UNIVERSITY OF SOUTHAMPTON COMP6208W1

---

SEMESTER 2 EXAMINATION 2015 - 2016

ADVANCED MACHINE LEARNING

DURATION 120 MINS (2 Hours)

---

This paper contains 4 questions

Answer THREE out of FOUR questions (each question is worth 33 marks)

An outline marking scheme is shown in brackets to the right of each question.

This examination is worth 60%. The coursework was worth 40%.

University approved calculators MAY be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct Word to Word translation dictionary AND it contains no notes, additions or annotations.

6 page examination paper.

**Question 1.**

Assuming a linear model $f(\boldsymbol{x}|\boldsymbol{w}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$ and given a set of data $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)|k = 1, 2, \ldots, P\}$, we can train the weights by performing a linear least square with a weight decay regularisation by minimising the "error"

$$E(\boldsymbol{w}) = \sum_{k=1}^{P} \left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - y_k\right)^2 + \nu\|\boldsymbol{w}\|^2.$$

(a) Explain what the regularisation term does. [5 marks]

(b) By defining a matrix $\mathbf{X}$ with columns equal $\boldsymbol{x}_k$ and a vector $\boldsymbol{y}$ with components $y_k$, rewrite the error function, $E(\boldsymbol{w})$ as a matrix equation. Rearrange this to bring together the terms in powers of the weight vector. [5 marks]

(c) Find the set of weights, $\boldsymbol{w}^*$, that minimise the error function, $E(\boldsymbol{w})$. [3 marks]

(d) By using the singular value decomposition, $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$, rewrite the equations for the optimum weights in terms of the matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$. [5 marks]

(e) Thus show that $\boldsymbol{w} = \mathbf{U}\mathbf{M}\mathbf{V}^{\mathsf{T}}\boldsymbol{y}$ where $\mathbf{M}$ is a diagonal matrix. What are the components of $\mathbf{M}$? [5 marks]

(f) Explain why using $\nu > 0$ leads to a better conditioned problem than if $\nu = 0$. [5 marks]

(g) Explain the Bias-Variance dilemma, and explain how introducing a regularisation term improves the generalisation performance. [5 marks]

 

**Question 2.**

(a) Explain in words how SVMs control over-fitting. [5 marks]

(b) Explain the kernel trick. [5 marks]

(c) Define what it means for a kernel to be positive semi-definite (giving different properties a positive semi-definite kernel satisfies). Explain why it is necessary for the kernel function of an SVM to be positive semi-definite. [5 marks]

(d) Briefly explain how to train a deep belief network. [8 marks]

(e) Explain why back-propagation is not an effective way to train a deep multi-layer perceptron. [5 marks]

(f) Explain the advantages and disadvantages of using an SVM as opposed to a deep belief network. [5 marks]

**TURN OVER**

**Question 3.**

(a) You are given a sequence $\mathcal{X} := \{x^{(1)}, \ldots, x^{(12)}\}$ of heads ($x^{(i)} = H$) and tails ($x^{(i)} = T$) which are the outcomes of 12 tosses of a (potentially biased) coin. Describe how you would fit the data to a binomial distribution $B(N, \theta)$ for $N = 12$ and $\theta = Pr(H)$, using maximum likelihood estimation.

[7 marks]

(b) Show that maximising the log-likelihood is equivalent to minimising the Kullback-Leibler divergence between $B(N, \theta)$ and the empirical distribution $\tilde{p}(\mathbf{x})$.

[7 marks]

(c) Discuss how a conjugate Beta prior

$$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

introduces "psuedo-counts" to affect the estimation of parameters of the binomial distribution. *Hint:* $\Gamma(n) = (n-1)!$ for integer arguments $n$.

[7 marks]

(d) You are given a data set containing observed values $v^{(n)}$ of variable $V$. You construct a probability model $p(V = v^{(n)}, H = h^{(n)}|\theta)$ that introduces model hidden variables $H$ that take values $h^{(n)} \in H$ that are associated with observations $v^{(n)}$. To infer the parameters $\theta$ you will need to maximise the log likelihood of the observed data

$$\sum_{n=1}^{N} \log p(V = v^{(n)}|\theta).$$

Describe how you would use the EM algorithm to perform this task. Describe how you can impute missing data using the EM algorithm.

[12 marks]

 

**Question 4.**

(a) In a graphical model a node is shaded if it represents an observed variable. By writing down the joint probabilities of the following graphical models determine when $A$ and $B$ are (conditionally or otherwise) independent.

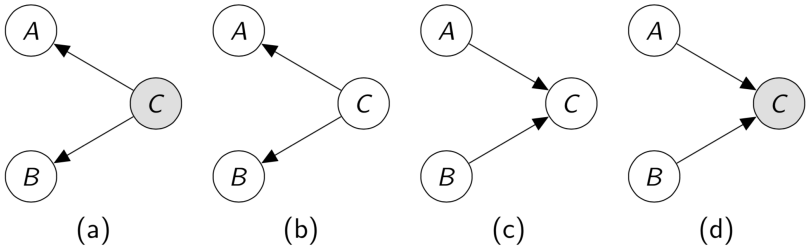


FIGURE 1: A shaded node indicates that a value taken by the corresponding variable is observed.

[8 marks]

(b) What does it mean for a classifier to have high variance? Explain what the concepts of bootstrapping and bagging are, and explain how the variance of a predictive classifier may be reduced if they are used appropriately. [10 marks]

(c) Describe the AdaBoost algorithm. The classification rule $F : X \to Y$ in AdaBoost takes $x \in X$ an input vector produces $Y = \{1, -1\}$ a binary output, where $F(x) = \text{sign}(\sum_m c_m f_m(x))$. What are $f_m(x)$ and $c_m$ and how are they used? Prove that the expectation, taken with respect to the conditional distribution $p(Y = y|X = x)$ of the loss function $L(y, f_m(x)) = \exp(-y f_m(x))$:

$$\mathbb{E}_{Y|X=x}(e^{-y f_m(x)})$$

is minimised at

$$f_m(x) = \frac{1}{2} \log \left( \frac{p(y = 1|x)}{p(y = -1|x)} \right).$$

[7 marks]

**TURN OVER**

(d) Describe the Metropolis-Hastings sampling method and show how the detailed balance condition explains the emergence of a stable probability distribution for the sample.

[8 marks]

**END OF PAPER**