UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

MACHINE LEARNING AND PATTERN RECOGNITION

Friday 20th May 2011

09:30 to 11:30

MSc Courses

Convener: C. Stirling
External Examiners: T. Attwood, R. Connor, R. Cooper, D. Marshall, M. Richardson

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

You MUST answer this question.

1. (a) Suppose you have a dataset $\mathcal{D} = \{(\mathbf{x}^n, \mathbf{y}^n), n = 1, 2, \ldots, N\}$, where we call the $\mathbf{y}^n$ terms the targets. Give a definition of the task of supervised learning for such a dataset. How can a supervised learning task be solved by building a model $P(\mathbf{x}|\mathbf{y})$? Specifically, what other part of the model is needed, and what equation is used to solve the task? [*4 marks*]

   (b) Naive Bayes for binary data taking values 0 and 1 uses a Bernoulli model for the data items. The parameter of the Bernoulli distribution is $p = P(1)$. Write out the Bernoulli likelihood for a binary dataset $\mathcal{D}$ of size N, and also the log-likelihood and hence show that the maximum likelihood value for p given data $\mathcal{D}$ corresponds to the proportion of 1s in the dataset. [*6 marks*]

   (c) The Beta distribution

$$P(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

   is a conjugate distribution to the Bernoulli. Explain what conjugacy means. Suppose you have counts of 20 and 10 for the number of 1s and the number of 0s (respectively) in a dataset. Using a Bernoulli likelihood and a beta prior with parameters $\alpha = \beta = 2$, write out the form of the posterior distribution for that data. [*4 marks*]

   (d) Suppose you wish to predict the apartment prices in Manhattan (which has a grid street structure) using information about the geographic location and size of the flat. You decide to use regression with a linear parameter model having geographically local features. Briefly discuss what form of features you would choose. [*3 marks*]

   (e) Describe the steps you would take to compute the principal component of a dataset $\{\mathbf{x}^1, \mathbf{x}_2, \ldots, \mathbf{x}^N\}$, where N denotes the total number of datapoints. Do so in such a way that someone else could recreate the process: be specific about whether you are doing operations over data points or over attributes. [*5 marks*]

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

2. (a) Define the Hessian matrix H of a function $f(\boldsymbol{\theta})$ of vector $\boldsymbol{\theta}$. Suppose we have a model $P(\mathcal{D}|\boldsymbol{\theta})$, where $\mathcal{D}$ denotes the data and $\boldsymbol{\theta}$ are parameters. Suppose you also have a prior distribution $P(\boldsymbol{\theta})$ for the parameters $\boldsymbol{\theta}$. Define a simple expression for the maximum posterior parameter $\boldsymbol{\theta}^*$ in terms of the prior and likelihood. Write out the logarithm of the posterior distribution, and give its Taylor expansion about $\boldsymbol{\theta}$ to second order. How is this used to determine the Laplace approximation? [*5 marks*]

   (b) Explain how gradient ascent optimization of a function $f(\boldsymbol{\theta})$ works. Explain the problems with gradient ascent that are overcome with conjugate gradient methods. [*4 marks*]

   (c) Suppose in a regression problem, you believed that each data point came from one of three possible linear functions of the features. However you did not know which function each data point belonged to. By considering the correspondence to Gaussian mixture models, describe how you could solve this process using an iterative procedure. [*6 marks*]

   (d) Define the variational approximation to the posterior distribution $P(\theta|\mathcal{D}, h)$, where h is a set of hyper-parameters. Show, using the fact $KL(Q||P) \geq 0$ for two distributions Q and P, that the marginal likelihood $P(\mathcal{D}|h)$ can be lower bounded through the use of the KL divergence, and that the variational approximation maximises that lower bound. [*7 marks*]

   (e) Why, in general, does a variational approximation to the posterior make a bad proposal distribution for importance sampling the posterior distribution? [*3 marks*]

3. You are developing a system that analyzes risk of hospitalisation from a primary healthcare dataset regarding visits to the doctor. The dataset includes the attributes for age, reason for visit (i.e. illness/issue), number of previous visits in two years, perceived severity (1 to 10), along with whether the patient was hospitalised.

   (a) The data for each customer is an array with each item following the order listed in the previous paragraph. For example, here is one data point from the database

   $$(44, \text{diabetes}, 6, 10, 1)$$

   To make a classifier you need to convert each entry into a numerical representation. However, you recognise that there is a potential problem with how the 'diabetes' attribute is represented. What simple but reasonable approach could you choose to encode this? [*2 marks*]

   (b) After you do your encoding, you are concerned that this is very high dimensional data. Explain why dimension reduction using PCA would be inappropriate. [*2 marks*]

   (c) You wish to use a neural network to predict the probability of hospitalisation for a patient presenting at his/her doctor. Explain how you would process the data, design, initialise, train, and validate a neural network for this data. Make sure you clearly state how you would choose between possible network structures, and how you would avoid local minima issues, and what use a prior distribution may be. How would you report your expected error? [*8 marks*]

   (d) A friend wishes to use this to see whether he is likely to be hospitalised. What is sample selection bias? Why is sample selection bias a problem for this application of your network? [*4 marks*]

   (e) You decide that you wish to sample from the posterior distribution for your network to improve its predictions. You decide to use Hamiltonian Monte-Carlo. Explain why this was a better choice than Metropolis-Hastings. [*3 marks*]

   (f) You consider the possibility of using a Gaussian process that must have a positive definite Kernel. Define what makes a Kernel positive definite. If a Kernel $K(\mathbf{x}.\mathbf{x}')$ is positive definite, show that the Kernel matrix $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) + \alpha\delta_{ij}$ is also positive definite. What does the $\alpha\delta_{ij}$ term correspond to in terms of a Gaussian process prediction? [*4 marks*]

   (g) Let $y(\mathbf{x})$ be modelled using a Gaussian process. Explain why using this model by equating $y(\mathbf{x})$ to 1 if the patient is hospitalised, and 0 if the patient is not, is the wrong thing to do. What should you do instead? [*2 marks*]