CSCI 567: Theory Assignment 2

Anurima Anil Padwal
USC Id: 4348819703

6 October 2019

Solutions

1.1. Given $y_i \mathbf{w}_k^T \mathbf{x}_i < 0$
To prove: $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\|$
Proof:
The perceptron weight update rule for a misclassification is given by:

$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k + y_i \mathbf{x}_i \\ \mathbf{w}_{k+1}^T \mathbf{w}_{opt} &= (\mathbf{w}_k + y_i \mathbf{x}_i)^T \mathbf{w}_{opt} \\ &= \mathbf{w}_k^T \mathbf{w}_{opt} + y_i \mathbf{x}_i^T \mathbf{w}_{opt} \\ &\geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \end{aligned}$$

The inequality follows from the fact that for $\mathbf{w}_{opt}$, the distance of any $\mathbf{x}_i$ from $\mathbf{w}_{opt}$ must be at least γ, i.e. $y_i(\mathbf{x}_i^T \mathbf{w}_{opt}) = |\mathbf{x}_i^T \mathbf{w}_{opt}| \geq \gamma$ and $\|\mathbf{w}_{opt}\| = 1$.
Hence, proved.

1.2. Given $y_i \mathbf{w}_k^T \mathbf{x}_i < 0$
To prove: $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$
Proof:
The perceptron weight update rule for a misclassification is given by:

$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k + y_i \mathbf{x}_i \\ \|\mathbf{w}_{k+1}\|^2 &= \|\mathbf{w}_k + y_i \mathbf{x}_i\|^2 \\ &= (\mathbf{w}_k + y_i \mathbf{x}_i)^T (\mathbf{w}_k + y_i \mathbf{x}_i) \\ &= \mathbf{w}_k^T \mathbf{w}_k + y_i \mathbf{w}_k^T \mathbf{x}_i + y_i \mathbf{x}_i^T \mathbf{w}_k + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \\ &= \mathbf{w}_k^T \mathbf{w}_k + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Now $\|\mathbf{w}_k\|^2 = \mathbf{w}_k^T \mathbf{w}_k$, $\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i = 1$,$y_i \mathbf{w}_k^T \mathbf{x}_i < 0$ since we only perform the update when $\mathbf{x}_i$ is misclassified.

$$\begin{aligned} \therefore \|\mathbf{w}_{k+1}\|^2 &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + 1 \\ &\leq \|\mathbf{w}_k\|^2 + 1 \end{aligned}$$

Hence proved.

1.3. From 1.1, it follows that every time we make a mistake, the dot product of the weight vector with the target increases by at least γ. So, after M mistakes, we have $\mathbf{w}_{M+1}^T \mathbf{w}_{opt} \geq M\gamma$.

From 1.2, it follows that every time we make a mistake the length squared of our weight vector increases by at the most 1. So, after M mistakes, $\|\mathbf{w}_{M+1}\|^2 \leq M$

$$
\begin{aligned}
&\|\mathbf{w}_{M+1}\| \leq \sqrt{M} \qquad (1)\\
&\mathbf{w}_{M+1}^T \mathbf{w}_{opt} \geq M\gamma\\
\texttt{Using Cauchy's inequality,}\quad&\\
&\|\mathbf{w}_{M+1}\|\|\mathbf{w}_{opt}\| \geq M\gamma\\
&\|\mathbf{w}_{opt}\| = 1\\
&\|\mathbf{w}_{M+1}\| \geq M\gamma \qquad (2)\\
\texttt{From (1) and (2),}\quad&\\
&\gamma M \leq \|\mathbf{w}_{M+1}\| \leq \sqrt{M}
\end{aligned}
$$

Hence proved.

1.4. From 1.3,

$$
\begin{aligned}
M\gamma &\leq \sqrt{M}\\
\sqrt{M} &\leq \gamma^{-1}\\
\texttt{Squaring both sides,}&\\
M &\leq \gamma^{-2}
\end{aligned}
$$

This we have proved that the perceptron algorithm makes finite number of mistakes that is at the most γ^{-2}, and hence it must converge.

