UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Monday 7 th May 2012

14:30 to 16:30

MSc Courses

Convener: B. Franke
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) You have a dataset $D = (\mathbf{x}^n, y^n; n = 1, 2, \ldots, N)$, and a conditional machine learning model of the form $P(y^n|\mathbf{x}^n, \mu_1, \mu_2, s_1, s_2)$, where μ_1, μ_2, s_1 and s_2 are all parameters of the model. The parameters μ_1 and μ_2 are *mean* parameters used within the model, and the s_1 and s_2 are *variance* parameters.

   i. Write down the log-likelihood maximisation (for the whole dataset D) in the form of a constrained minimisation problem for the parameters μ_1, μ_2, s_1 and s_2. A constrained minimisation is written in the form

$$\theta^* = \arg\min_\theta f(\theta) \text{ subject to } \begin{array}{c} g_1(\theta) > 0 \\ g_2(\theta) > 0 \\ \ldots \\ g_k(\theta) > 0 \end{array} \quad (1)$$

   for some k and $f, g_1, g_2, \ldots, g_k$. [*4 marks*]

   ii. Rewrite the log-likelihood maximisation as an equivalent unconstrained minimisation problem. [*3 marks*]

   iii. Very briefly explain why using line minimisation within a high dimensional optimisation procedure can be beneficial in terms of speed? [*2 marks*]

   (b) Suppose we have a posterior density $P(\theta|D)$ for parameters θ and for data denoted by D.

   i. Describe the Monte-Carlo approach for approximating the posterior mean

$$\int \theta P(\theta|D) d\theta, \quad (2)$$

   where the integral is a definite integral over the whole parameter space. [*3 marks*]

   ii. Describe the process for rejection sampling using a distribution $Q(\theta)$ that we are able to sample from, and where $Q(\theta) > \alpha P(D|\theta)P(\theta)$. [*4 marks*]

   iii. In importance sampling using a proposal distribution $Q(\theta)$ for a target distribution $P(\theta|D)$, what condition do we need on Q for the importance sampling procedure to be valid? [*1 mark*]

   (c) Write out the Bernoulli likelihood for a binary dataset $x^1, x^2, \ldots, x^N$ with Bernoulli probability p. From this, derive 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. [*5 marks*]

   (d) Naive Bayes assumes conditional independence. By considering the worst case of two attributes being identical given the class label, explain what effect a positive correlation between attributes (given the class) has on the inferred posterior probabilities. [*3 marks*]

2. (a) In standard degree courses, students can get one of a number of final marks (Fail, 3rd, Lower 2nd, Upper 2nd, 1st). You plan to use a neural network to predict the class of degree that a student will get dependent on the marks they obtained on coursework for their courses. Note that different people do different courses, each with different numbers of coursework. You could choose to represent the data by having one input attribute for each piece of coursework, and substituting the mean coursework value (computed across all those who did the course) in cases where individuals did not do that course.

   Alternatively, you could represent the data for an individual by having one input attribute for each range (0% to 10%,11% to 20%, ..., 99% to 100%) and making the input attribute value to be the proportion of the coursework the person did, that was given a mark in that range. For example the inputs for one individual might take the form of a vector $(0\%, 0\%, 0\%, 0\%, 10\%, 40\%, 40\%, 10\%, 0\%, 0\%)^T$.

   i. Give one brief argument against each of the alternative representations. [*6 marks*]

   ii. Suppose that you knew you were going to use a Naive Bayes model. Describe a representation that you could then use that is similar to, but arguably more elegant than, the first alternative above in that it explictly represents missing data, and briefly explain why that representation is suitable for Naive Bayes? [*2 marks*]

   (b) This question relates to neural networks in practice:

   i. Explain why it is important to standardise the data and start with small weights in neural networks. [*3 marks*]

   ii. Why should each of the initial weights for the different units be different from one another? [*2 marks*]

   iii. If we stop training early, what is the effective bias this induces on our learnt networks? [*1 mark*]

