Venkata Pradeep Kadubandi
3011069060

Homework 2

CS 567
July 10, 2019

1 Problem

1.1 Solution

By the definition of $\mathbf{w}_{opt}$, we know that all points are correctly classified when $\mathbf{w} = \mathbf{w}_{opt}$. Therefore, for any point $(\mathbf{x}_i, y_i)$:

$$y_i = sign(\mathbf{w}_{opt}^T \mathbf{x}_i) = sign(\mathbf{x}_i^T \mathbf{w}_{opt})$$

i.e.,

$$y_i \mathbf{x}_i^T \mathbf{w}_{opt} = \mathbf{x}_i^T \mathbf{w}_{opt} sign(\mathbf{x}_i^T \mathbf{w}_{opt}) = \left| \mathbf{x}_i^T \mathbf{w}_{opt} \right|$$

Also, by definition of γ, we have

$$\gamma = min_{i \in [N]} \frac{\left| \mathbf{w}_{opt}^T \mathbf{x}_i \right|}{\|\mathbf{w}_{opt}\|}$$

i.e., for any $(\mathbf{x}_i, y_i)$, we have:

$$\gamma \|\mathbf{w}_{opt}\| \leq \left| \mathbf{w}_{opt}^T \mathbf{x}_i \right| = \left| \mathbf{x}_i^T \mathbf{w}_{opt} \right|$$

From the above two, for any point $(\mathbf{x}_i, y_i)$:

$$y_i \mathbf{x}_i^T \mathbf{w}_{opt} \geq \gamma \|\mathbf{w}_{opt}\| \tag{1.1}$$

When perceptron algorithm makes a mistake and updates $\mathbf{w}$, we will have for some $(\mathbf{x}_i, y_i)$:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Applying transform and multiplying both sides by $\mathbf{w}_{opt}$, we get:

$$\begin{aligned} \mathbf{w}_{k+1}^T \mathbf{w}_{opt} &= \mathbf{w}_k^T \mathbf{w}_{opt} + y_i \mathbf{x}_i^T \mathbf{w}_{opt} \\ &\geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \qquad \text{...from 1.1} \end{aligned}$$

1.2 Solution

The below is the proof of inequality:

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 &= \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} = (\mathbf{w}_k + y_i \mathbf{x}_i)^T (\mathbf{w}_k + y_i \mathbf{x}_i) \qquad \text{...from perceptron update} \\ &= \mathbf{w}_k^T \mathbf{w}_k + \mathbf{w}_k^T y_i \mathbf{x}_i + y_i \mathbf{x}_i^T \mathbf{w}_k + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + 1 \qquad \text{...norm definition, } \mathbf{w}_k^T \mathbf{x}_i = \mathbf{x}_i^T \mathbf{w}_k,\ y_i \in (-1, 1),\ \|x_i\| = 1 \\ &\leq \|\mathbf{w}_k\|^2 + 1 \qquad \text{...given } y_i \mathbf{w}_k^T \mathbf{x}_i \leq 0 \end{aligned}$$

1.3 Solution

Using the inequality from sub section 1.1 over all the M steps where perceptron update happens, we have:

$$\begin{aligned}
\mathbf{w}_{k+1}^T \mathbf{w}_{opt} &\geq \mathbf{w}_l^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \\
\mathbf{w}_l^T \mathbf{w}_{opt} &\geq \mathbf{w}_m^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \\
&\ldots \\
\mathbf{w}_1^T \mathbf{w}_{opt} &\geq \mathbf{w}_0^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \qquad \ldots M \text{ equations}
\end{aligned}$$

where subscripts l,m indicate the value of $\mathbf{w}$ at the time of making a mistake.
By adding all the above equations and using telescopic sum, we will get:

$$\begin{aligned}
\mathbf{w}_{k+1}^T \mathbf{w}_{opt} &\geq \gamma M \|\mathbf{w}_{opt}\| + \mathbf{w}_0^T \mathbf{w}_{opt} \\
&= \gamma M \|\mathbf{w}_{opt}\| \qquad \ldots \text{given } \mathbf{w}_0 = 0
\end{aligned}$$

Using Cauchy-Schwartz inequality and combining with the above, we get:

$$\gamma M \|\mathbf{w}_{opt}\| \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \|\mathbf{w}_{opt}\|$$

