IFT6266 January 15, 2015

First assignment - Solution

Problem setting

Observations

Let

$$\mathbf{x} = [x_1, \ldots, x_N] \tag{1}$$

be a row vector of features and let

$$\mathbf{t} = [t_1, \ldots, t_C] \tag{2}$$

be a one-hot encoded row vector corresponding to the class of $\mathbf{x}$, *i.e.*

$$t_k \in \{0, 1\} \quad \forall k \tag{3}$$

and

$$\sum_{k=1}^{C} t_k = 1 \tag{4}$$

such that $t_k = 1$ if and only if $\mathbf{x}$ belongs to class k.

Model

Let $\mathbf{W}$ be a $N \times H$ matrix, $\mathbf{V}$ be a $H \times C$ matrix, $\mathbf{b}$ be a H-dimensional row vector and $\mathbf{d}$ be a C-dimensional row vector.

We define a one-hidden-layer MLP classifier as follows: let

$$\mathbf{h} = \sigma(\mathbf{x}\mathbf{W} + \mathbf{b}) \tag{5}$$

where

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

is the sigmoid elementwise nonlinearity, and let

$$\mathbf{y} = \mathrm{softmax}(\mathbf{h}\mathbf{V} + \mathbf{d}) \tag{7}$$

where

$$\mathrm{softmax}(\mathbf{z}) = \frac{e^{\mathbf{z}}}{e^{\mathbf{z}} \cdot \mathbf{1}} = \frac{e^{\mathbf{z}}}{\sum_i e^{z_i}} \tag{8}$$

is a normalized version of the exponential elementwise nonlinearity. Finally, let

$$\mathcal{L} = -\mathbf{t} \cdot \log \mathbf{y} = -\sum_{k=1}^{C} t_k \log y_k \tag{9}$$

be the loss function of the MLP classifier.

Solution

Function derivatives

Sigmoid

$$\begin{aligned}
\frac{d}{dz}\sigma(z) &= \frac{d}{dz}(1+e^{-z})^{-1} \\
&= -(1+e^{-z})^{-2}. - e^{-z} \\
&= \frac{1}{1+e^{-z}}\frac{e^{-z}+1-1}{1+e^{-z}} \\
&= \frac{1}{1+e^{-z}}\left(1-\frac{1}{1+e^{-z}}\right) \\
&= \sigma(z)(1-\sigma(z))
\end{aligned} \tag{10}$$

Softmax

Let $\mathbf{s} = \text{softmax}(\mathbf{z})$ be the softmax function. Then

$$\begin{aligned}
\frac{\partial s_k}{\partial z_l} &= \frac{\partial}{\partial z_l}\frac{e^{z_k}}{\sum_r e^{z_r}} \\
&= \frac{\delta_{k,l}e^{z_k} - e^{z_k}e^{z_l}}{\left(\sum_r e^{z_r}\right)^2} \\
&= s_k(\delta_{k,l} - s_l)
\end{aligned} \tag{11}$$

Scalar derivatives

Derivatives with respect to y

$$\frac{\partial \mathcal{L}}{\partial y_k} = \frac{\partial}{\partial y_k}\sum_{r=1}^{C} -t_r \log y_r = -\frac{t_k}{y_k} \tag{12}$$

Derivatives with respect to h

$$\begin{aligned}
\frac{\partial y_k}{\partial h_j} &= \sum_{r=1}^{C} \frac{\partial y_k}{\partial(\sum_{s=1}^{H} h_s V_{s,r} + d_r)} \frac{\partial(\sum_{s=1}^{H} h_s V_{s,r} + d_r)}{\partial h_j} \\
&= \sum_{r=1}^{C} y_k(\delta_{k,r} - y_r)V_{j,r} \\
&= y_k V_{j,k} - y_k \sum_{r=1}^{C} y_r V_{j,r}
\end{aligned} \tag{13}$$

This means

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_j} &= \sum_{k=1}^{C} \frac{\partial \mathcal{L}}{\partial y_k} \frac{\partial y_k}{\partial h_j} \\
&= \sum_{k=1}^{C} -t_k \left(V_{j,k} - \sum_{r=1}^{C} y_r V_{j,r} \right) \\
&= \sum_{k=1}^{C} -t_k V_{j,k} + \left[\sum_{k=1}^{C} t_k \right] \left[\sum_{r=1}^{C} y_r V_{j,r} \right] \\
&= \sum_{k=1}^{C} -t_k V_{j,k} + \sum_{r=1}^{C} y_r V_{j,r} \quad \text{(because } \sum_k t_k = 1 \text{ by definition)} \\
&= \sum_{k=1}^{C} V_{j,k} (y_k - t_k)
\end{aligned} \tag{14}$$