2.1. We have to minimize the cross entropy loss function to solve for w and b.

$$
\begin{aligned}
min_{\mathbf{w},b} L(\mathbf{w}, b) &= min_{\mathbf{w},b} - \sum_n \{y_n \log\left[p(y_n = 1|x_n)\right] + (1 - y_n)\log\left[p(y_n = 0|x_n)\right]\}\\
&= min_{\mathbf{w},b} - \sum_n \{y_n \log\left[\sigma(\mathbf{w}^T\mathbf{x}_n + b)\right] + (1 - y_n)\log\left[1 - \sigma(\mathbf{w}^T\mathbf{x}_n + b)\right]\}
\end{aligned}
$$

For simplicity, let $h(\mathbf{z_n}) = \sigma(\mathbf{w}^T\mathbf{x}_n + b)$. Now

$$\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \\
\frac{\partial\sigma(z)}{\partial z} &= \frac{(1+e^{-z})(0) - 1(e^{-z})(-1)}{(1+e^{-z})^2} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}-1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}}[1 - \frac{1}{1+e^{-z}}] \\
&= \sigma(z)(1-\sigma(z))
\end{aligned}$$

Thus $h'(\mathbf{z_n}) = h(\mathbf{z}_n)(1 - h(\mathbf{z}_n))$

$$min_{\mathbf{w},b} L(\mathbf{w}, b) = min_{\mathbf{w},b} - \sum_{i=1}^{n} \{y_n \log h(\mathbf{z}_n) + (1-y_n) \log{(1-h(\mathbf{z}_n))}\}$$

Take the partial derivative of L with respect to $\mathbf{w}$,

$$\begin{aligned}
\frac{\partial L}{\partial w} &= -\sum_n \{y_n \frac{h(\mathbf{z}_n)}{h(\mathbf{z}_n)}(1-h(\mathbf{z}_n))\mathbf{x}_n + (1-y_n)\frac{(-1)h(\mathbf{z}_n)(1-h(\mathbf{z}_n))}{1-h(\mathbf{z_n})}\mathbf{x}_n\} \\
&= -\sum_n \{(y_n(1-h(\mathbf{z}_n)) - (1-y_n)h(\mathbf{z}_n))\mathbf{x}_n\} \\
&= -\sum_n \{(y_n - y_n h(\mathbf{z}_n) - h(\mathbf{z}_n) + y_n h(\mathbf{z}_n))\mathbf{x}_n\} \\
&= -\sum_n \{(y_n - h(\mathbf{z}_n))\mathbf{x}_n\}
\end{aligned}$$

`Resubstituting,`

$$= -\sum_n \{(y_n - \sigma(\mathbf{w}^T\mathbf{x}_n + b))\mathbf{x}_n\}$$

Take the partial derivative of L with respect to b,

$$\begin{aligned}\frac{\partial L}{\partial b} &= -\sum_n \{y_n \frac{h(\mathbf{z}_n)}{h(\mathbf{z}_n)}(1-h(\mathbf{z}_n) + (1-y_n)\frac{(-1)h(\mathbf{z}_n)(1-h(\mathbf{z}_n))}{1-h(\mathbf{z_n})}\} \\ &= -\sum_n \{(y_n(1-h(\mathbf{z}_n) - (1-y_n)h(\mathbf{z}_n))\} \\ &= -\sum_n \{(y_n - y_n h(\mathbf{z}_n) - h(\mathbf{z}_n) + y_n h(\mathbf{z}_n))\} \\ &= -\sum_n \{(y_n - h(\mathbf{z}_n))\}\end{aligned}$$

`Resubstituting,`

$$= -\sum_n \{(y_n - \sigma(\mathbf{w}^T \mathbf{x}_n + b))\}$$

The Gradient Descent update rule for $\mathbf{w}$ is:

$$\begin{aligned}\mathbf{w}_{i+1} &= \mathbf{w}_i - \alpha * \frac{\partial L}{\partial w} \\ &= \mathbf{w}_i + \alpha * \sum_{i=1}^{n} \{(y_n - \sigma(\mathbf{w}^T \mathbf{x}_n + b))\mathbf{x}_i\}\end{aligned}$$

where α is learning rate.