From this we have the first half of inequality (given $\|\mathbf{w}_{opt}\| > 0$):

$$\gamma M \leq \|\mathbf{w}_{k+1}\| \tag{1.2}$$

In a similar fashion, we can write the inequality from sub section 1.2, M times and expand it inline to get:

$$\begin{aligned}
\|\mathbf{w}_{k+1}\|^2 &\leq \|\mathbf{w}_l\|^2 + 1 \\
&\leq \|\mathbf{w}_m\|^2 + 1 + 1 \qquad \ldots M \text{ terms till } \mathbf{w}_0 \\
&\leq \|\mathbf{w}_0\| + M = M \qquad \ldots \text{given } \mathbf{w}_0 = 0
\end{aligned}$$

The second half of inequality is simply obtained by taking square root on each side (considering only positive square root):

$$\|\mathbf{w}_{k+1}\| \leq \sqrt{M} \tag{1.3}$$

Combining 1.2 and 1.3, we have shown that:

$$\gamma M \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M} \tag{1.4}$$

1.4 Solution

This follows directly from result of sub section 1.3, if the algorithm made M mistakes before it converged, from 1.4, we will have (we know that M and γ are positive):

$$\begin{aligned}
\gamma M &\leq \sqrt{M} \qquad \ldots \text{i.e.,} \\
\sqrt{M} &\leq \frac{1}{\gamma} \qquad \ldots \text{i.e.,} \\
M &\leq \gamma^{-2}
\end{aligned}$$

2 Problem

2.1 Solution

The given loss function is:

$$L(\mathbf{w}, b) = -\sum_n \{y_n log(\sigma(z)) + (1 - y_n) log((1 - \sigma(z))\} \qquad \text{...where } z = \mathbf{w}^T\mathbf{x}_n + b$$

For update rule for w, we need to find the derivative of the above w.r.t w.

$$\begin{aligned}
\frac{\partial L}{\partial w_j} &= -\frac{\partial \sum_n \{y_n log(\sigma(z)) + (1 - y_n) log(1 - \sigma(z))\}}{\partial w_j} \\
&= -\sum_n \left\{ \frac{y_n}{\sigma(z)} \sigma(z)(1 - \sigma(z)) - \frac{(1 - y_n)}{(1 - \sigma(z))} \sigma(z)(1 - \sigma(z)) \right\} \frac{\partial z}{\partial w_j} \qquad \text{...derivative for log and sigmoid} \\
&= -\sum_n \{y_n(1 - \sigma(z)) - (1 - y_n)\sigma(z)\} x_{nj} \qquad \text{...simplification and } \frac{\partial z}{\partial w_j} = x_{nj} \\
&= \sum_n (\sigma(z) - y_n) x_{nj} \qquad \text{...more simplification}
\end{aligned}$$

So the update rule for w using Gradient Descent method is given by:

$$w_j^{t+1} = w_j^t - \lambda \sum_n x_{nj}[\sigma(\mathbf{w}^T\mathbf{x}_n + b) - y_n]$$

This can be written in the vector form like:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \lambda \sum_n \mathbf{x}_n[\sigma(\mathbf{w}^T\mathbf{x}_n + b) - y_n]$$

Interestingly, this update is quite similar to the one derived for linear regression in class for residual sum of squares loss function.

2.2 Solution

The model is $p(y = 1|x) = \sigma(wx)$ i.e., $p(y = 1|x) = \frac{1}{1+e^{-wx}}$. The learning rate $\lambda = 0.001$.

Initially we have $w_0 = 0$. We have 4 training samples: $(1, 0), (1, 1), (1, 1), (1, 1)$.

Since $w_0 = 0$, for any x, our initial prediction will be $\frac{1}{1+e^0} = \frac{1}{2}$.

Using the above training samples and update rule for w in sub section 2.1, we get:

$$\begin{aligned}
w_1 &= w_0 - (0.001)\sum_{n=1}^{4} x_n(\sigma(w_0 x_n) - y_n) \\
&= 0 - (0.001)[1.(\frac{1}{2} - 0) + 1.(\frac{1}{2} - 1) + 1.(\frac{1}{2} - 1) + 1.(\frac{1}{2} - 1)] \\
&= 0.001
\end{aligned}$$

Since all our training samples have $x = 1$, with the new w, our prediction for $p(y = 1)$ for all training samples will be $\frac{1}{1+e^{-0.001}} = 0.500249$. Since this is greater than 0.5, we will predict the y value as 1 for all training samples. Since the true/expected label for y is 0 in only one sample, we are accurately predicting 3 out of 4 samples i.e., with 75% training accuracy.

