the figure and draw the decision boundary that you think would result from training. Do not attempt to calculate the ML parameters; simply use your intuition. [*2 marks*]

ii. Consider the point marked ?. Let us denote the feature vector for this point by $\mathbf{x}$, and its true label by y. What is the probability $p(y = 1|\mathbf{x})$ according to the ML parameter estimates? Explain your answer. [*3 marks*]

2. For this problem, suppose that you are taking measurements x of an unknown one-dimensional quantity μ. Your equipment is unreliable. Most of the time, your equipment returns the value μ plus some Gaussian noise with mean zero and standard deviation 1. A small percentage of the time, however, your equipment malfunctions, and simply returns a Gaussian value with mean zero and standard deviation 1. The equipment returns no indication of whether it has malfunctioned or not. You wish to infer the value μ.

   (a) Write down the log likelihood for this problem. What are the parameters of this model? [*5 marks*]

   (b) Do you expect the log likelihood to be unimodal? [*2 marks*]

   (c) You decide to use the EM algorithm to estimate μ. Describe how the EM algorithm works, and give the specific EM update equations for this model. [*8 marks*]

   (d) A colleague suggests that instead of using EM, you place a prior on μ and use a Laplace approximation. Give a general description of how the Laplace approximation is computed. (You do not need to provide any specific equations for this particular situation.) [*7 marks*]

   (e) Would the Laplace approximation be a good idea in this case? [*3 marks*]

3. (a) Write out the procedure for Metropolis-Hastings, including the equation for the acceptance probability. Show that Gibbs sampling can be viewed as Metropolis-Hastings with a proposal distribution that is always accepted. [*7 marks*]

   (b) Recall that the pdf for a beta distribution is [*7 marks*]

$$p(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

   where $B(a,b)$ denotes the beta function.

   Suppose that you wish to sample from a beta distribution. You decide to use importance sampling to do so, with a uniform distribution as the proposal distribution. Describe how this importance sampler would work, giving specific equations. How would you use these samples to estimate the mean of x?

   You may assume that you have access to a library function that computes the Beta function $B(a,b)$ numerically, so you do not need to worry about how that is computed.

   (c) In a recent research paper presented at the ICML 2012 conference, Le and collaborators trained a nine-layer neural network on a data set of 10 million images. Suppose that they had been trying to solve a classification task, in which each of the 10 million images was labelled as belonging to either a positive class or a negative class. Which optimization algorithm would you prefer during training? Justify your answer. [*3 marks*]

   (d) Consider a binary classification problem with D continuous features for each training instance. The following questions consider a class-conditional Gaussian classifier. Let y refer to the class label of a particular instance, and $\mathbf{x}$ to the feature vector.

      i. Describe in equations the models used by the class-conditional Gaussian classifier. Clearly indicate the parameters of the model. [*3 marks*]

      ii. Suppose that you believed that the naive Bayes assumption was likely to be reasonable for this problem. How specifically could you incorporate this assumption into the classifier? [*2 marks*]

      iii. Consider the decision boundary for this classifier. Is it linear in the features $\mathbf{x}$? Why or why not? [*3 marks*]