UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2014/2015

ADVANCED MACHINE LEARNING

Duration: 120 mins

Answer THREE *out of* FIVE *questions (each question is worth 20 marks)*

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

University approved calculators MAY be used.

A foreign language translation dictionary (paper version) is permitted provided it contains no notes, additions or annotations.

Number of pages 8

Question 1 Given a set of data $\mathcal{D} = \{(\boldsymbol{x}_k, y_k) | k = 1, 2, \ldots, P\}$, we can perform ridge regression in an extended feature space by minimising the cost function

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

(a) What is the interpretation of these two terms?

(2 marks)

(b) The minimum cost weight vector will lie in a space spanned by the data-points in the extended feature space so that

$$\boldsymbol{w} = \sum_{\ell=1}^{P} \alpha_\ell \boldsymbol{\phi}(\boldsymbol{x}_\ell).$$

Explain why this is the case? *(3 marks)*

(c) Defining the kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y})$ rewrite the cost function in terms of the kernel and the parameters α_k.

(3 marks)

(d) Defining the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{y}$ with components α_k and y_k respectively, and the matrix **K** with components $K(\boldsymbol{x}_\ell, \boldsymbol{x}_k)$, rewrite the cost function in matrix form. *(2 marks)*

(e) Find the parameters $\boldsymbol{\alpha}^*$ that minimise the cost function.

(5 marks)

(f) Explain why it is important that the kernel function is positive semi-definite.

(2 marks)

(g) Give three conditions that a positive semi-definite kernel must satisfy.

(3 marks)

Question 2

(a) Describe a deep belief network and explained how it is trained. *(5 marks)*

(b) Explain what deep learning attempts to do. *(3 marks)*

(c) What are the main drawbacks of using deep learning? *(2 marks)*

(d) Explain how dropout is implemented in deep belief networks. *(5 marks)*

(e) Explain what dropout does and why it is used. *(5 marks)*

Question 3

This question will need you to use the KL divergence $KL(p\|q)$ between two distributions p and q:

$$KL(p\|q) = \mathbb{E}_p\left(\log(p/q)\right),$$

where $\mathbb{E}_p f = \sum_x p(x)f(x)$ denotes the expectation of f taken with respect to the distribution p.

(a) For the dataset $\mathcal{X} := \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, we can compute the *empirical distribution*

$$\tilde{p}(\mathbf{x}) = \frac{1}{N}(\sum_{n=1}^{N} \mathbb{I}[\mathbf{x} = \mathbf{x}^{(n)}]),$$

where

$$\mathbb{I}[x = y] = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

Find the probability distribution $p(x)$ that minimises the KL divergence $KL(\tilde{p}\|p)$. What is the log likelihood of $p(x)$?

(4 marks)

(b) 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 θ you will need to maximise the log likelihood, $\ell(\theta; \mathcal{V})$, of the observed data

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

Using the KL divergence between a variational distribution $q(h^{(n)}|v^{(n)})$ and the probability model for the hidden variables $p(H = h^{(n)}|V = v^{(n)}, \theta)$, show that the log likelihood $\ell(\theta; \mathcal{V})$ has a lower bound

$$\ell(\theta; \mathcal{V}) \geq \ell_{LB}(\{q\}, \theta) = \sum_{n=1}^{N} S_n + E_n,$$

where

$$S_n = -\sum_h q(h^{(n)}|v^{(n)}) \log\left(q(h^{(n)}|v^{(n)})\right)$$

and E_n is the expected complete (hidden and observed) data log likelihood:

$$E_n = \sum_h q(h^{(n)}|v^{(n)}) \log\left(p(h^{(n)}, v^{(n)}|\theta)\right).$$

For what choice of the variational distribution $q(h^{(n)}|v^{(n)})$ is the lower bound reached? *(6 marks)*

(c) A data set $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ is shown in the table below, where the variable $\mathbf{X} = (A, B)$ with A and B binary:

$\mathbf{X}$	A	B
$\mathbf{x}^{(1)}$	0	0
$\mathbf{x}^{(2)}$	1	0
$\mathbf{x}^{(3)}$	0	0
$\mathbf{x}^{(4)}$	1	?
$\mathbf{x}^{(5)}$	?	0
$\mathbf{x}^{(6)}$	1	1

This questions uses the EM algorithm to find the parameters $\theta_{ij} := p(A = i, B = j)$, (for example, $\theta_{00} := p(A = 0, B = 0)$) from $\mathcal{X}$ which contains missing values denoted by '?'.