2.3 Solution

When $x = -1$, with $w = 0.001$, our prediction for $p(y = 1)$ will be $\frac{1}{1+e^{0.001}} = 0.49975$ i.e., we will predict the label as 0 for the first test sample $(x_1, y_1) = (-1, 0)$ which is accurate.

The next two test samples both have $x = 1$, so following the same logic in the above section, we will end up predicting the label for y as 1. From the expected labels in the test set, 1 test point is correctly classified $\{(x_2, y_2)\}$ and the other one is not $\{(x_3, y_3)\}$.

So, out of 3 test samples, we will correctly classify 2 samples i.e., with 66.6% test accuracy.

3 Problem

3.1 Solution

We will calculate derivatives one step/layer at a time all the way from loss function node back to input nodes to obtain the required derivatives.

$$L(y, \hat{y}) = \sqrt{\frac{1}{2}((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2)}$$

$$\text{for } j \in \{1,2\}, \quad \frac{\partial L}{\partial \hat{y}_j} = \frac{1}{2L}.\frac{1}{2}.2(\hat{y}_j - y_j) \qquad ...\frac{\partial \sqrt{x}}{\partial x} = \frac{1}{2\sqrt{x}}, \text{ chain rule}$$

$$= \frac{1}{2L}(\hat{y}_j - y_j) \qquad ...\text{where } L = \sqrt{\frac{1}{2}((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2)}$$

$$\frac{\partial L}{\partial v_{jk}} = \frac{\partial L}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial v_{jk}} = \frac{1}{2L}(\hat{y}_j - y_j)z_k \qquad ...\text{each } v_{jk} \text{ is multiplied by } z_k \text{ and appears only in } \hat{y}_j$$

$$\frac{\partial L}{\partial z_k} = \sum_{j=1}^{2} \frac{\partial L}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial z_k} \qquad ...\text{each } z_k \text{ appears both in calculating } \hat{y}_1 \text{ and } \hat{y}_2\text{, hence the sum}$$

$$= \frac{1}{2L}\sum_{j=1}^{2}(\hat{y}_j - y_j)v_{jk} \qquad ...\text{took the constant term outside the summation}$$

$$\text{Let } a_k = \sum_{i=1}^{3} w_{ki}x_i \text{ , so that } z_k = arctan(a_k)$$

$$\frac{\partial L}{\partial a_k} = \frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial a_k} = \frac{1}{2L}\sum_{j=1}^{2}(\hat{y}_j - y_j)v_{jk}\left\{\frac{1}{a_k^2 + 1}\right\} \qquad ...\frac{\partial arctan(\alpha)}{\partial \alpha} = \frac{1}{\alpha^2 + 1}$$

$$= \frac{1}{2L(a_k^2 + 1)}\sum_{j=1}^{2}(\hat{y}_j - y_j)v_{jk} \qquad ...\text{taking the constant outside of summation}$$

$$\frac{\partial L}{\partial w_{ki}} = \frac{\partial L}{\partial a_k}\frac{\partial a_k}{\partial w_{ki}} = \left\{\frac{1}{2L(a_k^2 + 1)}\sum_{j=1}^{2}(\hat{y}_j - y_j)v_{jk}\right\}x_i \qquad ...\text{each } w_{ki} \text{ is multiplied by } x_i \text{ in getting } a_k$$

$$= \frac{x_i}{2L(a_k^2 + 1)}\sum_{j=1}^{2}(\hat{y}_j - y_j)v_{jk}$$

Finally writing down the terms that were asked:

$$\frac{\partial L}{\partial v_{jk}} = \frac{1}{2L}(\hat{y}_j - y_j)z_k \qquad \text{...where } L = \sqrt{\frac{1}{2}((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2)}$$

$$\frac{\partial L}{\partial w_{ki}} = \frac{x_i}{2L(a_k^2 + 1)} \sum_{j=1}^{2} (\hat{y}_j - y_j) v_{jk} \qquad \text{...where } a_k = \sum_{i=1}^{3} w_{ki} x_i$$