CS221 Fall 2018 Homework 2

SUNet ID: 05794739
Name: Luis Perez
Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

(a) We first need to know the $\nabla_{\boldsymbol{w}} \text{Loss}_{\text{hinge}}(x, y, \boldsymbol{w})$. We have:

$$\nabla_{\boldsymbol{w}} \text{Loss}_{\text{hinge}}(x, y, \boldsymbol{w}) = 0 \qquad (1 - \boldsymbol{w} \cdot \phi(x) y \leq 0)$$
$$\nabla_{\boldsymbol{w}} \text{Loss}_{\text{hinge}}(x, y, \boldsymbol{w}) = -\phi(x) y$$

We see then that $\boldsymbol{w}$ changes only when $\boldsymbol{w} \cdot \phi(x) y < 1$ such that the count for the word in that review either goes up by η (for a positive review) or down by η (for a negative review) – this is due to use subtracting the gradient, which has a scaling factor of $-\eta y$ on the feature vector.

We start with $w = [0, 0, 0, 0, 0, 0]$ where the features are {pretty, bad, good, plot, not, scenery }. On the first update, we note that $\boldsymbol{w} \cdot \phi(x) y = 0 < 1$, so we now have $w = [-0.5, -0.5, 0, 0, 0, 0]$.

For the second review, we now have $\boldsymbol{w} \cdot \phi(x) y = 0 < 1$, so we update to have $w = [-0.5, -0.5, 0.5, 0.5, 0, 0]$.

On the third review, we have $\boldsymbol{w} \cdot \phi(x) y = -0.5 < 1$, so we update to have $w = [-0.5, -0.5, 0, 0.5, -0.5, 0]$.

On the fourth review, we have $\boldsymbol{w} \cdot \phi(x) y = -0.5 < 1$, so we update to have $w = [0.0, -0.5, 0.0, 0.5, -0.5, 0.5]$.

(b) The data set we can use is as follows:

- (-1) not, $\phi(x_1) = [1, 0, 0]^T$
- (+1) good, $\phi(x_2) = [0, 1, 0]^T$
- (-1) bad, $\phi(x_2) = [0, 0, 1]^T$
- (+1) not bad, $\phi(x_3) = [1, 0, 1]^T$

The consider any linear classifier with weight vector $\boldsymbol{w} = [w_1, w_2, w_3]$ where we have $\hat{y}(x) = \boldsymbol{w} \cdot \phi(x)$ (we predict +1 if non-negative and -1 if negative). Suppose a linear classifier exists which can correctly classify the data above. Then we must have the

below be true, assuming correct classification:

$$\begin{aligned}
\boldsymbol{w}\phi(x_1) &= w_1 < 0 \\
\boldsymbol{w}\phi(x_2) &= w_2 \geq 0 \\
\boldsymbol{w}\phi(x_3) &= w_3 < 0 \\
\boldsymbol{w}\phi(x_4) &= w_1 + w_3 \geq 0
\end{aligned}$$

This is a contradiction, as we can't have $w_1 < 0, w_3 < 0$ and $w_1 + w_3 \geq 0$ (two negative values can't add to a non-negative value). Therefore, the above dataset is not linearly-seperable and a linear classifier cannot possible achieve zero loss.

To fix the problem, we could add an additional feature which is 0 for reviews with one word and 1 for reviews with two words.

- (-1) not, $\phi(x_1) = [1, 0, 0, 0]^T$
- (+1) good, $\phi(x_2) = [0, 1, 0, 0]^T$
- (-1) bad, $\phi(x_2) = [0, 0, 1, 0]^T$
- (+1) not bad, $\phi(x_3) = [1, 0, 1, 1]^T$

Then note that the weight vector $w = [-1, 1, -1, 3]$ on a linear classifier $w\phi(x)$ will now correctly classify the items in the data set.

$$\begin{aligned}
\boldsymbol{w}\phi(x_1) &= -1 \\
\boldsymbol{w}\phi(x_2) &= +1 \\
\boldsymbol{w}\phi(x_3) &= -1 \\
\boldsymbol{w}\phi(x_4) &= +1
\end{aligned}$$

Problem 2

(a) As described, we have the following loss:

$$\begin{aligned}
\text{Loss}(x, y, \boldsymbol{w}) &= (\sigma(\boldsymbol{w} \cdot \phi(x)) - y)^2 \\
&= \left(y - \frac{1}{1 + e^{-\boldsymbol{w}\dot{\phi}(x)}}\right)^2
\end{aligned}$$

