2 Artificial Intelligence (sbh11)

You have a supervised learning problem involving classification: a vector $x$ is to be assigned to one of $K$ classes. To do this you proceed in the usual way: you have a training set $s$ containing $m$ pairs $(x_i, y_i)$. However the labels $y_i$ are now vectors in $\{0, 1\}^K$ containing a single 1 representing the target class. So for example if there are 5 classes and some $x_i$ should be assigned to class 2 then $y_i = (0, 1, 0, 0, 0)$. To do this, it is proposed that you use $K$ neural networks. The $i$th network has parameters $w_i$ and computes the function $h(w_i, x)$. You may make no further assumptions regarding the function $h$.

(a) You aim to treat the output of the $i$th network as an estimate of the probability $\Pr(x \in \text{class } i|x, w)$ that $x$ should be in the $i$th class, where $w$ collects together all the $K$ vectors $w_1, \ldots, w_K$. It is proposed that to do this you should modify the setup described to compute

$$\Pr(x \in \text{class } i|x, w) = \text{prob}(i, x) = \frac{\exp(h(w_i, x))}{\sum_{j=1}^K \exp(h(w_j, x))}.$$  

Explain why this modification is required, and how it achieves the stated aim.  [4 marks]

(b) It is proposed that to train your networks, you should maximize the probability $\Pr(s|w)$ that a given collection of weights would produce the data in $s$. (You may consider the training inputs fixed.) Denote by $y_{i,j}$ the $j$th element of $y_i$. Show that training can be achieved by minimizing

$$E(w) = -\sum_{i=1}^m \sum_{j=1}^K y_{i,j} \log \text{prob}(j, x_i).$$

State any assumptions that you make.  [6 marks]

(c) You have previously applied the backpropagation algorithm for training the networks $h(w_i, x)$ and as a result of this you know how to compute derivatives $\partial h(w_i, x)/\partial w_{i,j}$ where $w_{i,j}$ is the $j$th element of $w_i$. Explain what further steps are necessary to use this knowledge to obtain derivatives of $E(w)$ with respect to the relevant weights.  [10 marks]