UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMPM090**
ASSESSMENT PATTERN	:	**COMPM090A**
MODULE NAME	:	**Applied Machine Learning (Masters Level)**
DATE	:	**20-May-14**
TIME	:	**10:00**
TIME ALLOWED	:	**2 Hours 30 Minutes**

Applied Machine Learning, GI09, 2014

Answer all THREE questions.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

1. Consider the linear regression problem for training data with vector input $\mathbf{x}^n$ and scalar output y^n, $n = 1, \ldots, N$:

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

where we define the regularised total square loss

$$E(\mathbf{w}) = \sum_{n=1}^{N} \left(y^n - \mathbf{w}^{\mathsf{T}}\mathbf{x}^n\right)^2 + \lambda \mathbf{w}^{\mathsf{T}}\mathbf{w}$$

a. Derive an explicit expression for the optimal weight vector $\mathbf{w}_{opt}$ in terms of the training data and regularisation constant λ. Give also an estimation for the computational complexity required to find $\mathbf{w}_{opt}$ using this expression.

[10 marks]

b. Describe the gradient descent procedure for finding the minimum of $E(\mathbf{w})$, explaining also any potential practical advantages or disadvantages of this approach.

[3 marks]

c. Describe the Newton procedure for finding the minimum of $E(\mathbf{w})$, explaining also any potential practical advantages or disadvantages of this approach. You should give an explicit update formula for the new weight vector in terms of the old weight vector so that this could be implemented directly by a computer programmer.

[5 marks]

d. Describe the conjugate gradients procedure for finding the minimum of $E(\mathbf{w})$, explaining also any potential practical advantages or disadvantages of this approach. You should give an explicit update formula for the new weight vector in terms of the old weight vector so that this could be implemented directly by a computer programmer.

[10 marks]

e. Consider the case that each training vector $\mathbf{x}$ is sparse, with only $0 \leq s \leq 1$ of the elements of the vector being non-zero. Give estimates for the computational complexity of finding $\mathbf{w}_{opt}$ based on the above procedures, namely (batch) gradient descent, Newton's method and conjugate gradients.

[7 marks]

f. Explain what is meant by an Auto-Regressive (AR) model for a time-series $y_1, \ldots, y_T$ and derive an explicit expression for the optimal AR coefficients in the least squares sense.

[7 marks]

[Total 42 marks]

2. Consider a set of training data with vector input $\mathbf{x}^n$ and vector output $\mathbf{y}^n$. For input vector $\mathbf{x}$, a neural network produces output vector

$$\mathbf{f}(\mathbf{x}|\mathcal{W}) = \sigma^L(\mathbf{W}^L\mathbf{h}^{L-1}),$$

where the hidden layer values are recursively defined by

$$\mathbf{h}^l = \sigma^l\left(\mathbf{W}^l\mathbf{h}^{l-1}\right), \quad l = 2,\dots,L-1, \qquad \mathbf{h}^1 = \sigma^1\left(\mathbf{W}^1\mathbf{x}\right)$$

The total set of parameters is given by $\mathcal{W} = \left\{\mathbf{W}^1,\dots,\mathbf{W}^L\right\}$ and the dimension of layer l is defined by d_l where $dim(\mathbf{W}^l) = d_l \times d_{l-1}$. You may assume that the transfer functions $\sigma^1,\dots,\sigma^L$ are known.

a. For the squared loss objective function

$$E(\mathcal{W}) = \sum_{n=1}^{N}\left(\mathbf{y}^n - \mathbf{f}(\mathbf{x}^n|\mathcal{W})\right)^2$$

derive an efficient recursive procedure to compute the gradient of $E(\mathcal{W})$ with respect to all the parameters $\mathcal{W}$.

[12 marks]

b. Explain what is meant by an autoencoder neural network.

