SEMESTER 2 EXAMINATION 2005/2006

MACHINE LEARNING

Duration: 120 mins

Answer ALL *questions from section A (20 marks)*
and ONE *question from section B (25 marks)*
and ONE *question from section C (25 marks).*

This examination is worth 70%. The coursework was worth 30%.

Calculators without text storage MAY be used.

Section A

Question 1

(a) Explain what is meant by generalisation error and describe how it is estimated. *(2 marks)*

(b) Give a Bayesian interpretation for minimising the sum of the mean squared error plus a regularisation term. *(3 marks)*

(c) Show that a MLP using linear nodes is no more powerful than a linear perceptron. *(5 marks)*

(d) Describe what is meant by the terms *training set*, *validation set* and *testing set*. *(3 marks)*

(e) Describe what is meant by the terms *classification*, *regression* and *density estimation*. *(3 marks)*

(f) Describe the *kernel trick* and how it is applied in machine learning. *(4 marks)*

Section B

Question 2

The linear perceptron with no bias has a response

$$y = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}.$$

We assume we have a set of training data

$$\{(\boldsymbol{x}^k, t^k) | k = 1, \ldots, P\}$$

where $\boldsymbol{x}^k$ are input patterns and t^k are targets.

(a) Write down an expression the mean square training error, $E(\boldsymbol{w})$, for the linear perceptron. *(3 marks)*

(b) By writing the inputs as a single matrix **X** whose k^{th} column is the input $\boldsymbol{x}^k$ and the targets as a vector $\boldsymbol{t}$ whose k^{th} element is t^k, express the mean squared training error in matrix form. *(3 marks)*

(c) By computing the gradient of the training error find the value of the weight vector which minimises the training error. *(4 marks)*

(d) Explain what it means for $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ to be ill-conditioned and argue why the weights found will be sensitive to the training data if this is the case. *(5 marks)*

(e) Show that by using a modified error function

$$\hat{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \nu\, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

the weight vector of the linear perceptron will be less sensitive to the training data. *(6 marks)*

(f) Explain in terms of the bias-variance dilemma why introducing a weight decay term can improve the generalisation performance. *(4 marks)*

Question 3

Assume we have a learning machine of the form

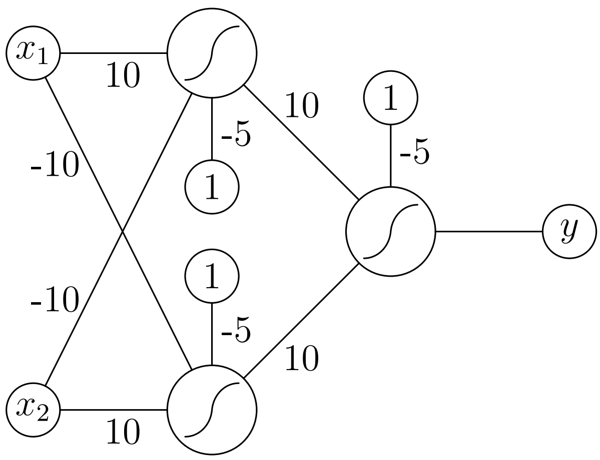
$$F(\boldsymbol{x}; b, \boldsymbol{w}, \mathbf{Q}) = b + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathbf{Q}\boldsymbol{x}$$

where the parameters to be learned, b, $\boldsymbol{w}$ and $\mathbf{Q}$, are a scalar, a vector and a symmetric matrix respectively.

(a) Given training data $\mathcal{D} = \{(\boldsymbol{x}^k, t^k) | k = 1, \ldots, P\}$ write down the mean squared error and compute the gradient with respect to b, w_i and $Q_{i,j}$. *(8 marks)*

(b) For two dimensional input patterns, sketch the contour lines where $F(\boldsymbol{x}; b, \boldsymbol{w}, \mathbf{Q})$ is constant. Compare this with similar surfaces for a linear perceptron and an RBF networks. *(8 marks)*

(c) Consider the MLP below



(i) Write an equation describing the response, y, in terms of the input $\boldsymbol{x} = (x_1, x_2)$ assuming a response function $g(x)$.

(ii) Show that the output of the network is constant along the line defined by $x_1 - x_2 = u$ and $x_1 - x_2 = -u$.

(iii) Sketch a contour diagram in the input space showing where the output of the network has equal response.

(9 marks)

Section C

Question 4

(a) What is an ill-posed problem? *(7 marks)*

(b) Describe the method of regularisation, making reference to examples in machine learning.

(9 marks)

(c) Show that the solution to the regularisation problem is equivalent to a Maximum A Posteriori (MAP) estimate (assume Gaussian noise). *(9 marks)*

Question 5

(a) Explain what is meant by the term over-parameterisation. State how it is removed from the hyperplane, $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$, in the linear Support Vector Machine formulation to produce a canonical hyperplane. *(3 marks)*

(b) State the condition for separability of the two-class data-set

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n},\ \boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \{-1, 1\}$$

with this canonical hyperplane. *(3 marks)*

(c) State the maximum margin principle and derive an expression for the Lagrangian of the resulting optimisation problem. *(10 marks)*

(d) Solve the Lagrangian problem,

$$\max_{\boldsymbol{\alpha}} \left(\min_{\boldsymbol{w}, b} \Phi(\boldsymbol{w}, b, \boldsymbol{\alpha}) \right),$$

to show that the solution for the Lagrange multipliers can be written as a quadratic program. *(9 marks)*

END OF PAPER