The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2012–2013

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE 2 hours

Answer THREE of the four questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

Registration number from U-Card (9 digits) — to be completed by student

1. This question concerns general concepts in machine learning.

a) *Bayes' rule* is used in many contexts in machine learning.

(i) Provide a definition of Bayes' rule. You should include mathematical formulae, and define any variables used. [10%]

(ii) Define the following components in Bayes' rule – *posterior*, *likelihood*, *prior*, *marginal likelihood* – and briefly describe their purpose (1-2 sentences for each). [20%]

(iii) Provide two examples where Bayes' rule is used in machine learning, and describe why Bayes' rule is used in this setting. [15%]

b) What does the term *marginalise* mean in relation to probability distributions? You should consider both the *discrete* and *continuous* settings, and provide mathematical formulae to support your answer. [15%]

c) The Binomial-Beta is said to be an example of a *conjugate* prior relationship.

(i) Give a definition of a *conjugate prior*, and motivate why conjugate priors are desirable. [15%]

(ii) Prove that conjugacy holds for the Binomial and Beta distributions. Show your working. [25%]

For your reference, the Binomial distribution is defined as

$$P(k, n|u) = \binom{n}{k} u^k (1-u)^{n-k}$$

where k is the number of successes after n trials (both positive integers), and u is the binomial parameter (real number between 0 and 1). The Beta distribution is defined as

$$P(u|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} u^{\alpha-1}(1-u)^{\beta-1}$$

where u is a real number between 0 and 1 and α and β are the Beta parameters. The function $B(\cdot)$ is a normalising constant.

2. This question is based on the following data. We are trying to predict whether or not it will rain, and have identified two features which might be important—whether the sky is clear or cloudy, and whether or not we hear birds singing. Over three days, we observed the following:

   1. cloudy sky; birds singing; rain
   2. clear sky; birds singing; no rain
   3. clear sky; birds quiet; rain

a) We have decided to model the data using the perceptron algorithm.

(i) Illustrate this data using a graph, denoting each day as a point. Draw a separating hyperplane on the graph and label the regions for the two classes. [10%]

(ii) State the weight update rule used in the perceptron algorithm. [5%]

(iii) Now apply the perceptron algorithm for training the model parameters. First, represent your training data as a matrix, $\mathbf{X}$, for the data points and vector $\mathbf{t}$ for their target values (+1 = rain and -1 = no rain). Now perform just one pass of the perceptron algorithm over the training set, showing your working and the final parameter values. Don't forget to include a bias term. [15%]

(iv) The following day there is a cloudy sky and the birds are quiet. What is your model's prediction (i.e., rain or not)? Include your working. [5%]

b) It turns out on the fourth day that it doesn't rain. We now elect to include this new example (cloudy sky; birds quiet; no rain) into our training set and re-train our model.

(i) The above model will no longer be appropriate for this dataset. Justify why this is the case. [10%]

(ii) Would you be able to solve the problem using radial basis functions? If so, how many RBFs are needed and where could they be placed? Please justify your answer, either way. [20%]

c) *Support Vector Machines (SVMs)* refine the perceptron by including the notion of margin of separation.

(i) Illustrate the concept of the separating margin using a diagram assuming binary classification with two dimensional input data. Now highlight all the support vectors, and annotate the margin. State which points violate the margin constraints (ensure you include a few). [15%]

(ii) What is the loss function being minimised by the soft-margin SVMs? Include the mathematical formula. How does this differ from the zero-one loss and logistic loss? Provide a diagram to illustrate your answer. [20%]

3. This question concerns regression and the Bayesian approach to regression with basis functions.

a) For each pair of terms below define and contrast the two terms. Use at least one example to illustrate your answer.

(i) overdetermined and underdetermined systems [15%]

(ii) epistemic and aleatoric uncertainty [15%]

b) A typical linear model could have the form

$$t_i = mx_i + c + \epsilon_i$$

where t_i is the regression target observation, x_i is the input location and ϵ_i is the noise. All are associated with the ith observation.

(i) Write the form of the basis set for the ith data point, $\boldsymbol{\phi}_i$, such that this model can be written:

$$t_i = \mathbf{w}^\top \boldsymbol{\phi}_i + \epsilon_i.$$

[5%]

(ii) The linear model is a 1st order polynomial. What would the basis, $\boldsymbol{\phi}_i$, be for a 4th order polynomial? [5%]

c) In a regression problem we are given a vector of real valued targets, $\mathbf{t}$, consisting of N observations $t_1 \ldots t_N$ which are associated with a unidimensional input $x_1 \ldots x_N$. We are to perform the regression by minimizing the following error function with respect to $\mathbf{w}$

$$E(\mathbf{w}) = \sum_{i=1}^{N} (t_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2.$$

Write down the *likelihood* that corresponds to this error function, introducing any additional parameters as necessary. Describe how the error function is related to the likelihood. [20%]

d) Consider the following Gaussian prior density for $\mathbf{w}$,

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

where k is the length of the vector $\mathbf{w}$ and α is the variance of the prior.

(i) Multiply the prior by the likelihood from (c). Show that the result is of the form of an exponentiated quadratic, and describe why that means the posterior density for $\mathbf{w}$ is Gaussian. [20%]

(ii) Show that the mean, $\boldsymbol{\mu}_w$, and covariance, $\mathbf{C}_w$, of the posterior density are given by

$$\mathbf{C}_w = \left[\alpha^{-1}\mathbf{I} + \sigma^{-2}\sum_{i}^{N} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top\right]^{-1}$$

$$\boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2} \sum_{i=1}^{N} t_i \boldsymbol{\phi}_i$$

[20%]

4. This question concerns the concepts behind data modelling such as generalization and model selection. In your answers, when it is appropriate, you may want to make use of the regression example we saw in the lectures and lab class involving the gold medal winning 100m times from the Olympic games between 1896 and 2008.

   a) Give short definitions for the following terms associated with a model fitting and generalisation capability.

      (i) overfitting [5%]

      (ii) extrapolation [5%]

      (iii) interpolation [5%]

   b) In this part we will cover approaches to model selection.

      (i) What is a validation set? [10%]

      (ii) What is the difference between hold out validation and cross validation? [20%]

      (iii) What are the relative advantages and disadvantages of leave-one-out cross validation and five fold cross validation? [25%]

   c) What is the Bayesian approach to model selection and why is it less susceptible to overfitting? [30%]

END OF QUESTION PAPER