[3 marks]

c. Consider an autoencoder with a single hidden layer ($L = 2$). When the transfer function at the output layer is the identity, $\sigma^L(x) = x$, derive an expression for the optimal weight matrices $\mathbf{W}^2, \mathbf{W}^1$ and relate this to Principal Components Analysis. What would be the optimal weights for an autoencoder with a larger number of layers, $L > 2$ but with the identity transfer function on the output layer?

[10 marks]

[Total 25 marks]

3. This question concerns nearest neighbour methods.

a. i. Explain what is meant by nearest neighbour classification for a dataset of N examples, $\mathcal{D} = \{(\mathbf{x}^n, c^n), n = 1, \ldots, N\}$, for D-dimensional inputs $\mathbf{x}$ and discrete class labels c. Explain how to classify a new input $\mathbf{x}$.

[2 marks]

ii. For the Euclidian distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{D}(x_i - y_i)^2}$$

and the dataset $\mathcal{D}$ above, describe the computational complexity (both time and storage) of computing the nearest neighbour classifier for a novel input $\mathbf{x}$.

[2 marks]

iii. Explain what is meant by a metric distance and show that the Euclidean distance is a metric.

[4 marks]

b. Orchard's algorithm is a way to speed up the calculation of the nearest neighbour (for a metric distance) of a query $\mathbf{q}$ to a set of D-dimensional training vectors $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$.

i. For a metric distance show that if $d(\mathbf{q},\mathbf{x}^i) \leq \frac{1}{2}d(\mathbf{x}^i,\mathbf{x}^j)$ then $d(\mathbf{q},\mathbf{x}^i) \leq d(\mathbf{q},\mathbf{x}^j)$. Draw a picture to describe this mathematical result.

[4 marks]

ii. Explain in detail how Orchard's algorithm works.

[4 marks]

iii. Give an example dataset and query for which Orchard's algorithm will not be faster than simply explicitly evaluating the nearest neighbour by computing all distances from the query point $\mathbf{q}$.

[2 marks]

iv. Give an example dataset and query for which Orchard's algorithm will be faster than simply explicitly evaluating the nearest neighbour by computing all distances from the query point $\mathbf{q}$.

[2 marks]

c. Consider a metric distance and a set of datapoints $\mathbf{x}^i$, $i \in I$ for which $d(\mathbf{q},\mathbf{x}^i)$ has already been computed.

i. Show that we may form the lower bound

$$d(\mathbf{q},\mathbf{x}^j) \geq \max_{i \in I} \left\{ d(\mathbf{q},\mathbf{x}^i) - d(\mathbf{x}^i,\mathbf{x}^j) \right\} \equiv L_j$$

[2 marks]

ii. Explain how the Approximating and Elimination Search Algorithm (AESA) uses the above result to attempt to speed up the computation of the nearest neighbour of a query point $\mathbf{q}$.

[2 marks]

d. The KD tree is a hierarchical data-structure that can be used to potentially speed up nearest neighbour search.

i. Explain how to form a KD tree, giving an example of a dataset (of two-dimensional data) and the corresponding KD tree.

[3 marks]

ii. Consider a query vector $\mathbf{q}$. Let's imagine that we have partitioned the datapoints into those with first dimension x_1 less than a defined value v (to its 'left'), and those with a value greater or equal to v (to its 'right'):

$$\mathcal{L} = \{\mathbf{x}^n : x_1^n < v\}, \qquad \mathcal{R} = \{\mathbf{x}^n : x_1^n \geq v\}$$

Let's also say that our current best nearest neighbour candidate is $\mathbf{x}^i$ and that this point has squared Euclidean distance $\delta^2 = \left(\mathbf{q} - \mathbf{x}^i\right)^2$ from $\mathbf{q}$. Show that if $q_1 \geq v$ and $(v - q_1)^2 > \delta^2$, then $(\mathbf{x} - \mathbf{q})^2 > \delta^2$.

[3 marks]

iii. Explain how the above result can be used with the KD tree to search for the nearest neighbour of a query. Use your example KD tree above and an example query point to describe how to find the nearest neighbour using the KD tree.

[3 marks]

[Total 33 marks]

END OF PAPER