(i) Write down the conditional probabilities for the missing values $p(A = a|B = b)$ and $p(B = b|A = a)$ for a given (perhaps random) choice of parameters $\theta_{ij}^{(old)}$.

(ii) Write down the complete log likelihood of the data $\mathcal{X}$ using the estimated conditional probabilities for the missing values $q_t(A = a|B = b)$ and $q_t(B = b|A = a)$.

(iii) For what set of parameters is the log likelihood of the hidden and visible data maximised?

(10 marks)

Question 4

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

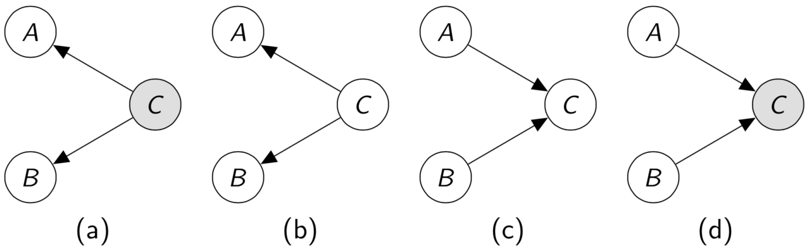

FIGURE 1: A shaded node indicates that a value taken by the corresponding variable is observed.

(4 marks)

(b) A dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$ consists of a 1-dimensional output $y^{(n)}$ for each D-dimensional input data point $\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_D^{(n)})$ for $n = 1, \ldots, N$, and the collections of inputs and outputs are denoted $\mathcal{X}$ and $\mathcal{Y}$ respectively. The data is fit by a linear function $f(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} = \beta_0 + \sum_j \beta_j x_j$. The residual errors $\epsilon^{(n)} := (y^{(n)} - f(\mathbf{x}^{(n)}))$ are identically and independently distributed (iid) according to a Gaussian distribution with zero mean and standard deviation σ. Write down the log likelihood $f(\mathcal{Y}|\beta_0, \boldsymbol{\beta}, \mathcal{X})$ of the data in this model.

(3 marks)

(c) If the weights $\boldsymbol{\beta}$ are drawn from a prior distribution,

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{D} p(\beta_j),$$

write down an expression proportional to the posterior distribution

$$p(\boldsymbol{\beta}|\mathcal{X}, \mathcal{Y}, \beta_0)$$

given the data set $\mathcal{D}$.

(3 marks)

(d) Show that by choosing a Gaussian $p(\beta_j|a) \propto \exp(-\beta_j^2/a)$ or a double-exponential (Laplace) prior $p(\beta_j|a) \propto \exp(-|\beta_j|/a)$, the mode of the posterior distribution in the previous part gives the ridge regression or lasso formulation of the parameter estimation task framed as an optimisation problem. Contrast the effects of altering $a > 0$ on the coefficients $\boldsymbol{\beta}$ in ridge regression and lasso.

(5 marks)

(e) In the AdaBoost algorithm, the classification rule $F : X \to Y$ where $x \in X$ is an input vector and $Y = \{1, -1\}$ a binary output, and $F(x) = \text{sign}(\sum_m c_m f_m(x))$. Prove that the expectation, taken with respect to the conditional distribution $p(Y = y|X = x)$ of the loss function $L(y, f_m(x)) = \exp(-y f_m(x))$:

$$E_{Y|X=x}(e^{-y f_m(x)})$$

is minimised at

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

(5 marks)

Question 5

(a) **Provide** a description of the MapReduce framework. **Include** details of both the *programming model* and underlying *distributed data model*.

(8 marks)

(b) Assume that you have a large number of feature vectors stored on a distributed file system across a cluster of machines. **Describe** how you might apply the MapReduce programming model to distribute the problem of applying *k-means clustering* to the data. Also **describe** any *limitations* of your approach and any possible *workarounds*.

(12 marks)