Flipped Classroom Questions on Feedforward Neural Networks
Michael Collins

Question 1: Consider a neural network

$$\phi(x;\theta) = g(Wx + b)$$

where $x \in \mathbb{R}^d$, $W \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$, and g is a transfer function defined as

$$g(z) = \alpha \times z + c$$

where $\alpha \in \mathbb{R}$ is a constant, and $c \in \mathbb{R}^m$ is a vector.

The following relationship will be useful in this question: given vectors v and x, and a matrix A,

$$v \cdot (Ax) = v' \cdot x$$

where $v' = A^\top v$.

Proof:

$$\underbrace{v}_{m\times 1} \cdot \left(\underbrace{A}_{m\times d}\underbrace{x}_{d\times 1}\right) = \underbrace{v^\top}_{1\times m}\underbrace{A}_{m\times d}\underbrace{x}_{d\times 1} = (v^\top A)x = (A^\top v)^\top x = v' \cdot x$$

where $v' = \underbrace{A^\top}_{d\times m}\underbrace{v}_{m\times 1}$

Question 1a: Now say we define a model

$$p(y|x;\theta,v) = \frac{\exp\{v(y)\cdot\phi(x;\theta) + \gamma_y\}}{\sum_{y'}\exp\{v(y')\cdot\phi(x;\theta) + \gamma_{y'}\}}$$

Show that for any parameter values $v(y)$ and γ_y for $y \in \mathcal{Y}$, there are parameter values $v'(y)$ and γ'_y such that for all x, y,

$$p(y|x;\theta,v) = \frac{\exp\{v'(y)\cdot x + \gamma'_y\}}{\sum_{y'}\exp\{v'(y')\cdot x + \gamma'_{y'}\}}$$

Question 1b: Now assume the transfer function is

$$g(z) = Az + c$$

where $A \in \mathbb{R}^{m\times m}$ is a matrix, and $c \in \mathbb{R}^m$ is a vector. Show that under this model, for any parameter values $v(y)$ and γ_y for $y \in \mathcal{Y}$, there are parameter values $v'(y)$ and γ'_y such that for all x, y,

$$p(y|x;\theta,v) = \frac{\exp\{v'(y)\cdot x + \gamma'_y\}}{\sum_{y'}\exp\{v'(y')\cdot x + \gamma'_{y'}\}}$$

Question 1c: Now assume we have an instance of the XOR problem, with examples

$$x = [0, 0] \quad y = -1$$

$$x = [0, 1] \quad y = +1$$

$$x = [1, 0] \quad y = +1$$

$$x = [1, 1] \quad y = -1$$

Show geometrically why a neural network with two neurons and transfer function

$$g(z) = \alpha \times z + c$$

or

$$g(z) = Az + c$$

fails to model this data.

Question 2: Consider the function LS : $\mathbb{R}^m \rightarrow \mathbb{R}^m$ that maps a vector $l \in \mathbb{R}^m$ to a vector $\text{LS}(l) \in \mathbb{R}^m$ with the following components:

$$\text{LS}_y(l) = l_y - \log \sum_{y'} \exp\{l_{y'}\}$$

We will refer to this as the "log softmax function".

Question 2a: What is the value for

$$\frac{\partial \text{LS}_y(l)}{\partial l_y}$$

for each value of y?

What is the value for

$$\frac{\partial \text{LS}_y(l)}{\partial l_{y'}}$$

for any y, y' such that $y \neq y'$?

Question 2b: Now consider the following sequence of equations that defines the value of the output o given an input x^i and label y^i:

$$\begin{aligned} z \in \mathbb{R}^m &= Wx^i + b \\ h \in \mathbb{R}^m &= g(z) \\ l \in \mathbb{R}^K &= Vh + \gamma \\ q \in \mathbb{R}^K &= \text{LS}(l) \\ o \in \mathbb{R} &= -q_{y_i} \end{aligned}$$

Here we define $K = |\mathcal{Y}|$ where $\mathcal{Y}$ is the set of possible labels. Here V is a matrix of parameters $V \in \mathbb{R}^{K \times m}$, and $\gamma \in \mathbb{R}^K$.

Recall also that for a scalar $z = w \cdot x + b$, and a scalar $h = g(z)$ for some transfer function, we have:

$$\frac{dh}{dw_j} = \frac{dg(z)}{dz} x_j$$

We can write the derivative of o with respect to parameter $W_{j,k}$ using the chain rule:

$$\frac{\partial o}{\partial W_{j,k}} = \sum_y \frac{\partial o}{q_y} \frac{\partial q_y}{\partial W_{j,k}}$$

$$\frac{\partial q_y}{\partial W_{j,k}} = \sum_{y'} \frac{\partial q_y}{\partial l_{y'}} \frac{\partial l_{y'}}{\partial W_{j,k}}$$

Complete the following expressions:

$$\frac{\partial l_{y'}}{\partial W_{j,k}} = \sum_{k=1}^{m}$$

$$\frac{\partial o}{\partial q_y} =$$

$$\frac{\partial q_y}{\partial l_{y'}} =$$

$$\frac{\partial l_{y'}}{\partial h_k} =$$

One hint: note that with $l = Vh + \gamma$, we have

$$l_y = \sum_{k=1}^{m} V_{y,k} h_k + \gamma_y$$

Question 3: Assume we have a model with input $f(x) \in \mathbb{R}^d$, and parameters $v(y) \in \mathbb{R}^d$ for each label y. The set of possible labels is $\mathcal{Y} = \{1, 2, \ldots K\}$. Give definitions of a feature vector $f(x, y)$ such that for all x, y,

$$v(y) \cdot f(x) = w \cdot f(x, y)$$

where v is the concatenation of parameter vectors

$$w = [v(1); v(2); \ldots; v(K)]$$

Now assume that in addition to the $v(y)$ parameters, we have a parameter $\gamma_y \in \mathbb{R}$ for each label y. How would you define $f(x, y)$ and w so that for all x, y

$$v(y) \cdot f(x) + \gamma_y = f(x, y) \cdot w$$

Question 4: Assume we have an instance of the XOR problem, with examples

$$x = [0, 0] \quad y = -1$$

$$x = [0, 1] \quad y = +1$$

$$x = [1, 0] \quad y = +1$$

$$x = [1, 1] \quad y = -1$$

Assume that we have a neural network with three neurons, which take values

$$h_1 = x_1, \quad h_2 = x_2, \quad h_3 = x_1 \times x_2, \quad h_4 = x_1^2, \quad h_5 = x_2^2$$

where $[x_1, x_2]$ is the input vector. Is it possible to model the data using these definitions of the neurons?