CS221 Spring 2019 Homework 2

SUNet ID: yijiez
Name: Yijie Zhuang
Last updated: April 16, 2019

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) Given:

$$Loss(x, y, w) = max\{1 - w \cdot \phi(x)y, 0\} \tag{1}$$

and

$$margin = w\phi(x)y \tag{2}$$

$$\nabla_w Loss_{hinge}(x, y, w) = \begin{cases} -\phi(x)y & margin < 1 \\ 0 & margin \geq 1 \end{cases} \tag{3}$$

$w = [0, 0, 0, 0, 0, 0]$
1. $\phi(x) = [1, 0, 1, 0, 0, 0]; y = -1; \phi(x)y = [-1, 0, -1, 0, 0, 0]; w = [-0.5, 0, -0.5, 0, 0, 0]$
2. $\phi(x) = [0, 1, 0, 1, 0, 0]; y = 1; \phi(x)y = [0, 1, 0, 1, 0, 0]; w = [-0.5, 0.5, -0.5, 0.5, 0, 0]$
3. $\phi(x) = [0, 1, 0, 0, 1, 0]; y = -1; \phi(x)y = [0, -1, 0, 0, -1, 0]; w = [-0.5, 0, -0.5, 0.5, -0.5, 0]$
4. $\phi(x) = [1, 0, 0, 0, 0, 1]; y = 1; \phi(x)y = [1, 0, 0, 0, 0, 1]; w = [0, 0, -0.5, 0.5, -0.5, 0.5]$

After the classifier is trained on the four data points, the weights of the six words are: $[0, 0, -0.5, 0.5, -0.5, 0.5]$.

(b) Mini-reviews:
1. (1) good
2. (-1) bad
3. (-1) not good
4. (1) not bad

If we want to get zero error on the dataset, we'll need to have all "good", "bad", "not" be classified correctly. Assume we already assigned positive weights to "good" and negative weights to "bad". If we assign negative weights to "not", then "not bad" will have negative value; if we assign positive weights to "not", then "not good" will have positive value.

To fix the problem, we can add feature called "not good". The feature set becomes:

$$'good', 'bad', 'not', 'notgood'$$

Assume $w = [a, b, c, d]$, we have:
1. $a = 1$
2. $b = -1$
3. $a + c + d = -1$
4. $b + c = 1$

Thus, $w = [1, -1, 2, -4]$, which means w={'good':1,'bad':-1,'not':2,'not good':-4}

Problem 2: Predicting Movie Ratings

(a)

$$Loss(x, y, \mathbf{w}) = (\sigma(\mathbf{w} \cdot \phi(x)) - y)^2 = (\frac{1}{1 + e^{-\mathbf{w} \cdot \phi(x)}} - y)^2 \tag{4}$$

(b)

$$\nabla_w Loss(x, y, \mathbf{w}) = 2\phi(x)(\frac{1}{1 + e^{-w\phi(x)}} - y) \cdot \frac{e^{-w\phi(x)}}{(1 + e^{-w\phi(x)})^2} = 2\phi(x)(p - y)p(1 - p) \tag{5}$$

$p = \frac{1}{1+e^{-w\phi(x)}}$ and $p \in (0, 1)$

(c) Given y = 1, the gradient of the loss becomes:

$$2\phi(x)(p - 1)p(1 - p) = -2\phi(x)p(1 - p)^2 \tag{6}$$

To make the gradient small, we can make p tend to 1 or 0, by making w tend to ∞ or $-\infty$ multiple of $\phi(x)$, so the minimum magnitude of the gradient tends to be 0, but it cannot be 0.

(d) To make the gradient large, we can maximize $p(1 - p)^2$.
Let $G(p) = \ln p(1 - p)^2$, we can get $G(p) = \ln p + \ln (1 - p)^3 = \ln p + 2 \ln (1 - p)$.
Let $G'(p) = \frac{1}{p} - \frac{2}{1-p} = 0$, we can get $p = \frac{1}{3}$. As $0 < p < 1$, this value will maximize the result.

Given $p = \frac{1}{3}$, the max gradient of the loss is $\frac{8}{27} \parallel \phi(x) \parallel$.