2.2. $p(y=1|x) = \sigma(\mathbf{w}^T \mathbf{x})$
$p(y=0|x) = 1 - \sigma(\mathbf{w}^T \mathbf{x})$
Initially, we set $w = 0$ and learning rate $= 0.001$.

$$\sum_{i=1}^{4} (y_i - \sigma(w_i.x_i))x_i$$

$$\begin{aligned}&= (0 - \frac{1}{1+e^{-1*0}})1 + (1 - \frac{1}{1+e^{-1*0}})1 + (1 - \frac{1}{1+e^{-1*0}})1 + (1 - \frac{1}{1+e^{-1*0}})1 \\ &= 1\end{aligned}$$

Updating w,

$$\begin{aligned}w_1 &= w_0 + \texttt{learning rate} * \sum_{i=1}^{4} (y_i - \sigma(w_i.x_i))x_i \\ &= 0 + 0.001 * 1 \\ &= 0.001\end{aligned}$$

For point $(x_1, y_1) = (1, 0)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$

For point $(x_2, y_2) = (1, 1)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$
For point $(x_3, y_3) = (1, 1)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$
For point $(x_4, y_4) = (1, 1)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$

The point (x_1, y_1) is misclassified while all other points are correctly classified. Hence the training accuracy after one batch iteration is $3/4 = 0.75$ or 75%

2.3.For point $(x_1, y_1) = (-1, 0)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{0.001*1}} = 0.49975 < 0.5$. So our prediction is $y* = 0$
For point $(x_2, y_2) = (1, 1)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$
For point $(x_3, y_3) = (1, 0)$, we have the prediction as $\sigma((w.x) = \frac{1}{1+e^{-0.001*1}} = 0.50025 > 0.5$. So our prediction is $y* = 1$
Hence, only one point, i.e. (x_3, y_3) is misclassified, and our test accuracy is 2/3 = 0.6667 or 66.67%.

3. $$L(y, \hat{y}) = -\sum_{j=1}^{3} y_j \log \hat{y}_j$$

$$\frac{\partial L}{\partial \hat{y}_j} = \frac{-y_j}{\hat{y}_j} \quad \text{——— (1)}$$

$$\frac{\partial L}{\partial v_{jk}} = \sum_{i=1}^{3} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial v_{jk}} \quad \text{——— (2)}$$

$$\hat{y}_i = \text{SOFTMAX}(o_i)$$

$$= \frac{e^{o_i}}{\sum_{m=1}^{3} e^{o_m}}$$

$$\frac{\partial L}{\partial v_{jk}} = \sum_{i=1}^{3} \frac{\partial L}{\partial \hat{y}_i} \left[\sum_{l=1}^{3} \frac{\partial \hat{y}_i}{\partial o_l} \cdot \frac{\partial o_l}{\partial v_{jk}} \right] \quad \text{——— (3)}$$

$$\frac{\partial \hat{y}_i}{\partial o_l} = \frac{\partial}{\partial o_l} \left(\frac{e^{o_i}}{\sum_{m=1}^{3} e^{o_m}} \right)$$

$$\frac{\partial \hat{y}_i}{\partial o_l} = \frac{\left(\sum_{m=1}^{3} e^{o_m}\right) e^{o_i} - e^{o_i} \cdot e^{o_l}}{\left(\sum_{m=1}^{3} e^{o_m}\right)^2} \quad (i == l)$$

$$= \frac{e^{o_i}}{\sum_{m} e^{o_m}} - \left(\frac{e^{o_i}}{\sum_{m=1}^{3} e^{o_m}} \right)^2$$

