CS221 Fall 2018 - 2019 Homework 2

Name: Dat Nguyen

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

Problem 1: Building intuition

(a) The table below show the weights of the six words ("pretty", "good", "bad", "plot", "not", "scenery") after each iteration.

weight/iter	1	2	3	4
pretty	-0.5	-0.5	-0.5	0
good	0	0.5	0	0
bad	-0.5	-0.5	-0.5	-0.5
plot	0	0.5	0.5	0.5
not	0	0	-0.5	-0.5
scenery	0	0	0	0.5

(b) Proposed data set

"good"	1
"bad"	-1
"not good"	-1
"not bad"	1

From first row $w_{\text{good}} > 0$ and from second row $w_{\text{bad}} < 0$. But from third row $w_{\text{not}} + w_{\text{good}} < 0$ and from forth row $w_{\text{not}} + w_{\text{bad}} > 0$, which is a contradiction.
If we add a feature which is the count the phrase "not good". Then we can make the margin of "not good" as negative as we like while not affecting the margin of "not bad".

Problem 2: Predicting Movie Ratings

(a)

$$\text{Loss}(x, y, \mathbf{w}) = \big(\sigma(\mathbf{w} \cdot \phi(x)) - y\big)^2$$

(b)

$$\begin{aligned}
\frac{\partial \text{Loss}(x, y, \mathbf{w})}{\partial \mathbf{w}} &= 2\big(\sigma(\mathbf{w} \cdot \phi(x)) - y\big)\frac{\partial \sigma(\mathbf{w} \cdot \phi(x)) - y}{\partial \mathbf{w}} \\
&= 2\big(\sigma(\mathbf{w} \cdot \phi(x)) - y\big)\sigma(\mathbf{w} \cdot \phi(x))(1 - \sigma(\mathbf{w} \cdot \phi(x)))\phi(x) \\
&= 2(p - y)p(1 - p)\phi(x)
\end{aligned}$$

(c) Because when $p \to y$ then $p - y \to 0$ and $p(1-p) \to y(1-y) = 0$, we can make the magnitude of the derivative as close to 0 as we want by adjusting w nearer to y. However the magnitude cannot ever be exactly 0 because the sigmoid function cannot achieve the value of 0.

(d) We have

$$\begin{aligned}\Big|\Big|\frac{\partial \text{Loss}(x, y, \mathbf{w})}{\partial \mathbf{w}}\Big|\Big| &= 2||(p-y)p(1-p)||||\phi(x)|| \\ &= 2||(p-1)^2 p||\phi(x)||\end{aligned}$$

Take derivative of $(p-1)^2 p$

$$\begin{aligned}\frac{\partial (p-1)^2 p}{\partial p} &= 2(p-1)p + (p-1)^2 \\ &= (p-1)(3p-1) \\ \frac{\partial (p-1)^2 p}{\partial p} &= 0 \\ \leftrightarrow p = \frac{1}{3} \text{or} 1\end{aligned}$$

So we conclude that $||(p-1)^2 p|| \leq ||(\frac{1}{3}-1)^2 * \frac{1}{3}|| = \frac{4}{27}$ and therefore

$$\Big|\Big|\frac{\partial \text{Loss}(x, y, \mathbf{w})}{\partial \mathbf{w}}\Big|\Big| \leq \frac{8}{27}||\phi(x)||$$

(e) Considering a particular pair (x, y) in $\mathbf{D}$. Since $\mathbf{w}$ yields zero loss, we have

$$\begin{aligned}\frac{1}{1 + e^{-\mathbf{w}\cdot\phi(x)}} &= y \\ \leftrightarrow 1 + e^{-\mathbf{w}\cdot\phi(x)} &= \frac{1}{y} \\ \leftrightarrow e^{-\mathbf{w}\cdot\phi(x)} &= \frac{1}{y} - 1 \\ \leftrightarrow \mathbf{w}\cdot\phi(x) &= -\log\big(\frac{1}{y} - 1\big)\end{aligned}$$

Therefore we can make the transformation $y \to -\log\big(\frac{1}{y} - 1\big)$ to create the dataset $\mathbf{D}'$ of (x, y'), so that in $\mathbf{D}'$ the old weight vector $\mathbf{w}$ still yields zero loss.

Problem 3: Sentiment Classification

(d)

Wrong sentence	Explanation
wickedly funny , visually engrossing , never boring , this movie challenges us to think about the ways we consume pop culture ."	the weight for "boring" is very negative. The classifier needs to recognize that before "boring" is the negated word "never"
the best thing i can say about this film is that i can't wait to see what the director does next .	the weight for "can't" is very negative. The classifier does not realize that "can't wait" does not necessarily imply negative thing
. . . standard guns versus martial arts cliche with little new added .	weight for "new" is very positive that the negative weight of "little" can hardly compensate.
as ex-marine walter , who may or may not have shot kennedy , actor raymond j . barry is perfectly creepy and believable .	weight for "or" is relatively negative while it does not convey useful meaning. It may be better to remove conjunction words like "or", "and"...
o timo esforo do diretor acaba sendo frustrado pelo roteiro , que , depois de levar um bom tempo para colocar a trama em andamento , perde-se de vez a partir do instante em que os estranhos acontecimentos so explicados .	not english comment so there is little information to help with the prediction.

(f) I conduct 5 experiments corresponding to n = 2, 3, 4, 5, 6. Among those n = 5 produces the smallest test error of 0.27. My explanation is that the majority of words can be identifiable by 5 characters so the set of 5-grams words contain the correspondents of the words in the set of all words and more.
One sentence where n-grams might be better than word feature is "not bad". In word feature both weight of "not" and "bad" are probably negative so the overall score is negative while it is a favorable review. In 6-grams it may recognize that "notbad" is actually a good word so the weight may be positive corresponding to label 1.

Problem 4: K-means clustering

(a)
Center assignments of points in each iteration in the first case

iteration/point	1	2	3	4
1	1	0	1	0
2	1	0	1	0

Center assignments of points in each iteration in the second case

iteration/point	1	2	3	4
1	0	0	1	1
2	0	0	1	1

(c) In the center assignment step, for every group of points which should belong to the same cluster, we consider the sum of distance of all points in that group to each center and assign the center with minimal distance to all points of that group. (d) If we just run once, we may get into local optimal since K-means alternates optimizing variable while keeping the other constant. By running K-means multiple times we can increase the chance of getting global optimal if we choose the setting with lowest construction loss.

(e) If we scale all dimensions in our initial centroids and data points, then every distance between point and centroid is also scaled by some constant factor α and so the sum of distance or loss is also scaled by α. Suppose we find the set S of optimal centroids with optimal loss d in the scaled dataset. If we scale back those centroids to get S' with loss d' then we will show that d' must be optimal in the initial dataset. We suppose the contrary that there exists $d'' < d'$ in the initial dataset then if we scale the centroids of d''we get $\alpha d'' < \alpha d' = d$ in the scaled dataset which contradicts with the fact that d is optimal in the scaled dataset. Therefore the clusters of before and after scaling are the same.
If we just scale only certain dimension then that is no longer true.