(e) There exists a w to make Loss(x,y,w) = 0:

$$\frac{1}{1 + e^{-\mathbf{w} \cdot \phi(x)}} - y = 0 \tag{7}$$

$$1 + e^{-\mathbf{w} \cdot \phi(x)} = \frac{1}{y} \tag{8}$$

$$e^{-\mathbf{w} \cdot \phi(x)} = \frac{1}{y} - 1 \tag{9}$$

$$\mathbf{w} \cdot \phi(x) = \log \frac{y}{1-y} \tag{10}$$

Thus, by making $y \rightarrow \log \frac{y}{1-y}$, we are able to create D' where w still yield 0 loss.

Problem 3: Sentiment Classification

(a) N/A

(b) N/A

(c) N/A

(d)

Sentence	Wrong Reason
home alone goes hollywood , a funny premise until the kids start pulling off stunts not even steven spielberg would know how to do . besides , real movie producers aren't this nice. (Truth: -1, Prediction: 1)	Positive words like "funny", "real" contributed to the positive results, words performs as turning points like "not even" and "aren't" cannot outweigh them.
a perfectly competent and often imaginative film that lacks what little lilo & stitch had in spades – charisma .(Truth: 1, Prediction: -1)	Words like "lacks" and "little" are treated as negative words wrongly.
a heady , biting , be-bop ride through nighttime manhattan , a loquacious videologue of the modern male and the lengths to which he'll go to weave a protective cocoon around his own ego .(Truth: 1, Prediction: -1)	Neutral words "around" and "male" are treated as negative words.
it's painful to watch witherspoon's talents wasting away inside unnecessary films like legally blonde and sweet home abomination , i mean , alabama . (Truth: -1, Prediction: 1)	Negative word "painful" is treated as positive word, "sweet" has the highest positive weight, but it's just part of a movie name
wickedly funny , visually engrossing , never boring , this movie challenges us to think about the ways we consume pop culture .(Truth: 1, Prediction: -1)	'never boring' are treated as two negative words.

To improve the accuracy of the classifier, we could add N continuous words as a feature instead of just using a single word.

(e) N/A

(f) The test passed when $n > 3$, and got smallest value when $n = 5$. Explanation: By using N gram, we are able to capture more than one word in a feature, words like 'not

good' or 'not bad' will be classified correctly.
An example review: "This movie is not good". In this case, word features may assign a positive result to it if "good" outweights "not". However, n-gram will be able to capture it and give 'notgood' a negative weight.

Problem 4: K-means clustering

(a) μ_1 has (x_1, x_3), μ_2 has (x_2, x_4), then we update $\mu_1 = [2, 0], \mu_2 = [2, 1]$. Final assignments stay the same.
μ_1 has (x_1, x_2), μ_2 has (x_3, x_4), then we update $\mu_1 = [0, \frac{1}{0}], \mu_2 = [1, \frac{2}{3}]$. Final assignments stay the same.

(b) N/A

(c) Given

$$Loss = \sum_{i=1}^{n} \parallel \mu_{z_i} - \phi(x) \parallel^2 \tag{11}$$

When setting centroid, given z is fixed, to minimize loss, we'll have

$$2 \sum_{x \in cluster_\mu} (\mu - \phi(x)) = 0 \tag{12}$$

$$\mu = \frac{\sum_{x \in cluster_\mu} \phi(x)}{\mid x \mid^2} \tag{13}$$

During assignment, when μ is fixed, to minimize loss with z, we can do

$$z_i = argmin \sum_{j(g_i = g_j)} \parallel \phi(x_j) - \mu_z \parallel^2 \tag{14}$$

where g_i means group that i is assigned to. If $(i, j) \in S$ and $(j, k) \in S$, then $g_i = g_j = g_k$.

(d) K-means can only converge to local minimum. By running k-means multiple times on the same dataset with different random initializations, we'll be able to find the global minimum easier.

(e) If we scale all dimensions in our initial centroids and data points by some factor, we are guarenteed to retrieve the same clusters after running k-means because the loss / distance will be scaled by the same factor.
However, if we just scale some dimensions, the clustring results may be different. For example, assume centroids are [0, 1] and [1, 1]. If $x \to 0.0001x$, then centroids become [0, 1] and [0.0001, 0], which are so closed that we cannot treat them as two clusters at all.