# CIE6032 and MDS6232 Homework 1 Solution

October 2018

## Problem 1

**[25 Points]**

Cross entropy is often used as objective function when training neural networks in classification problems. Suppose the training set includes $N$ training pairs $D = \{(\mathbf{x}_i^{(\text{train})}, \quad y_i^{(\text{train})})\}_{i=1}^N$, where $\mathbf{x}_i^{(\text{train})}$ is a training sample and $y_i^{(\text{train})} \in \{1, \ldots, c\}$ is its class label. $\mathbf{z}_i$ is the output of the network given input $\mathbf{x}_i^{(\text{train})}$ and the nonlinearity of the output layer is softmax. $\mathbf{z}_i$ is a $c$ dimensional vector, $z_{i,k} \in [0, 1]$ and $\sum_{k=1}^c z_{i,k} = 1$. Please
(1) write the objective function of cross entropy with softmax activation function, and the gradient of hidden-to-output weights (**Hits: two conditions of softmax**) (20 points)
(2) it is equivalent to the negative log-likelihood on the training set, assuming the training samples are independent. (5 points)

**Answer:**
(1). Donate the training label $y_i$ by an one-hot target vector $\mathbf{t}_i \in \{0, 1\}^c$, where $\mathbf{t}_{i,y_i} = 1$ and $\mathbf{t}_i^T \mathbf{t}_i = 1$.

Then the cross-entropy between the two distributions $\mathbf{t}_i$ and $\mathbf{z}_i$ is thus

$$L(\mathbf{t}_i, \mathbf{z}_i) = -\sum_{k=1}^{c} \mathbf{t}_{i,k} \log \mathbf{z}_{i,k}. \qquad (\textbf{5 points}) \tag{1}$$

Donate hidden-to-output transformation before and after softmax activation function as $a_k = W_{kj} h_j + b_k$ and $z_k = \text{softmax}(a_k) = \frac{a_k}{\sum_{l=1}^c a_l}$, for the derivative of softmax activation function, we have

$$\frac{\partial z_k}{\partial a_k} = \begin{cases} z_k(1 - z_l) & k = l \qquad (\textbf{5 points}) \\ -z_k z_l & k \neq l \qquad (\textbf{5 points}) \end{cases} \tag{2}$$

Then through chain-rule, we have the derivative of hidden-to-output weights as follows.

$$\frac{\partial L}{\partial W_{kj}} = \frac{\partial L}{\partial z_k} \cdot \frac{\partial z_k}{\partial a_k} \cdot \frac{\partial a_k}{\partial W_{kj}}$$
$$= (-\sum_{k=1}^{c} t_k \frac{1}{z_k}) \cdot \frac{\partial z_k}{\partial a_k} \cdot h_j \quad (3)$$

Then import Eq. (1) and Eq. (2) into Eq. (3), we have

$$\frac{\partial L}{\partial W_{kj}} = (-\frac{t_k}{z_k} z_k(1 - z_k) + \sum_{l \neq k} \frac{t_l}{z_l} z_k z_l) \cdot h_j$$
$$= (z_k - t_k) \cdot h_j \quad \textbf{(5 points)} \quad (4)$$

(2). Eq. (1) can be rewritten as

$$L(\mathbf{t}_i, \mathbf{z}_i) == -\log \mathbf{z}_{i,y_i} = -\log P(y = y_i | \mathbf{x}_i), \quad (5)$$

which is the negative log-likelihood of the $i$-th training sample. As the training samples are independent, it can be extended to the whole training set. **(5 points)**

# Problem 2

**[25 points]**

$x_1$ and $x_2$ are two input variables, and $y$ is the target variable to be predicted. The network structure is shown in Figure 1(a). $h_{11} = f_{11}(x_1)$, $h_{12} = f_{12}(x_2)$, and $y = g(h_{11}, h_{12})$.

