CS221 Fall 2015 Homework Sentiment]

SUNet ID: prabhjot
Name: Prabhjot Singh Rai

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) Mapping reviews into feature vectors as follows,

$$\begin{aligned}
\phi_{x1} &= \{pretty: 1, bad: 1\}, y_1 = -1 \\
\phi_{x2} &= \{good: 1, plot: 1\}, y_2 = +1 \\
\phi_{x3} &= \{not: 1, good: 1\}, y_3 = -1 \\
\phi_{x4} &= \{pretty: 1, scenery: 1\}, y_4 = +1
\end{aligned}$$

Recalling from the graph, gradient of hinge loss, for margin less than one, will be $-\phi_{(x)}y$ and 0 for margin greater than one.

$$\nabla_w Loss_{hinge}(x, y, w) = \begin{cases} -\phi_{(x)}y & \text{when}(w.\phi)y < 1 \\ 0 & \text{when}(w.\phi)y > 1 \end{cases}$$

Stochastic gradient descent is defined as

$$w \leftarrow w - \eta \nabla_w Loss_{hinge}(x, y, w)$$

Initialising $\mathbf{w} = [0, \ldots 0]$, or $\mathbf{w} = \{pretty: 0, bad: 0 \ldots scenery: 0\}$, and iterating over each feature vector to update w
First iteration, $w.\phi_{x1}y = 0, \nabla Loss = -\phi_{(x)}y = \{pretty: 1, bad: 1\}$

$$\begin{aligned}
w &= w - \eta(\{pretty: 1, bad: 1\}) \\
w &= \{pretty: 0, bad: 0 \ldots scenery: 0\} - \{pretty: 0.5, bad: 0.5\} \\
w &= \{pretty: -0.5, bad: -0.5\}
\end{aligned}$$

Second iteration, $w.\phi_{x2}y = 0, \nabla Loss = -\phi_{(x)}y = \{good: -1, plot: -1\}$

$$\begin{aligned}
w &= w - \eta\{good: -1, plot: -1\} \\
&= \{pretty: -0.5, bad: -0.5\} - 0.5\{good: -1, plot: -1\} \\
&= \{pretty: -0.5, bad: -0.5, good: 0.5, plot: 0.5\}
\end{aligned}$$

Third iteration, $w.\phi_{x3}y = -0.5, \nabla Loss = -\phi_{(x)}y = \{not: 1, good: 1\}$

$$\begin{aligned}
w &= w - 0.5\{not: 1, good: 1\} \\
&= \{not: -0.5, bad: -0.5, plot: 0.5, pretty: -0.5\}
\end{aligned}$$

Fourth iteration, $w.\phi_{x4}y = -0.5, \nabla Loss = -\phi_{(x)}y = \{pretty : -1, scenery : -1\}$

$$\begin{aligned} w &= w - \{pretty : -0.5, scenery : -0.5\} \\ &= \{scenery : 0.5, plot : 0.5, bad : -0.5, not : -0.5\} \end{aligned}$$

Therefore, weights of the six words are $\{pretty : 0, good : 0, bad : -0.5, plot : 0.5, not : -0.5, scenery : 0.5\}$

(b) Labelled dataset:
1. "not good" (-1)
2. "good" (+1)
3. "bad" (-1)
4. "not bad" (+1)

Let $\{not : x, good : y, bad : z\}$ be the weights assigned to each feature(our feature extractor considering only single words as per the question), where x, y and z can be both positive and negative. In order to get a total of zero error on all data points, we need to get zero error on each data point. For "good" review, score should be positive, therefore, $y > 0$. For bad, score should be negative, therefore, $z < 0$. For "not good", $x + y$ should be < 0, therefore, since $y > 0$, $x < 0$ and $x < -y$. For "not bad", $x + z$ should be > 0, but through former calculations, $x + z$ will be < 0. Therefore, no linear classifier using word features can get zero error on this dataset.

In order to get a zero error, we can append our feature vector with bi-grams(taking two continuous words into account). This additional feature would add weights to "not bad" in the first data and "not bad" in the last. In such a scenario, weights of each feature would be "not good" < 0, "good" > 0, "bad" < 0 and "not bad" > 0.

Problem 2

(a)

$$\begin{aligned} f_w(x) &= \sigma(w.\phi_x) \\ &= (1 + e^{-w.\phi_x})^{-1} \\ Loss_{squared}(x, y, w) &= (f_w(x) - y)^2 \\ &= ((1 + e^{-w.\phi_x})^{-1} - y)^2 \end{aligned}$$

(b) Let $p = \sigma(w.\phi_x) = (1 + e^{-w.\phi_{(x)}})^{-1}$

$$\begin{aligned}
\nabla_w Loss &= \frac{d(p-y)^2}{dw} \\
&= \frac{d(p-y)^2}{dp}\frac{dp}{dw} \\
&= 2(p-y)\frac{dp}{dw} && \textbf{...(1)} \\
\frac{dp}{dw} &= \frac{d(1+e^{-w.\phi_{(x)}})^{-1}}{dw} \\
&= (-1)(1+e^{-w.\phi_{(x)}})^{-2}e^{-w.\phi(x)}(-\phi_{(x)}) \\
&= (p)^2\frac{(1-p)}{p}\phi_{(x)} && \text{(since } p = (1+e^{-w.\phi_x})^{-1}\text{, therefore } e^{-w.\phi_{(x)}} = \frac{1-p}{p}\text{)} \\
&= p(1-p)\phi_{(x)}
\end{aligned}$$

Therefore, substituting the value of $\frac{dp}{dw}$ in (1), we get gradient of the loss is

$$\nabla_w Loss = 2(p-y)p(1-p)\phi_{(x)}$$

(c) Substituting $y = 1$ and arbitary $\phi_{(x)}$ in the above equation,

$$\begin{aligned}
\nabla_w Loss &= 2(p-1)p(1-p)\phi_{(x)} \\
&= -2(p-1)^2 p\phi_{(x)}
\end{aligned}$$

In order to make the magnitude of the gradient of the loss arbitrarily small, we need to make the above equation close to zero. That can happen when p approaches 1 and p approaches zero.
When p approaches 1:

$$\begin{aligned}
p &= 1 \\
(1+e^{-w.\phi_{(x)}})^{-1} &= 1 \\
e^{-w.\phi_{(x)}} &= 0
\end{aligned}$$

w approaches ∞.
When p approaches 0:

$$\begin{aligned}
p &= 0 \\
(1+e^{-w.\phi_{(x)}})^{-1} &= 0 \\
e^{-w.\phi_{(x)}} &\text{ approaches } \infty
\end{aligned}$$

w approaches $-\infty$. Therefore, for w approaching ∞ and $-\infty$, magnitude of the gradient of the loss is arbitrarily small(approaching zero).
No, the magnitude of the gradient can never be zero.