Derivatives with respect to V

$$\begin{aligned}
\frac{\partial y_r}{\partial V_{j,k}} &= \frac{\partial y_r}{\partial (\sum_{s=1}^{H} h_s V_{s,k} + d_k)} \frac{\partial (\sum_{s=1}^{H} h_s V_{s,k} + d_k)}{\partial V_{j,k}} \\
&= y_r (\delta_{k,r} - y_k) h_j
\end{aligned} \tag{15}$$

This means

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial V_{j,k}} &= \sum_{r=1}^{C} \frac{\partial \mathcal{L}}{\partial y_r} \frac{\partial y_r}{\partial V_{j,k}} \\
&= \sum_{r=1}^{C} -t_r (\delta_{k,r} - y_k) h_j \\
&= \left[\sum_{r=1}^{C} t_r \right] y_k h_j - t_k h_j \\
&= (y_k - t_k) h_j \quad \text{(because } \sum_r t_r = 1 \text{ by definition)}
\end{aligned} \tag{16}$$

Derivatives with respect to d

$$\begin{aligned}
\frac{\partial y_r}{\partial d_k} &= \frac{\partial y_r}{\partial (\sum_{s=1}^{H} h_s V_{s,k} + d_k)} \frac{\partial (\sum_{s=1}^{H} h_s V_{s,k} + d_k)}{\partial d_k} \\
&= y_r (\delta_{k,r} - y_k)
\end{aligned} \tag{17}$$

This means

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial d_k} &= \sum_{r=1}^{C} \frac{\partial \mathcal{L}}{\partial y_r} \frac{\partial y_r}{\partial d_k} \\
&= \sum_{r=1}^{C} -t_r(\delta_{k,r} - y_k) \\
&= \left[\sum_{r=1}^{C} t_r \right] y_k - t_k \\
&= y_k - t_k \quad \text{(because } \sum_r t_r = 1 \text{ by definition)}
\end{aligned} \tag{18}
$$

Derivatives with respect to W

$$
\begin{aligned}
\frac{\partial h_j}{\partial W_{i,j}} &= \frac{\partial h_j}{\partial (\sum_{s=1}^{A} x_s W_{s,j} + b_j)} \frac{\partial (\sum_{s=1}^{A} x_s W_{s,j} + b_j)}{\partial W_{i,j}} \\
&= h_j(1 - h_j) x_i
\end{aligned} \tag{19}
$$

This means

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_{i,j}} &= \frac{\partial \mathcal{L}}{\partial h_j} \frac{\partial h_j}{\partial W_{i,j}} \\
&= \sum_{k=1}^{C} V_{j,k}(y_k - t_k) h_j (1 - h_j) x_i
\end{aligned} \tag{20}
$$

Derivatives with respect to b

$$
\begin{aligned}
\frac{\partial h_j}{\partial b_j} &= \frac{\partial h_j}{\partial (\sum_{s=1}^{A} x_s W_{s,j} + b_j)} \frac{\partial (\sum_{s=1}^{A} x_s W_{s,j} + b_j)}{\partial b_j} \\
&= h_j(1 - h_j)
\end{aligned} \tag{21}
$$

This means

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b_j} &= \frac{\partial \mathcal{L}}{\partial h_j} \frac{\partial h_j}{\partial b_j} \\
&= \sum_{k=1}^{C} V_{j,k}(y_k - t_k) h_j (1 - h_j)
\end{aligned} \tag{22}
$$

Matrix and vector derivatives

From previous results, it is straightforward to verify that

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{V}} &= \mathbf{h}^T(\mathbf{y}-\mathbf{t}), \\
\frac{\partial \mathcal{L}}{\partial \mathbf{d}} &= \mathbf{y}-\mathbf{t}, \\
\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \mathbf{x}^T\left[(\mathbf{y}-\mathbf{t})\mathbf{V}^T \odot \mathbf{h} \odot (\mathbf{1}-\mathbf{h})\right], \\
\frac{\partial \mathcal{L}}{\partial \mathbf{b}} &= (\mathbf{y}-\mathbf{t})\mathbf{V}^T \odot \mathbf{h} \odot (\mathbf{1}-\mathbf{h})
\end{aligned}
\tag{23}
$$