UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 10)

MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Friday 1 May 2009

14:30 to 16:30

Year 4 Courses

Convener: D K Arvind
External Examiners: A Frisch, J Gurd

MSc Courses

Convener: A Smaill
External Examiners: R Connor, R Cooper, D Marshall, I Marshall

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

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

(a) Why would you use 1 of M encoding? When should you not use 1 of M encoding? [*4 marks*]

(b) Write out the form of the likelihood for a Naive Bayes model for data $\{(\mathbf{x}^n, y^n), n = 1, 2, \ldots, N\}$, where $y^n \in \{0, 1\}$ is a binary target variable, and $\mathbf{x}$ is a vector of binary attributes. By writing out the log likelihood and taking derivatives, derive the maximum likelihood solution for the prior term ($P(y = 1)$) in the Naive Bayes model above. [*8 marks*]

(c) What does it mean to say a prior distribution is a conjugate prior?

Given a problem with dataset $\mathcal{D}$ and parameters $\boldsymbol{\theta}$, define the marginal likelihood (or evidence).

Why is the marginal likelihood useful? Describe how the 'Google Sets' algorithm uses the marginal likelihood to obtain a ranking of the best additional elements to add to a given set. [*9 marks*]

(d) You are given the data

$$\mathbf{x}_1 = (1, -1, 5)^T,\ \mathbf{x}_2 = (5, 1, 1)^T,\ \mathbf{x}_3 = (1, -1, -1)^T,\ \mathbf{x}_4 = (-3, 5, -1)^T.$$

Write down the covariance of the maximum likelihood general Gaussian model fit to this data. To help you, some of the values are given below.

$$\begin{pmatrix} 8 & -4 & ? \\ ? & 6 & ? \\ 2 & ? & 6 \end{pmatrix}$$

[*4 marks*]

2. (a) You wish to model some data using a Gaussian process. Write down and describe the form of a Gaussian process model with covariance function (or kernel) $K(\mathbf{x}, \mathbf{x}')$ and zero mean, as defined over any finite dataset $\{(\mathbf{x}^n, y^n)|n = 1, 2, \ldots, N\}$, with $y^n \in \mathbb{R}$. You only need to write the prior distribution over this set of datapoints. You do not need to give the predictive equations. [*6 marks*]

(b) The Gaussian process can be used for classification, but is no longer analytic. One possible approach is to use a Laplace approximation. Describe in enough detail to recreate the method, what the Laplace approximation is, how to do it and why it has the form it does (There is no need to describe its specific application in the Gaussian process framework - a general description is enough). [*7 marks*]

(c) In many practical uses of Gaussian processes, the larger eigenvalues of the covariance matrix of functions correspond to the eigenvectors that are 'smoother'. By 'smoother' we mean that components corresponding to nearby positions are more likely to be similar.

Consider a zero mean Gaussian process evaluated at a fixed set of data points. The noisy measurements at those points are jointly denoted by $\mathbf{y}$ and the latent noise-free function values are denoted by $\mathbf{f}$. The covariance matrix for function values $\mathbf{f}$ is denoted $\mathbf{K}_f$, and the covariance matrix for the noisy observations $\mathbf{y}$ is denoted $\mathbf{K}_y = \mathbf{K}_f + \sigma^2\mathbf{I}$.

Given the values $\mathbf{y}$ at the data points, the predicted posterior mean value for function $\mathbf{f}$ at those same data points takes the form $\boldsymbol{\mu} = \mathbf{K}_f(\mathbf{K}_y)^{-1}\mathbf{y}$.

It is possible to decompose $\mathbf{K}_f$ using an eigen-decomposition $\mathbf{K}_f = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues, and $\mathbf{V}$ is the orthonormal matrix of eigenvectors (i.e. $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, hence $\mathbf{V}^{-1} = \mathbf{V}^T$).

i. Show that $\mathbf{K}_y$ and $\mathbf{K}_f$ share the same eigenvectors.

ii. By considering the noise-free predictions of $\mathbf{f}$ at the data points, show that, commonly, the Gaussian process predictions are increasingly smooth as the noise parameter σ^2 gets larger.

[*6 marks*]

(d) Describe a way to test the relevance of particular attributes to the prediction of the target value using the hyper-parameters of Gaussian processes with a squared exponential covariance function. [*6 marks*]

3. (a) Describe one benefit the use of neural networks provides over logistic regression. Give two disadvantages. [*3 marks*]

(b) Describe, at a high level, a good algorithm for optimising the parameters of a neural network for a given choice of network size and starting parameter. [*6 marks*]

(c) Precisely describe the two main properties that are desirable of a Markov chain for it to be used in an MCMC procedure. [*6 marks*]

(d) To sample from a neural network posterior, the Hamiltonian Monte-Carlo procedure is generally preferable to Metropolis Hastings.

Illustrate how the Hamiltonian Monte-Carlo sampler overcomes the random walk behaviour of the Metropolis Hastings algorithm. You should do this qualitatively by depicting in a drawing the way each process would sample from a square uniform distribution in 2D (you can assume this has been slightly smoothed by convolving with a Gaussian so that the derivatives are everywhere well defined). In other words you should illustrate the dynamics of the Markov chain. Assume the Metropolis-Hastings is using a Gaussian proposal with a width significantly less than the region of non-zero probability. Write a sentence or two to explain your drawings. [*7 marks*]

(e) You are sampling from a posterior distribution that you know is bimodal (i.e. has two peaks in the distribution). You also have the positions of the maximum posterior values for the two modes. You plan to augment a standard Metropolis Hastings sampler (with a local Gaussian proposal distribution) with an additional proposal every 20 steps. This additional proposal involves a jump between the regions of the two modes if the current sampler is within a certain distance of either mode.

Briefly describe how you might design a sampler to do this. You should state what form the proposal for the additional jump would take and what the acceptance probability should be. In giving the description, you may find it useful to provide a drawing illustrating this method in 2 dimensions.

[*3 marks*]