2 Artificial Intelligence (sbh11)

You have a supervised learning problem involving *classification*: a vector $\mathbf{x}$ is to be assigned to one of K classes. To do this you proceed in the usual way: you have a training set $\mathbf{s}$ containing m pairs $(\mathbf{x}_i, \mathbf{y}_i)$. However the labels $\mathbf{y}_i$ 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 $\mathbf{x}_i$ should be assigned to class 2 then $\mathbf{y}_i = (0,1,0,0,0)$. To do this, it is proposed that you use K neural networks. The ith network has parameters $\mathbf{w}_i$ and computes the function $h(\mathbf{w}_i, \mathbf{x})$. You may make no further assumptions regarding the function h.

(*a*) You aim to treat the output of the ith network as an estimate of the probability $\Pr(\mathbf{x} \in \text{class } i|\mathbf{x}, \mathbf{w})$ that $\mathbf{x}$ should be in the ith class, where $\mathbf{w}$ collects together all the K vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$. It is proposed that to do this you should modify the setup described to compute

$$\begin{aligned}\Pr(\mathbf{x} \in \text{class } i|\mathbf{x}, \mathbf{w}) &= \text{prob}(i, \mathbf{x}) \\ &= \frac{\exp(h(\mathbf{w}_i, \mathbf{x}))}{\sum_{j=1}^{K} \exp(h(\mathbf{w}_j, \mathbf{x}))}.\end{aligned}$$

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 $\Pr(\mathbf{s}|\mathbf{w})$ that a given collection of weights would produce the data in $\mathbf{s}$. (You may consider the training inputs fixed.) Denote by $y_{i,j}$ the jth element of $\mathbf{y}_i$. Show that training can be achieved by minimizing

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

State any assumptions that you make. [6 marks]

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