   (c) Describe a simple gradient ascent procedure for optimising the weights $\mathbf{w}$ (which includes the biases) of a neural network, given the network log-likelihood, denoted $L(D|\mathbf{w})$ for data D. Describe the problems associated with setting the learning rate. You do not need to say how to compute any derivatives you might need. [*5 marks*]

QUESTION CONTINUES ON NEXT PAGE

(d) Let $\sigma(x)$ be the logistic function, and so $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$. Consider the single datum likelihood for a logistic regression model:

$$P(y_i = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) \tag{3}$$

(where the bias is included in $\mathbf{w}$ by augmenting the data $\mathbf{x}$ with a unit attribute). Show the derivative of the negative log-likelihood with respect to w_i is

$$-(1 - \sigma(\mathbf{w}^T\mathbf{x}))x_i \tag{4}$$

and hence derive the form for an element of the Hessian matrix (the matrix of second derivatives) for the logistic regression model. Using the fact that a convex function has no *local* minima, and the fact that the matrix $\mathbf{x}\mathbf{x}^T$ has all eigenvalues greater than or equal to zero, show that parameter estimation for logistic regression has at most a single maximum. [*6 marks*]

3. (a) For real $\mathbf{y}$ and binary c, let $P(\mathbf{y}|c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ be a class conditional Gaussian distribution with mean $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$:

$$P(\mathbf{y}|c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \frac{1}{|2\pi\boldsymbol{\Sigma}_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{y} - \boldsymbol{\mu}_c)\right) \tag{5}$$

Consider using a class-conditional model for a problem with two classes $c \in \{0, 1\}$. You are given learnt $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, where the $\boldsymbol{\Sigma}_c$ are constrained to be identical covariance matrices for both classes $c = 0, 1$

i. How would you estimate $P(c)$ from a dataset of size N? What assumptions are you making? [*2 marks*]

ii. Write down how to obtain the probability of the class label $P(c|\mathbf{y})$ in terms of $P(\mathbf{y}|c)$ and $P(c)$. [*1 mark*]

iii. If $P(c = 1) = P(c = 0)$ in this situation, show that the decision boundary for the classification $P(c|\mathbf{y})$ is linear (assume we classify using the most probable class). [*6 marks*]

(b) It is possible to combine class-conditional modelling and Gaussian processes to make a class conditional Gaussian process model. In this model $P(\mathbf{y}|c, \mathbf{x}) = \prod_i P(y_i|c, \mathbf{x})$ ($\mathbf{y} \in \mathbb{R}$, $c \in \{0, 1\}$) where, for each c, $P(y_i|c, \mathbf{x})$ is a Gaussian process model. A prior $P(c|\mathbf{x}) = P(c)$ is also given. The aim is to predict $P(c|\mathbf{x}, \mathbf{y})$. Alternatively, $P(c|\mathbf{x}, \mathbf{y})$ can be modelled directly with a Gaussian process classifier.

i. Discuss the computational advantages of the class conditional Gaussian process model over the Gaussian process classifier (give at least two important advantages for full marks). [*4 marks*]

ii. What is one modelling disadvantage of the class conditional Gaussian process model over the Gaussian process classifier (give a problematic assumption of a class conditional Gaussian process described above). [*2 marks*]

(c) 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 that the $KL(Q||P) \geq 0$ 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. [*5 marks*]

(d) Define detailed balance for a Markov chain, and show that if detailed balance holds for a distribution $Q(X)$ and transition $P(X_t|X_{t-1}) = P(X_1|X_0)$, then that distribution must be a fixed point of the Markov Chain. In other words, if X_t is distributed as $Q(X_t)$ then taking one step of the Markov Chain $X_t \rightarrow X_{t+1}$, leaves X_{t+1} distributed as $Q(X_{t+1})$. [*5 marks*]