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 14th May 2010

09:30 to 11:30

Year 4 Courses

Convener: D. K. Arvind
External Examiners: K. Eder, A. Frisch

MSc Courses

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

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) Describe the difference between categorical data, ordinal data and discrete data. Give an example of a case where the existence of one attribute can be dependent on the value of another attribute. [*4 marks*]

(b) For data $\mathcal{D}$, define the posterior distribution (over parameters θ) in terms of the prior distribution, the marginal likelihood and the likelihood. How can the marginal likelihood be practically computed in the case of a small discrete set of parameter values? [*3 marks*]

(c) Consider using a logistic regression model to model the data in each figure displayed on the next page (where we use the horizontal and vertical axes as the features – we don't compute any additional nonlinear features). Parameters are learnt for each of the three cases using maximum likelihood, and using the illustrated points as training data. The + and ◯ denote the two classes.

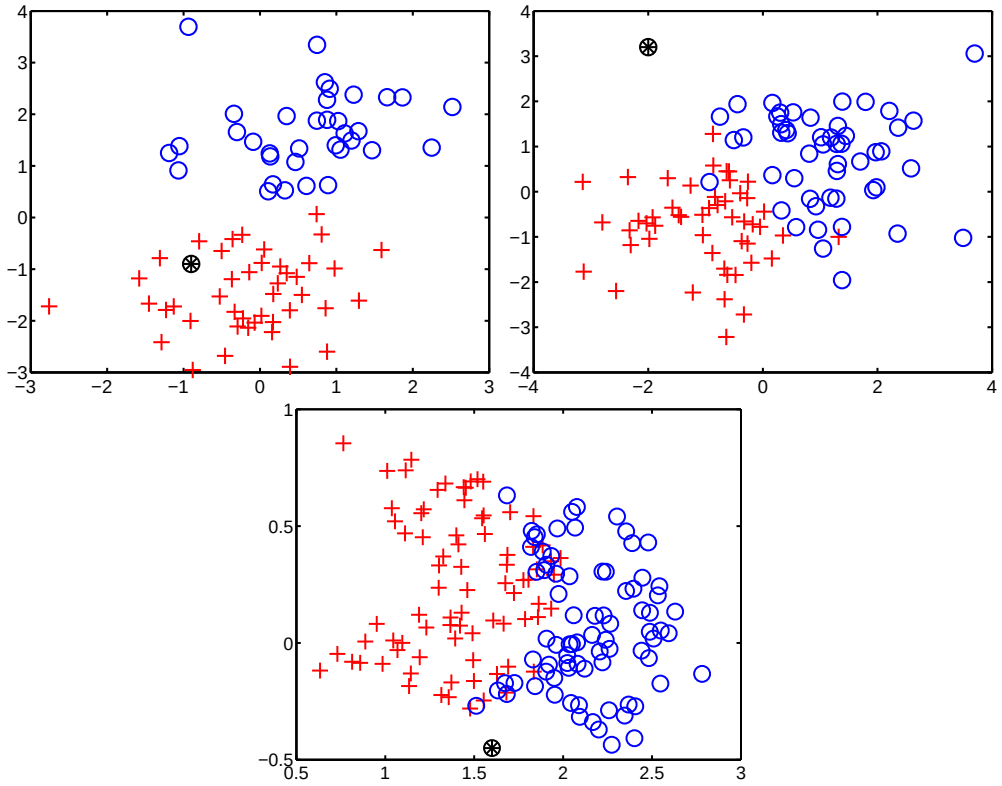
Copy each of the three figures and draw approximately the likely position of the decision boundary. You only need to copy the figures to sufficient accuracy to reasonably illustrate your boundary.

Estimate by eye the probability of the point labelled with an encircled star being in the + class for each case. For two of these cases there is a problem with using maximum likelihood. State which two, describe the problem and propose a solution. [*10 marks*]

(d) You have data consisting of records of the form (x_1, x_2, x_3) where x_3 is a class label we wish to predict, and x_1 and x_2 are the covariates (the variables we will be given). All data is binary.