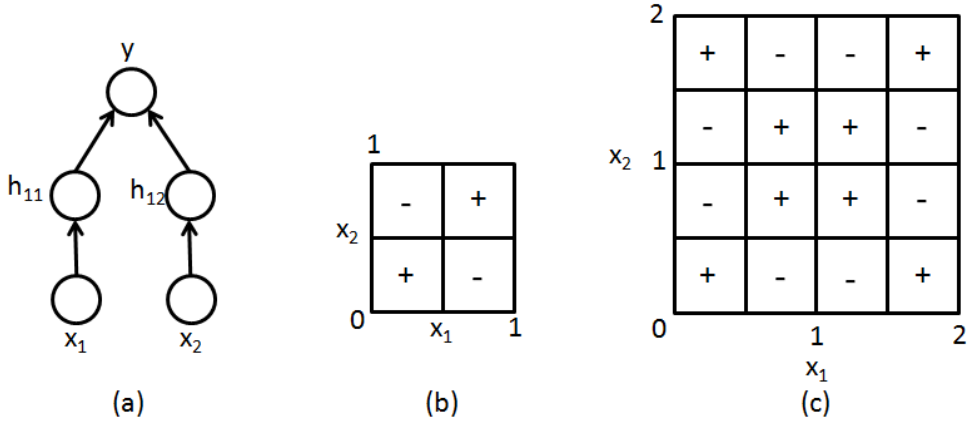


Figure 1: Problem 2

- Assuming $x_1 \in [0,1]$ and $x_2 \in [0,1]$, in order to obtain the decision regions in Figure 1(b), decide functions $f_{11}$, $f_{12}$, and $g$. [**5 points**]

**Answer:**
A possible solution could be

$$f_{11}(x_1) = \begin{cases} 1 & \text{if } x_1 \geq 0.5 \\ -1 & \text{if } x_1 < 0.5 \end{cases}, \quad f_{12}(x_2) = \begin{cases} 1 & \text{if } x_2 \geq 0.5 \\ -1 & \text{if } x_2 < 0.5 \end{cases}, \quad \text{and } g(h_{11}, h_{12}) = h_{11} h_{12}. \tag{6}$$

- Now we extend the range of $x_1$ and $x_2$ to $[0,2]$. Please add one more layer to Figure 1(a) in order to obtain the decision regions in Figure 1(c). [**5 points**]

**Answer:**
Note that the decision regions are symmetric with respect to $x_1 = 1$ and $x_2 = 1$. Then a possible solution could be adding a layer $z$ between $h$ and $x$ that makes

$$z_i = 1 - |x_i - 1| = \begin{cases} 2 - x_i & \text{if } x_i \geq 1 \\ x_i & \text{if } x_i < 1 \end{cases}, \quad \text{for } i = 1, 2. \tag{7}$$

- Although the decision boundaries in Figure 1(c) look complicated, there exist regularity and global structure. Please explain such regularity and global structure. Based on your observation, draw the decision boundaries when the range of $x_1$ and $x_2$ are extended to [0, 4]. [**5 points**]

**Answer:**
As pointed out in the previous answer, we can extend the symmetry property by "unfolding" Figure 1(c) again to $[0,4]^2$, which results in the decision regions shown in Figure 2.

- Following the question above and assuming the range of $x_1$ and $x_2$ is extended to $[0, 2^n]$, draw the network structure and the transform function in each layer, in order to obtain the decision regions with the same regularity and global structure in Figure 1 (b) and (c). The complexity of computation units should be $O(n)$. [**5 points**]

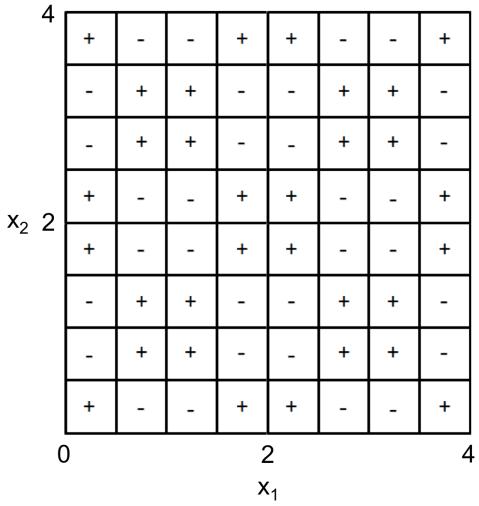


Figure 2: Answer to problem 2

**Answer:** We could insert $n$ consecutive layers $z^{(1)}, \cdots z^{(n)}$ between $x$ and $h$, such that

$$z_i^{(j)} = \begin{cases} 2^j - z_i^{(j-1)} & \text{if } z_i^{(j-1)} \geq 2^{j-1} \\ z_i^{(j-1)} & \text{if } z_i^{(j-1)} < 2^{j-1} \end{cases}, \quad \text{for } j = 1, \cdots, n, \text{and } i = 1, 2, \tag{8}$$

and let $z^{(0)} = x$ for convenience.

- Assuming the range of $x_1$ and $x_2$ is $[0, 2^n]$ and only one hidden layer is allowed, specify the network structure and transform functions. [**5 points**]

**Answer:** The pattern of decision regions also has a period of 2 along each axis, meaning that the decision regions for $x$ are the same as ($x \mod 2$). Thus formally, one possible hidden layer function could be

$$h_{1i}(x_i) = f_{1i}(1 - |x_i - 2\lfloor x_i/2 \rfloor - 1|), \text{for } i = 1, 2, \tag{9}$$

where $f_{1i}$ is the same as defined in Eq. (6).