Homework 2 Sentiments

1. a. $\nabla_w \text{Loss}_{\text{hinge}} =$

① $y=1$, $\phi(x) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ ~~margin = $\vec{w} \cdot \phi(x) = 0$~~.

margin $= \vec{w} \cdot \phi(x) y = 0 < 1$

$$\nabla_w = -1\, \phi(x) y = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\vec{w} \leftarrow \vec{w} - \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

② $y=-1$, $\phi(x) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$

margin $= \vec{w} \cdot \phi(x) y = 0 \times (-1) = 0 < 1$

$$\nabla_w = \cancel{-1}\,[\text{margin} < 1]\, \phi(x)\, \cancel{y}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \vec{w} = \vec{w} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

③ $y=-1$, $\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$, $\vec{w} \cdot \phi(x) y = 1 \times (-1) = -1$

$$\nabla_w = -1\,[-1 < 1]\,(-1)\, \phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\vec{W} \leftarrow \vec{w} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

④ $y=1$, $\phi(x) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$, $\vec{w} \cdot \phi(x) y = 1(1) = 1 = \text{margin}$

$$\nabla_W = \cancel{-1 \begin{bmatrix} 1<1 \end{bmatrix}}\ 0$$

$$\vec{w} = \vec{W} - 0 = \vec{w} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$ ANSWER.

b.

① good (+1) ② bad (−1) ③ ~~bad s~~ not bad (+1) ④ not good (−1)

~~[0 1 0~~ We need to ensure that the margin is greater than ~~or equal to~~ 0.

$\vec{\omega} \cdot \phi(x) y > 0$. Vector $= \begin{bmatrix} \text{good} \\ \text{bad} \\ \text{not} \end{bmatrix}$

$$\begin{cases} 1 \overset{\times}{\cancel{0}} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \cdot \vec{W} > 0 \\ (-1) \overset{\times}{\cancel{0}} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \cdot \vec{W} > 0 \\ 1 \overset{\times}{\cancel{0}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \cdot \vec{w} > 0 \\ (-1) \overset{\times}{\cancel{0}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \cdot \vec{W} > 0 \end{cases} \Rightarrow \begin{cases} W_1 > 0 \\ -W_2 > 0 \Rightarrow W_2 < 0 \\ W_2 + W_3 > 0 \\ W_1 + W_3 < 0 \end{cases}$$

If $W_1 > 0$, $W_2 < 0$, then if $\omega_1 + W_3 < 0 \Rightarrow W_3 < 0$.

but then it is impossible for $W_2 + W_3 > 0$, since W_2 & W_3 are both negative.

Thus it is impossible to find $\vec{w}$ that satisfies this system of equations.

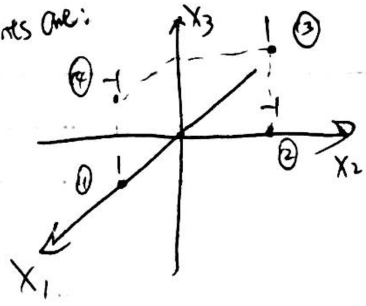
~~Since we have proven that a linear classifier won't work, we must try non-linear.~~

~~From the dataset, we can see that if we want $W_1 > 0$, $W_2 < 0$, and $W_1 + W_3 \leq 0$, then W_3 is negative. It's only possible that $W_2 - W_3 > 0$, but not $W_2 + W_3 > 0$.~~

~~So $\phi(x) = (x_1, x_2, x_3, x_1^2 + x_2^2 + x_3^2 +$~~

~~$\phi(x) = [x_1, x_2, x_3, x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_2x_3]$~~

The points are:



There is no way a hyperplane can seperate these 4 points.

New feature:

$X_4 = 1[X_2 == X_3]$ then the points can be seperated

since $Y = 1$ if $X_2 == X_3$, $Y = ~~0~~ -1$ if $X_2 \neq X_3$

2.

a. $\text{Loss}(x, y, \vec{w}) = \left[(1+e^{-\vec{w}\cdot\phi(x)})^{-1} - y\right]^2$

b. $\nabla_w \text{Loss} = 2(\vec{p} - y) \cancel{\frac{\partial \vec{p}}{\partial \vec{w}}} \nabla_w \vec{p}$

$$\cancel{\frac{\partial \vec{p}}{\partial \vec{w}} =} \nabla_w \vec{P} = -(1+e^{-\vec{w}\cdot\phi(x)})^{-2}(-\phi(x))e^{-\vec{w}\cdot\phi(x)}$$

$$\Rightarrow \nabla_w \text{Loss} = 2\left[(1+e^{-\vec{w}\cdot\phi(x)})^{-1} - y\right]\left[(1+e^{-\vec{w}\cdot\phi(x)})^{-2}\phi(x)\, e^{-\vec{w}\cdot\phi(x)}\right]$$

c. From the result in b, we have $e^{-\vec{w}\cdot\phi(x)}$ in the numerator and $[1+e^{-\vec{w}\cdot\phi(x)}]^{-3}$ in the denominator.