Treating x_3 as a class label, define clearly and precisely the maximum likelihood naive Bayes classification method for obtaining a probability of class x_3 given the values of the other attributes. Write out the full log likelihood for the Naive Bayes model, and hence show that the maximum likelihood estimate of the parameter $p(x_i = 0|x_3 = 0)$ is proportional to the number of times attribute i is 0 for class 0 data. [*8 marks*]

QUESTION CONTINUES ON NEXT PAGE


4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
1
0.5
0
−0.5
0.5
1
1.5
2
2.5
3

2. (a) Describe how to build a sample from a distribution P using importance sampling, and how to use it to compute a Monte-Carlo estimate for the mean of the distribution P. [*7 marks*]

   (b) 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*]

   (c) Suppose you use a Gaussian process with no noise, a zero mean function and a squared exponential covariance of the precise form

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^2\right)$$

   Approximately what would your posterior distribution be for a predictive point that is more than 10 units away from any data point. [*4 marks*]

   (d) Define the KL divergence, and state how the KL divergence is used in the variational approximation to find a good approximation to the posterior distribution. [*3 marks*]

   (e) Can a Gaussian be fit to a uniform distribution using a variational approximation? Explain. [*1 mark*]

   (f) The figure overleaf illustrates a true posterior distribution. Copy the figure and illustrate on the same axes the form of approximate posterior distribution that you would most likely get for each of the following approximations to the true posterior. Label any important features of each (also, be sure to label which is which).

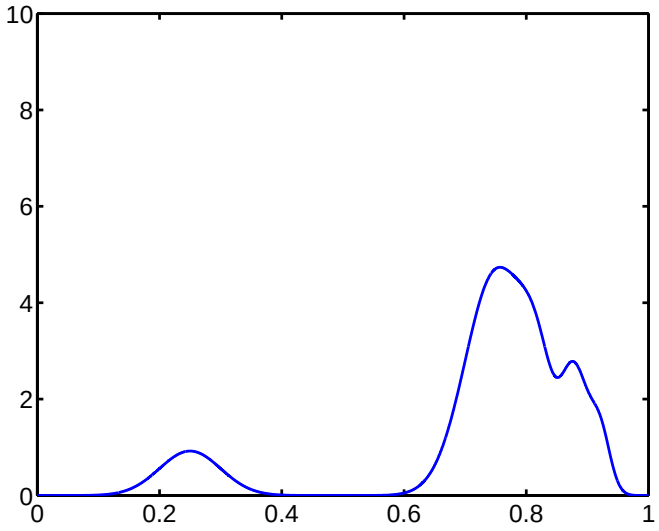
      i. A variational approximation with a Gaussian distribution.

      ii. A Laplace approximation.

   You only need to copy the figures to sufficient accuracy to reasonably illustrate your answers. [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

QUESTION CONTINUED FROM PREVIOUS PAGE



3. (a) Write out how to compute the two major principal components of data $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ in a way that someone could implement it from your description. [*5 marks*]

   (b) Consider a one dimensional real valued dataset which you believe contains two classes. You have learnt a class conditional Gaussian model for the data. Let x denote a point in the space of the data, and let A and B denote the two classes. Class A is modelled with a class conditional Gaussian with mean $\mu_A = 2$ and variance $\sigma_A^2 = 1$. Class B is modelled with a class conditional Gaussian with mean $\mu_B = 1$ and variance $\sigma_B^2 = 1/4$. The priors are presumed equal.

   Sketch the probability density for the two Gaussian distributions corresponding to class A and class B. Do so on the same graph. Be sure to label the critical points, including the central point of each curve and the height of each curve. Make sure the places where the distributions become negligibly small are broadly in the right places, and the shapes of the curve are broadly correct. You may use the fact that $1/\sqrt{2\pi} \approx 0.4$. Mark whatever decision boundaries there are. [*6 marks*]

   (c) You have two models for univariate data of the form $0 < x \leq 1$:

   $$H_1 : P(x) = -\log(x)$$
   $$H_2 : P(x|a) = 1/a \text{ iff } 0 < x \leq a \text{ and } P(x) = 0 \text{ otherwise}$$

   where all logarithms are natural logarithms.

   You choose a uniform prior over the parameter a: $0 < a \leq 1$. You give each model equal prior weight.

   i. Can you choose between these two models after seeing one data point? Explain your answer. [*3 marks*]

   ii. Suppose the first two data points just happened to both be 0.2 (or in reality very close to 0.2). Work through how you would choose between the two hypotheses using Bayesian model selection. ($[\log(0.2)]^2 \approx 2.6$) [*4 marks*]

   (d) Briefly describe the Hamiltonian Monte-Carlo approach to sampling (e.g. for sampling the hyperparameters of a Gaussian process). Explain the benefits it has over Metropolis Hastings, and how it achieves those benefits. [*7 marks*]