(b) We can take the gradient directly, letting $p = \sigma(\boldsymbol{w} \cdot \phi(x))$

$$\begin{aligned}
\nabla_{\boldsymbol{w}}\text{Loss}(x, y, \boldsymbol{w}) &= -2(y - p)\nabla_w \sigma(\boldsymbol{w} \cdot \phi(x)) && \text{(chain rule)} \\
&= -2(y - p)p(1 - p)\phi(x) && \text{(derivative of sigmoid as detailed here)}
\end{aligned}$$

(c) Suppose we have some arbitrary $\phi(x)$ and $y = 1$. Then we can simplify the gradient expression slightly.

$$\begin{aligned}\nabla_{\boldsymbol{w}}\text{Loss}(x, 1, \boldsymbol{w}) &= -2(1-p)^2 p\phi(x) \\ &= -2[p - 2p^2 + p^3]\phi(x)\end{aligned}$$

We can make the above arbitrarily small by taking $||\boldsymbol{w}|| \to \infty$ with the additionally restriction that $\boldsymbol{w} \cdot \phi(x) \neq 0$. To see why, let us see how p is affected as $||\boldsymbol{w}||$ changes.

$$\begin{aligned}\lim_{||w||\to\infty} p &= \lim_{||w||\to\infty} \frac{1}{1 - e^{-\boldsymbol{w}\cdot\phi(x)}} \\ &= \lim_{||w||\to\infty} \frac{1}{1 - e^{-||w||\boldsymbol{u}\cdot\phi(x)}} \qquad (\boldsymbol{u} = \tfrac{\boldsymbol{w}}{||w||})\end{aligned}$$

At this point, we have two options. The first, is $\boldsymbol{u} \cdot \phi(x) > 0$ or $\boldsymbol{u} \cdot \phi(x) < 0$ (the $= 0$ case is not possible by our constraints). In the first case, we'll have:

$$\lim_{||w||\to\infty} p = 1$$

while in the second case, we have:

$$\lim_{||w||\to\infty} p = 0$$

In either scenario, we have:

$$\lim_{||w||\to\infty} \nabla_{\boldsymbol{w}}\text{Loss}(x, y, -c\frac{\phi(x)}{||\phi(x)||_2^2}) = 0$$

From the above, we see that we can make the gradient be as small as we'd like. The intuition is that we can make the gradient arbitrarily small as long as we can make $||\boldsymbol{w}||$ arbitrarily large.

However, we note that the magnitutde of the gradient will never be exactly zero.

(d) In terms of making the gradient be large, this is achieved when $p-2p^2+p^3$ is maximized in the interval $[0, 1]$. We note that the derivative is $1-4p+3p^2 = (3p-1)(p-1)$ which has roots at $p = 1$ and $p = \frac{1}{3}$. From the results above, we know that $p = 1$ is a local minimum. We note that the function is convex on $[0, 1]$, and as such, $p = \frac{1}{3}$ is a local maximum.

Therefore, maximum magnitude that the gradient can take occurs at $p = \frac{1}{3}$ and is given by:

$$2\left(\frac{4}{27}\right)||\phi(x)||_2 = \frac{8}{27}||\phi(x)||_2$$

(e) In order to generate our new dataset, we simply transform $y \to y'$ as follows:

$$y' = \ln\left(\frac{y}{1-y}\right)$$

We claim that there $\exists \boldsymbol{w}^*$ such that $\boldsymbol{D'}$ has zero loss when the prediction is given by $\hat{y'} = \boldsymbol{w}^* \cdot \phi(x)$ (a linear predictor). To see why, first let use solve for y given our above defintion of y':

$$\begin{aligned}
y' &= \ln\left(\frac{y}{1-y}\right) \\
\implies e^{y'} &= \frac{y}{1-y} \\
\implies e^{y'} &= \frac{1}{\frac{1}{y} - 1} \\
\implies \frac{1}{y} &= e^{-y'} + 1 \\
\implies y &= \frac{1}{1+e^{-y'}}
\end{aligned}$$

With the above in hand, let is now see why we can achieve zero loss on our new dataset with standard linear regression. We note that:

$$\begin{aligned}
&\left(y - \frac{1}{1+e^{-\boldsymbol{w}\cdot\phi(x)}}\right)^2 = 0 \qquad \text{(given in the problem that such a } \boldsymbol{w} \text{ exists)} \\
\implies &\frac{1}{1+e^{-y'}} - \frac{1}{1+e^{-\boldsymbol{w}\cdot\phi(x)}} = 0 \\
&\qquad\qquad \text{(solving for } y \text{ given our definition of } y' \text{ and substituting)} \\
&\implies y' = \boldsymbol{w}\cdot\phi(x) \\
&\implies (y' - \boldsymbol{w}\cdot\phi(x))^2 = 0
\end{aligned}$$

We therefore have $\boldsymbol{w}^* = \boldsymbol{w}$ which converges to zero loss on $\boldsymbol{D'}$.