UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2008/2009

MACHINE LEARNING

Duration: 120 mins

Answer THREE *questions out of* FOUR

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

University approved calculators MAY be used.

Question 1

(a) Explain what you understand by *supervised learning*, *unsupervised learning* and *novelty detection*. Briefly describe one potential application of each of these.

(6 marks)

(b) A linear regression model given by

$$f = \boldsymbol{w}^t \boldsymbol{x}$$

is to be estimated from N items of data, $\{\boldsymbol{x}_n, f_n\}_{n=1}^N$, where the usual offset term w_0 is ignored. Show how minimising the average squared error leads to a closed form solution for the parameter vector $\boldsymbol{w}$. Derive your answer in the form of a pseudo inverse of a matrix.

(7 marks)

(c) How would you modify your solution above to obtain an *online* algorithm for estimating $\boldsymbol{w}$. *(4 marks)*

(d) Comment on the convergence properties of the online algorithm.

(3 marks)

(e) Show how, by choosing a suitable substitute for the squared error in the regression problem, the *perceptron* algorithm for classification may be derived. *(7 marks)*

(f) Comment on the convergence properties of the perceptron algorithm.

(3 marks)

(g) What is its main limitation in solving pattern classification problems?

(3 marks)

Question 2

(a) Bayes rule for conditional probabilities, as commonly used in statistical pattern classification problems, is

$$P\left[A|\boldsymbol{x}\right] = \frac{P\left[A\right]\, p\left(\boldsymbol{x}|A\right)}{p\left(\boldsymbol{x}\right)}$$

With reference to a practical problem of your choice, explain the different terms in the above expression. *(10 marks)*

(b) Explain, using two dimensional sketches, the difference between *Principal Component Analysis* and *Fisher Linear Discriminant Analysis*. Briefly describe a potential application of each of these. *(10 marks)*

(c) A two dimensional two-class pattern classification problem is defined by the following data:

- Class A:
$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2.2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1.8 \end{bmatrix}, \begin{bmatrix} 0.8 \\ 2 \end{bmatrix}, \& \begin{bmatrix} 1.2 \\ 2 \end{bmatrix}$$
- Class B:
$$\begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2.7 \\ 1 \end{bmatrix}, \begin{bmatrix} 3.3 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 0.7 \end{bmatrix}, \& \begin{bmatrix} 3 \\ 1.3 \end{bmatrix}$$

Assuming the data are distributed according to Gaussian probability densities, derive an expression for the Bayes optimum class boundary.

Draw a neat sketch of the data and the class boundary.

Illustrate on your sketch the class boundary of a *nearest-neighbour* decision rule. *(13 marks)*

Question 3

(a) What is the difference between learning error and generalisation error? *(2 marks)*

(b) Explain how you can accurately estimate the generalisation error given a limited data set. *(6 marks)*

(c) Why are regularisation terms added to the error function? *(6 marks)*

(d) A weight decay term has the form $\lambda \sum_i w_i^2$. Show how adding such a term modifies the update rule for the weights and hence explain why it is known as a weight decay term. *(10 marks)*

(e) Explain the drawback of using a weight decay term and explain how an SVM avoids the need for such a term. *(9 marks)*

Question 4

(a) Explain how an SVM can be made to separate linearly separable data. Provide a schematic sketch of how this is done.

(5 marks)

(b) Mercer's theorem states that

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \lambda_i \, \psi_i(\boldsymbol{x}) \, \psi_i(\boldsymbol{y}).$$

Show that if the eigenvalues λ_i are non-negative (i.e. $\lambda_i \geq 0$) then for any real function $f(x)$

$$\int f(\boldsymbol{x}) \, K(\boldsymbol{x}, \boldsymbol{y}) \, f(\boldsymbol{y}) \, \mathrm{d}\,\boldsymbol{x} \, \mathrm{d}\,\boldsymbol{y} \geq 0.$$

(5 marks)

(c) Explain why positive semi-definiteness is an important property of kernels used in SVMs. *(5 marks)*

(d) Show that if $K_1(\boldsymbol{x}, \boldsymbol{y})$ and $K_2(\boldsymbol{x}, \boldsymbol{y})$ are positive semi-definite kernels then so is $K_3(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$. *(5 marks)*

(e) Using the fact that any positive semi-definite kernel can be decomposed as

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \phi_i(\boldsymbol{x}) \, \phi_i(\boldsymbol{y})$$

show that the product of two kernel functions is positive semi-definite.

(5 marks)

(f) Using the previous results show that the exponential of a positive semi-definite kernel function is also positive semi-definite.

(4 marks)

(g) Prove that the Gaussian kernel is positive semi-definite.

(4 marks)

END OF PAPER