COMS 4771-2 Fall 2018 Homework 3

Zhuangyu Ren(zr2209)

October 29, 2018

Problem 1

@articlescikit-learn, title=Scikit-learn: Machine Learning in Python, author=Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., journal=Journal of Machine Learning Research, volume=12, pages=2825–2830, year=2011

(a) My fourth algorithm is skip-grams, choose two words that has only one word in between, because I think that this way can remove some useless phases like "the", "a" and so on. I came up with this algorithm by searching Google and found some like this.

(b) Training error rate:
1. Unigram representation: 0.101567
2. Term frequency-inverse document frequency: 0.095156
3. Bigram representation: 0.049628
4. Skip-grams: 0.063369

(c) Test error rate:
1. Unigram representation: 0.10546291726279355
2. Term frequency-inverse document frequency: 0.10586276482091203
3. Bigram representation: 0.10005560380105084
4. Skip-grams: 0.08923863487987324

(d) The lists of words with highest weights, described above.
'disappointd', 'exceed', 'gema', 'heavenly', 'hooker', 'incredible', 'perfectionism', 'phenomenal', 'skeptically', 'worrier'
The lists of words with lowest weights, described above.
'downhill', 'flavorlessly', 'hopfenweisse', 'inedible', 'lacker', 'mediocrely', 'poison', 'underwhelming', 'underwhelmingly', 'worst'

(e) The text of two misclassified examples, with associated lists of words, and explanations for why the examples were misclassified.

text1: everytime i go to las vegas i would find a morning to have dim sum at ping pang pong even i live in la a place have tons of dim sum restaurant for my taste preference ping pang pong offers much better taste than any other dim sum place in la you should go early unless you want to wait for at least an hour

highest: 'in', 'everytime', 'early', 'place', 'you', 'to', 'preference', 'dim', 'vegas', 'wait'

value: 16.0, 17.0, 17.0, 24.0, 36.0, 39.0, 23.0, 27.0, 19.0, 32.0

lowest: 'taste', 'sum', 'should', 'want', 'offers', 'least', 'other', 'hour', 'la', 'unless'

value: -54.0, -51.0, -25.0, -28.0, -27.0, -40.0, -30.0, -49.0, -30.0, -39.0

text2: i love dim sum those little small bite sized dishes always get me i enjoyed eating at the cathay house restaurant with my boyfriend and his mother it was a good place to hit up my favorite dish here was the steamed bums and those sesame balls that are like extremely dense but delicious i left feeling like i had a food baby inside of me because i was stuffed beyond belief i guess you can get carried away eatings dim sum because they are served in the small plates that you can lose track of what you have eaten the service was not very well i think one of the ladies that was pushing the cart was having a bad day and they never refilled our waters

highest: 'same', 'of', 'and', 'those', 'love', 'favorite', 'stuffed', 'can', 'you', 'delicious'

value: 31.0, 33.0, 36.0, 34.0, 38.0, 42.0, 39.0, 44.0, 51.0, 72.0

lowest: 'sum', 'our', 'but', 'waters', 'left', 'hit', 'dense', 'dish', 'think', 'the'

value: -34.0, -30.0, -33.0, -49.0, -32.0, -28.0, -25.0, -26.0, -25.0, -24.0

I think it's because that the chosen words are not of much relation with emotions, so the prediction is not that accurate.

Problem 2

(a) Suppose Perceptron does not exit the loop in the t-th iteration. Then there is a labeled example $(x_t, y_t) \in S$ such that

$y_t \langle w_*, x_t \rangle \geqslant 1$

$y_t \langle \widehat{w}_t, x_t \rangle < 1$, as is the definition of margin perceptron

We bound $\langle w_*, w_{t+1} \rangle$ from above and below to deduce a bound on the number of loop iterations. First, we bound $\langle w_*, \widehat{w}_t \rangle$ from below:

$\langle w_*, \widehat{w}_{t+1} \rangle = \langle w_*, \widehat{w}_t \rangle + \eta y_t \langle w_*, x_t \rangle \geqslant \langle w_*, \widehat{w}_t \rangle + \eta$

since $\widehat{w}_1 = 0$, we have $\langle w_*, \widehat{w}_t \rangle \geqslant \eta t$

We now bound $\langle w_*, \widehat{w}_{t+1} \rangle$ from above.

$\langle w_*, \widehat{w}_{t+1} \rangle \leqslant \|w_*\|_2 \|\widehat{w}_{t+1}\|_2$

$\|\widehat{w}_{t+1}\|_2^2 = \|\widehat{w}\|_2^2 + 2\eta y_t \langle \widehat{w}_t, x_t \rangle + \eta^2 y_t^2 \|x_t\|_2^2 \leqslant \|\widehat{w}_t\|_2^2 + \eta^2 L^2 + 2\eta$

since $\widehat{w}_1 = 0$, we have $\|\widehat{w}_{t+1}\|_2^2 \leqslant (\eta^2 L^2 + 2\eta)t$

so $\langle w_*, \widehat{w}_{t+1} \rangle \leqslant \|w_*\|_2^2 \sqrt{\eta^2 L^2 + 2\eta}\sqrt{t}$

Combining the upper and lower bounds
$\eta t \leqslant \langle w_*, \widehat{w}_{t+1} \rangle \leqslant \|w_*\|_2^2 \sqrt{\eta^2 L^2 + 2\eta}\sqrt{t}$
since $\eta = \frac{1}{L^2}$,
$t \leqslant 3 \|w_*\|_2^2 L^2$

(b) From part a we know that this loop halts at t-th step, then $\|\widehat{w}_{t+1}\|_2 = \|\widehat{w}_t\|_2$, the last output is $\|\widehat{w}_t\|_2$
And also from part a, we know that $\|\widehat{w}_t\|_2 \leqslant \frac{3t}{L^2}$
plug in $t \leqslant 3 \|w_*\|_2^2 L^2$
then, $\|\widehat{w}_t\|_2 \leqslant 3 \|w_*\|_2$

(c) As above in part b, $\|\widehat{w}_t\|_2 \leqslant 3 \|w_*\|_2$
The margin is $\frac{1}{\|w\|_2}$, so the margin achieved by $\widehat{w}_t$ is (almost, within a factor of three) as large as the margin achieved by w_*

Problem 3

(a) As α is fixed, then compute the partial derivative of L with respect to w and ξ
$\frac{\partial L(w,\alpha,\xi)}{\partial w} = w^T - \sum_{i=1}^n \alpha_i x_i^T = 0$
$w^T = \sum_{i=1}^n \alpha_i x_i^T$
$w = X^T \alpha$
$\frac{\partial L(w,\alpha,\xi)}{\partial \xi} = C\xi - \sum_{i=1}^n \alpha_i = 0$
$\xi = \frac{\alpha}{C}$

(b)

$$\begin{aligned} g(\alpha) &= \frac{1}{2}(X\alpha)^T(X\alpha) + \frac{\|\alpha\|_2^2}{2C} + \alpha^T(Y - X^T X\alpha - \frac{\alpha}{C}) \\ &= \frac{1}{2}(X\alpha)^T(X\alpha) + \frac{\alpha^T\alpha}{2C} + \alpha^T Y - (X\alpha)^T X\alpha - \frac{\alpha^T\alpha}{C} \\ &= -\frac{1}{2}(X\alpha)^T(X\alpha) - \frac{\alpha^T\alpha}{2C} + \alpha^T Y \end{aligned}$$

$g'(\alpha) = -\alpha^T X^T X - \frac{\alpha^T}{C} + Y^T = 0$
$\alpha = (K + \frac{I}{C})^{-1} Y$