~~When~~ So it basically comes down to ~~maximi~~ minimize $\frac{1}{(e^{-\vec{w}\phi(x)})^2}$. This happens when $\vec{\omega} = -\vec{\infty}$

So that this becomes $\frac{1}{e^{\infty}} = 0$. ~~$\vec{w} = \vec{0}$~~

$||\nabla_{\text{Loss}}|| = 0$ in this case.

~~2d~~ d.

$$\|\nabla_w Loss\|_{y=0} = \| 2(1+e^{-\vec{w}\cdot\phi(x)})^{-3}\phi(x)e^{-\vec{w}\cdot\phi(x)}\|$$

If $\vec{w} = \begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\0\end{bmatrix}$ then it reaches maximum.

$$\|\nabla_w Loss\|_{y=0} = \| 2\cdot(1+1)^{-3}\phi(x)\cdot 1\| = \|\frac{1}{4}\phi(x)\|$$

e. logistic regression $\longleftrightarrow$ linear regression

~~We know logi~~ So y can take the inverse of sigmoid.

$$\Rightarrow y' = \text{sigmoid}^{-1}(y) = \underline{\underline{\log\left(\frac{y}{1-y}\right)}}$$

3 – d:

This document provides response to Question 3d.

Analysis of incorrect predictions:

1. === 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 [WRONG]

Reason for error: This review is clearly sarcastic with quite a few positive words, such as "funny", "real", "nice", etc, which could have confused the predictor.

The classifier needs to use phrases rather than single words as features to "understand" the sarcastic reviews that sometimes use positive words.

2. === 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 [WRONG]

Reason for error: This review uses too many rare words such that they were not learned from the training sample.

To fix this, the classifier needs example reviews containing these words.

3. === '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 [WRONG]

Reason for error: It appears that the movie names contain positive words with heavy weights, such as sweet, films, home, etc.

If we were provided with the names of all movies, we could filter them out.

4. === wickedly funny , visually engrossing , never boring , this movie challenges us to think about the ways we consume pop culture .

Truth: 1, Prediction: -1 [WRONG]

Reason for error: Context wasn't taken into account so that e.g. "never boring" drags the score down a lot whilst it's actually a positive phrase.

If we also take phrases as features it would improve the predictions in this situation.

5. === patchy combination of soap opera , low-tech magic realism and , at times , ploddingly sociological commentary .

Truth: -1, Prediction: 1 [WRONG]

Reason for error: Context was not taken into account, some rare words weren't learned in the first place but they convey a lot of meaning.

We could provide more training samples containing those rare words, and extract more features such as phrases.

6. === only in its final surprising shots does rabbit-proof fence find the authority it's looking for .

Truth: -1, Prediction: 1 [WRONG]

Reason for error: This review doesn't have adjectives describing directly how the movie is. It is using an analog/something in the movie to convey how ridiculous the plot is.

The prediction in the case can be improved by having more training samples.

7. === alternative medicine obviously has its merits . . . but ayurveda does the field no favors .

Truth: -1, Prediction: 1 [WRONG]

Reason: The review uses "but" to make a turn in its attitude.

We should extract phrases containing "but" and the words following it.

8. === no screen fantasy-adventure in recent memory has the showmanship of clones' last 45 minutes .

Truth: 1, Prediction: -1 [WRONG]

Reason: Positive words such as fantasy-adventure and showmenship were not learned.

We need more training samples. Also, we need to split the dash to make two separate words, which are more likely to have been learned.

9. o ótimo esforço 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 são explicados .

Truth: -1, Prediction: 1 [WRONG]

Reason: The vast majority of the training samples are in English, while this review is in Spanish. As a result, most of the words' weights were not learned at all.

We need to have more training samples in Spanish.

