2011 Paper 4 Question 1

Artificial Intelligence I

A perceptron takes inputs $\mathbf{x}^T = (\ x_1\ x_2\ \cdots\ x_n\ ) \in \mathbb{R}^n$ and computes its output

$$h(\mathbf{x};\mathbf{w}) = w_0 + \sum_{i=1}^{n} w_i x_i$$

using weight vector $\mathbf{w}^T = (\ w_0\ w_1\ w_2\ \cdots\ w_n\ ) \in \mathbb{R}^{n+1}$. We aim to use it to solve a regression problem using a training set $\mathbf{s}^T = (\ (\mathbf{x}_1, y_1)\ (\mathbf{x}_2, y_2)\ \cdots\ (\mathbf{x}_m, y_m)\ )$ with $y_i \in \mathbb{R}$. The approach will be to minimise the error function

$$E(\mathbf{w}) = \sum_{i=1}^{m} (y_i - h(\mathbf{x}_i, \mathbf{w}))^2$$

by gradient descent.

(*a*) Derive the gradient descent learning algorithm for this problem. [5 marks]

(*b*) The application dictates that the learning process sets as many weights as possible to zero, with the possible side effect that E is increased. It has been suggested that the error function used above might be modified by adding a further term

$$\lambda \sum_{i=0}^{n} f(w_i, \theta)$$

to E where

$$f(w,\theta) = \begin{cases} 1 & \text{if } |w| > \theta \\ 0 & \text{if } |w| \leq \theta \end{cases}$$

(*i*) Explain the purpose of the parameters λ and θ in the extra term. [4 marks]

(*ii*) Assuming we continue to use a gradient descent approach, explain why this term might be inappropriate. [1 mark]

(*c*) Suggest a function that is appropriate for a gradient descent approach, having a shape similar to that of f, and derive the associated gradient descent learning algorithm. [10 marks]