$$= \hat{y}_i - (\hat{y}_i)^2$$

$$= \hat{y}_i (1 - \hat{y}_i) \quad \text{——— (4)}$$

$$\frac{\partial \hat{y}_i}{\partial o_l} = \frac{\left(\sum_{m=1}^{3} e^{o_m}\right) 0 - e^{o_i} \cdot e^{o_l}}{\sum_{m=1}^{3} e^{o_m}} \quad (i \neq l)$$

$$= -\hat{y}_i \cdot \hat{y}_l \quad \text{——— (5)}$$

$$o_l = \sum_{k=1}^{4} v_{lk} z_k$$

$$\frac{\partial o_l}{\partial v_{pq}} = z_k I(q == k) \quad \text{——— (6)}$$

where $I(\cdot)$ is identity function.

Date:

$$\frac{\partial L}{\partial v_{jk}} = -\sum_{i=1}^{3} \frac{y_i}{\hat{y_i}} \sum_{l=1}^{3} \frac{\partial \hat{y_i}}{\partial o_l} \cdot \frac{\partial o_l}{\partial v_{jk}}$$

$$= -\sum_{i=1}^{3} \left\{ \frac{y_i}{\hat{y_i}} \sum_{l=1}^{3} \left[\hat{y_i}(1-\hat{y_i}) I(l==i) + (-\hat{y_i}\hat{y_l}) I(i \neq l) \right] * z_{k'} \cdot I(k'==k) \right\}$$

$$\text{where } O_l = \sum_{k'=1}^{4} V_{lk'} \cdot z_{k'}$$

$$\therefore \frac{\partial L}{\partial v_{jk}} = -\sum_{i=1}^{3} \left\{ y_i \cdot \sum_{l=1}^{3} \left[(1-\hat{y_i}) I(l==i) - \hat{y_l} I(i \neq l) \right] \cdot z_{k'} I(k'==k) \right.$$

* To compute $\frac{\partial L}{\partial w_{ki}}$

$$\frac{\partial L}{\partial w_{ki}} = \sum_{m=1}^{3} \left[\frac{\partial L}{\partial y_m} \sum_{n=1}^{4} \left(\frac{\partial y_m}{\partial z_k} \cdot \frac{\partial z_n}{\partial w_{ki}} \right) \right]$$

$$= \sum_{m=1}^{3} \frac{\partial L}{\partial y_m} \sum_{l=1}^{3} \frac{\partial y_m}{\partial o_l} \sum_{n=1}^{4} \frac{\partial o_l}{\partial z_n} \cdot \frac{\partial z_n}{\partial w_{ki}} \quad — (1)$$

$$\text{Let } z_n = \tanh\left(\sum_{t=1}^{4} w_{nt} \cdot x_t \right)$$

$$\frac{\partial z_n}{\partial w_{pq}} = \left[1 - \tanh^2\left(\sum_{t=1}^{4} w_{nt} \cdot x_t \right) \right] x_t \cdot I(t==q)$$

$$= (1 - z_n^2) \cdot x_t \cdot I(t==q) \quad —(2)$$

$$\text{Let } O_l = \sum_{k'=1}^{4} V_{lk'} z_{k'}$$

$$\frac{\partial O_l}{\partial z_{q'}} = V_{lk'} I(k'==q)$$

$$\frac{\partial L}{\partial w_{ki}} = -\left\{ \sum_{m=1}^{3} \frac{y_m}{\hat{y}_m} \sum_{l=1}^{3} \hat{y}_m (1-\hat{y}_m) I(l==m) \right.$$

$$+ (-\hat{y}_m \cdot \hat{y}_l) \cdot I(l \neq m) \sum_{n=1}^{4} V_{lk'} I(k'==n) \cdot$$

$$\left. (1 - z_n)^2 \cdot x_t \cdot I(t==i) \right\}$$

where: $O_l = \sum_{k'=1}^{4} V_{lk'} \cdot z_{k'}$

Notes: $\frac{\partial (\tanh(x))}{\partial x} = \frac{\partial}{\partial x}\left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)$

$$= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

$$= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2}$$

$$= 1 - \tanh^2 x$$