10. === this may be the dumbest , sketchiest movie on record about an aspiring writer's coming-of-age .

Truth: -1, Prediction: 1 [WRONG]

Reason: The very last few words are about the content of the movie, but most of them are positive. However, they have no impact on the overall positiveness of the review.

We could include phrases consist of "about" and the words following it.vvv

3 – f:

From my experiment with test 3b-2, I found that the test error is minimized with n = 7.

Since the character n-gram gives much more features than the word features, it can overfit. Also, although the character n-gram is able to capture some of the contextual information such as "not good", which the word feature extractor cannot catch, it also produces a lot of redundant features that do not contribute to the prediction.

Review: The movie does a decent job at esthetics.

Reason: This review uses the spelling "esthetics" which is an uncommon form of the word "aesthetics". The character n-gram is better because it'd still be able to relate this review to the examples in the learning dataset that use "aesthetics".

4.

a.

①

(Euclidean distance)² μ ϕ	$\mu_1 = [-1,0]$	$\mu_2 = (3,2)$
$\phi_A(x_1) = (0,0)$	1	13
$\phi(x_2) = (0,1)$	2	10
$\phi(x_3) = (2,0)$	~~1~~ 9	5
$\phi(x_4) = (2,2)$	5	1.

Group:

$\mu_1 = (-1,0)$	$\mu_2 = (3,2)$
$\phi(x_1)$	$\phi(x_3)$
$\phi(x_2)$	$\phi(x_4)$

$\Rightarrow z_1 = 1, z_2 = 1$
$z_3 = 2, z_4 = 2$

② $\mu_1 = \frac{(0,1)}{2} = (0, \frac{1}{2})$

$\mu_2 = \frac{(4,2)}{2} = (2,1).$

③

	μ_1	μ_2
$\phi(x_1)$	[illegible]	μ_2
$\phi(x_2)$		
$\phi(x_2)$		
$\phi(x_4)$		

③

	μ_1	μ_2
$\phi(x_1)$	$\frac{1}{4}$	5
$\phi(x_2)$	$\frac{1}{4}$	4
$\phi(x_3)$	$\frac{17}{4}$	1
$\phi(x_4)$	$25/4$	1

$z_1 = 1,\ z_2 = 1$

$z_3 = 2,\ z_4 = 2$

⇒ Converges !

2. ~~μ_1~~

①

	$\mu_1 = (1,-1)$	$\mu_2 = (0,2)$
$\phi(x_1)$	2	4
$\phi(x_2)$	5	1
$\phi(x_3)$	2	8
$\phi(x_4)$	2	4

$z_1 = 1,\ z_2 = 2, z_3 = 1\ ,\ z_4 = 1$

② So $\mu_1 = \dfrac{(0,0)+(2,0)+(2,2)}{3} = \dfrac{(4,2)}{3} = \left(\frac{4}{3}, \frac{2}{3}\right)$

$\mu_2 = (0,1)$

(2)

	$\mu_1 = (\frac{4}{3}, \frac{2}{3})$	$\mu_2 = (0,1)$
$\phi_1(x_1)$	2/3	1
$\phi_2(x_2)$	17/9	0
$\phi(x_3)$	8/9	5
$\phi(x_4)$	20/9	5.

$z_1 = 1,\ z_2 = 2,\ z_3 = \overset{1}{\cancel{2}}\ z_4 = \cancel{2}\ 1$

$\Rightarrow$ Convergence achieved !!

C.

We should avoid incorrect grouping in our algorith

We can devise a modified algorithm to fix incorrect grouping if it happens.

Steps:

① Run regular k-means until convergence.

② Check if points that should be in the same cluster(s) are grouped separately.

4 – c:

We can run the regular K-means algorithm until convergence, then fix the incorrect clustering.

1. Run regular k-means algorithm until convergence.

2. If two points are assigned to different clusters while it's pre-specified that they should belong to the same cluster, continue to step 3. Otherwise simply return. We call these two points p1 and p2 in this algorithm description. Suppose after step 1, p1 belongs to centroid A and p2 belongs to centroid B.

3. Find the midpoint between p1 and p2. We call this point midpt.

4. Find if A or B is closer to midpt than the other. We suppose that A is closer to midpt than B in this example.

5. Move A to midpt, while assign some random data points to B to re-initialize it.

6. Run k-means again with these two new centroids until convergence. See if the precondition is met.

7. If not, start over from step 3.