UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMPGI09**

ASSESSMENT PATTERN : **COMPGI09D**

MODULE NAME : **Applied Machine Learning**

DATE : **28-May-15**

TIME : **10:00**

TIME ALLOWED : **2 Hours 30 Minutes**

TURN OVER

Applied Machine Learning, GI09, 2015

Answer all THREE questions.

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

Calculators are NOT permitted

1. a. Explain what is meant by Forward Automatic Differentiation (AutoDiff) and give two procedures (one exact and the other an approximation) that compute the gradient of a subroutine $\mathbf{x}$ with respect to its arguments $\mathbf{x}$, giving time complexities of the approaches.

[5 marks]

b. Explain what is meant by Reverse AutoDiff. Describe the algorithm and its time complexity. Give an example to illustrate how the algorithm works.

[7 marks]

c. 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

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

The total set of parameters is given by $\mathcal{W} = \left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}$ 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 $\sigma^1,\ldots,\sigma^L$ are known.

i. 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 $\mathcal{W}$ and relate this to reverse mode AutoDiff.

[5 marks]

ii. Explain how to compute the gradient of $E(\mathcal{W})$ when parameters of the network are tied.

[3 marks]

[Total 20 marks]

2. a. Explain why the gradient of a function points along the direction of maximal change.

[2 marks]

b. Prove that for a convex function $f(\mathbf{x})$ which has a Hessian with eigenvalues less than L, then under the gradient descent procedure

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{1}{2\varepsilon T}(\mathbf{x}_1 - \mathbf{x}^*)^2$$

where $\mathbf{x}_1$ is the starting point, $\mathbf{x}_T$ is the value after $T-1$ gradient descent steps and the learning rate $\varepsilon \leq 1/L$.

[20 marks]

c. Explain what is meant by Nesterov's Accelerated Gradient procedure and compare its theoretical convergence rate with that of standard gradient descent for a convex function.

[4 marks]

d. Explain what is meant by Newton's method for optimisation of a function $f(\mathbf{x})$ and show that the position $\mathbf{x}_T$ obtained after T updates of the algorithm is invariant with respect to a linear coordinate transformation $\mathbf{x} = \mathbf{M}\mathbf{y}$.

[10 marks]

e. Explain what is meant by the Gauss-Newton optimisation method and what advantages this has over the standard Newton method.

[5 marks]

[Total 41 marks]

3. Principal Components Analysis (PCA) is a method to form a lower dimensional representation of data. For datapoints $\mathbf{x}^n, n = 1, \ldots, N$, define the matrix

$$\mathbf{X} = \left[\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\right]$$

That is, for datapoints $\mathbf{x}$ with dimension D, then $\mathbf{X}$ is $D \times N$ dimensional. The data is such that the mean is zero, that is

$$\sum_{n=1}^{N} \mathbf{x}^n = \mathbf{0}$$

The covariance matrix of the data, $\mathbf{S}$, has elements

$$S_{ij} = \frac{1}{N} \sum_{n=1}^{N} x_i^n x_j^n$$

a. Explain why PCA is often used as a pre-processing step in machine learning and explain what the geometric meaning of PCA is.

[4 marks]

b. Explain how to write $\mathbf{S}$ in terms of matrix multiplication of $\mathbf{X}$.

[3 marks]

c. PCA is typically described in terms of the eigen-decomposition of $\mathbf{S}$. Explain how this procedure works and also discuss the computational complexity of performing PCA based on directly computing the eigen-decomposition of $\mathbf{S}$.

[4 marks]

d. Consider the situation in which the datapoints $\mathbf{x}^n$ are very sparse – that is, only a few elements of each vector $\mathbf{x}^n$ are non-zero, resulting also in a sparse matrix $\mathbf{S}$. Describe a computationally efficient procedure to estimate the principal direction (the largest eigenvector of $\mathbf{S}$) and explain why this is efficient.

[4 marks]

e. Continuing with the sparse datapoint scenario, can you conceive a technique that would enable one to perform full PCA (not just the principal eigenvector) efficiently?

[4 marks]

f. An alternative way to perform PCA is based on the singular value decomposition (SVD) of the matrix $\mathbf{X}$. Explain why this is related to the eigen-decomposition of $\mathbf{S}$ and explain the computational complexity of this approach to performing PCA compared to directly computing the eigen-decomposition of $\mathbf{S}$.

[5 marks]

g. PCA can be considered as representing the i^{th} component of the n^{th} datapoint using

$$x_i^n \approx \sum_j y_j^n b_{ji}$$

where b_{ji} are the elements of the basis vectors for the j^{th} basis vector, and y_j^n is the corresponding coefficient. In the case that some of the components of x_i^n are missing, we cannot find the optimal PCA solution by the standard eigen-approach. In this case, define the least squares objective

$$E(\mathbf{B}, \mathbf{Y}) = \sum_{n=1}^{N} \sum_{i=1}^{D} \gamma_i^n \left[x_i^n - \sum_j y_j^n b_{ji} \right]^2$$

where

$$\gamma_i^n = \begin{cases} 1 & \text{if } x_i^n \text{ exists} \\ 0 & \text{if } x_i^n \text{ is missing} \end{cases}$$

Derive a procedure for minimising $E(\mathbf{B}, \mathbf{Y})$ that is guaranteed to decrease the objective function at each stage of the iteration, and for which each iteration corresponds to the solution of a set of linear equations.

[9 marks]

h. Explain the method of Fisher's Linear Discriminants and derive an explicit formula for the projection vector.

[6 marks]

[Total 39 marks]

END OF PAPER