UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2010/2011

MACHINE LEARNING

Duration: 120 mins

Answer all parts of the question in section A (20 marks)
and TWO *questions from section B (25 marks each)*

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

University approved calculators MAY be used.

Section A

Question 1

(a) Briefly describe what principal component analysis (PCA) does and how it is performed. *(6 marks)*

(b) Briefly describe K-means clustering. *(6 marks)*

(c) Describe (without mathematics) the Bias-Variance Dilemma. *(6 marks)*

(d) Explain how dimensionality reduction (e.g. using PCA or K-means) can reduce the generalisation error. *(2 marks)*

Section B

Question 2

(a) How would you represent the following categories in a numerical feature vector (input pattern)? Explain your decisions.

(i) Over 18 or not

(ii) Colour preference (red, green, yellow, blue)

(iii) Experience (none, little, medium, very)

(6 marks)

(b) Describe three different methods for handling missing data (i.e feature vectors with missing features). Discuss their relative benefits and problems.

(6 marks)

(c) Describe K-fold cross validation and explain why it is used?

(6 marks)

(d) Describe how to compute an ROC curve and how it can be interpreted?

(7 marks)

Question 3

(a) Assume you have a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ write down the squared training error for a **linear perceptron** with weights $\boldsymbol{w}$.

(3 marks)

(b) By writing the training patterns as a matrix $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ and the targets in a vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\mathsf{T}$ write down an expression for the squared training error in matrix form.

(3 marks)

(c) Compute the weight vector $\boldsymbol{w}^*$ that minimises the sum of the squared training error plus a regularisation term $\nu\|\boldsymbol{w}\|^2$.

(8 marks)

(d) Explain why without regularisation the $\boldsymbol{w}^*$ is ill defined if there are fewer training patterns than features (i.e. the size of the input vectors) and how adding a regularisation term cures this.

(5 marks)

(e) Explain why adding a regularisation term would make a linear perceptron less sensitive to the training data. Why might this improve the expected generalisation performance? *(6 marks)*

Question 4

(a) Given training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n, \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ what condition is required that the two classes can be separated by a perceptron. *(2 marks)*

(b) Assuming the data is separable there are usually many hyperplanes, $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b = 0$, that will separate the data. Explain what criteria is used in the linear Support Vector Machine to choose a unique hyperplane. Explain why this is a good choice?

(4 marks)

(c) Show (e.g. by drawing a diagram) that the distance between the separating plane defined by $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b = 0$ with $|\boldsymbol{w}| = 1$ and a data point $\boldsymbol{x}_i$ is equal to $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - b$ for points on the positive side (with respect to $\boldsymbol{w}$) of the separating plane and $-\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b$ for points on the other side. By rescaling $\boldsymbol{w}$ and b by the margin size show that the maximum margin hyper-plane can be found from the Lagrangian,

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left[\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - b\right] - 1\right), \quad \alpha_i \geq 0.$$

(8 marks)

(d) Solve the Lagrangian problem, $\max_{\boldsymbol{\alpha}} \left(\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha})\right)$, to show that the solution for the Lagrange multipliers can be written as a quadratic program,

$$\max_{\boldsymbol{\alpha}} \tfrac{1}{2}\boldsymbol{\alpha}^\mathsf{T}\mathbf{H}\boldsymbol{\alpha} + \boldsymbol{c}^\mathsf{T}\boldsymbol{\alpha},$$
$$\text{subject to the constraints,}$$
$$\alpha_i \geq 0, \quad \textstyle\sum_{j=1}^n \alpha_j y_j = 0.$$

(8 marks)

(e) What are the Support Vectors and how do these relate to the Lagrange multipliers?

(3 marks)

END OF PAPER