MOCK QUESTIONS (30/3/2009)

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

Monday 30 March 2009

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

MOCK QUESTIONS (30/3/2009)

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

   (a) Define what is meant by categorical data. Define what is meant by ordinal data. [*2 marks*]

   (b) Clearly write out how to compute the two major principal components of data $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$. What scaling issues are there with principal component analysis and what is a practical way of dealing with this is many circumstances? [*7 marks*]

   (c) Define the posterior distribution in terms of the prior distribution, the marginal likelihood and the likelihood. How can the marginal likelihood be used? [*4 marks*]

   (d) Write out the form of an exponential family distribution, stating what must be known about the different parts. Write down the conjugate prior corresponding to an exponential family likelihood of the form you wrote earlier. Show that the resulting posterior for the case of a single data point is analytically computable and takes the same form as the prior. Why is this enough to allow a straightforward proof for the case of any number of data points? [*7 marks*]

   (e) Define the KL divergence between two distributions. Give three important properties of the KL divergence. How does a variational approach use the KL divergence to approximate the posterior? Show how the variational method provides a lower bound to the likelihood. [*5 marks*]

2. **You should either answer this question or question 3.**

(a) Describe the form of a Gaussian process model for regression. Why is it harder to use Gaussian processes for classification? Give one approach for overcoming this difficulty. [*8 marks*]

(b) If a Gaussian process model for a function $\mathbf{f}$ corresponds to a finite feature space then after a sufficiently large number of data points, some of the eigenvalues of the covariance matrix will be zero. We investigate this here.

Consider using a linear model $f = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$ and a finite feature model with M features $\boldsymbol{\phi} = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$. Show that using a $N(\mathbf{0}, \mathbf{I})$ Gaussian prior on the parameter vector $\mathbf{w}$, the resulting Gaussian process has a covariance with at least $N - M$ zero values, where N is the number of data points. To do this you will want to collect together the values of f for all the data points into the vector $\mathbf{f}$, and the features into matrix $\boldsymbol{\Phi}$ (you should describe how this is done), resulting in equation

$$\mathbf{f} = \mathbf{w}^T\boldsymbol{\Phi}$$

You will want to compute the mean and covariance of $\mathbf{f}$ and use the fact that a matrix $\mathbf{C} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ where $\boldsymbol{\Phi}$ is N by M, and $M < N$ will have at least $N - M$ zero eigenvalues. [*5 marks*]

(c) What is Automatic relevance determination? Describe its application to Gaussian processes. [*5 marks*]

(d) How would you interpret a covariance function of the form:

$$K(x, y) = \exp\left(-\frac{2\sin^2\left(\frac{x-y}{2}\right)}{l^2}\right)$$

for one-dimensional real-valued x, y? [*3 marks*]

(e) Discuss the multi-modality of the marginal likelihood (of the length scales) for Gaussian process regression with a squared-exponential kernel. You may find it useful to draw diagrams to illustrate this. [*4 marks*]

3. **You should either answer this question or question 2.**

   (a) Define a Naive Bayes model, and derive estimates for the parameters of the model given suitable data. [*6 marks*]

   (b) Discuss a sensible process for optimising the parameters and number of units in a neural network. [*5 marks*]

   (c) Describe how Markov chains can be used to produce samples from an intractable posterior distribution. You will be expected to mention important properties of the Markov chains. [*3 marks*]

   (d) Describe three different MCMC sampling methods, and state the disadvantages of each. [*5 marks*]

   (e) Roughly draw the contours of the negative log likelihood of a two dimensional Gaussian distribution with covariance

$$\begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

   Draw the same diagram again below. On the top diagram, illustrate the sample paths of a Gibbs sampler (draw these as a connected path). On the bottom diagram draw the sample paths of a Hamiltonian Monte-Carlo sampler (you may presume that the accuracy and step size is such that all the proposals are accepted. You should plot points at each sample. [*4 marks*]

   Given samples x_i from $P(x) = 1 - x/2$, $0 \leq x \leq 2$, and the same number of unbiased coin throws $h_i \in \{0, 1\}$, how could you obtain the same number of samples from the distribution

$$P(x) = \begin{cases} 1/2 - x/4 & \text{for } 0 \leq x < 1 \\ 1/4 & \text{for } 1 \leq x < 2 \\ 1/2 - (3-x)/4 & \text{for } 2 \leq x \leq 3 \end{cases}$$

   [*2 marks*]