

CS 373 HW 3 PDF:
 I want to use 2 late days
 Cr statement: Lingbin Sun

2 perceptron (45pt)
 2.1 theory (15 pt)
 1(5 pt) perceptron model discussed in class is shown as:
 $f(x) = \begin{cases} 1 & \text{if } \sum w_j x_j > 0 \\ 0 & \text{if } \sum w_j x_j \leq 0 \end{cases}$

The bias term is missing in the shown equation. Write the equation with the bias term included

Is perceptron with a bias term more expressive (can represent more classification scenarios) compared to the one without bias? Why or why not?

Solution:

Yes, the perceptron with a bias term is more expressive, because it is often convenient to have a non-zero threshold. In other words, we might want to predict positive if $a > \theta$ for some value θ . The way that is most convenient to achieve this is to introduce a bias term into the neuron, so the activation is always increased by some fixed value b , therefore, we compute

$$a = \sum_{d=1}^D w_d x_d + b$$

This is the complete neural model of learning. The model is parameterized by D many weights w_1, w_2, \dots, w_D , and a single scalar bias value b

Hence,

$$f(x) = \begin{cases} 1 & \text{if } \sum w_j x_j + b > 0 \\ 0 & \text{if } \sum w_j x_j + b \leq 0 \end{cases}$$

Is the equation with bias term included.

2. 5pt) given below are four figures that show the distribution of data points with binary classes. The two colors denote the different classes. For each of these, reason whether the following would give a high (>0.95) classification accuracy.

- i) a perceptron without bias
- ii) a perceptron with bias

	a	b	c	d
I	yes, It can, because the center is in origin	No, without bias the graph cannot be centered at the origin	yes, It pass origin, the line is almost y axis	no, The separate line will be around $x=3$, not passing origin

	a	b	C	D
ii	No, because with bias, for example, if $b > 0$ the classification line (basically a circle) will be shifted upward.	yes, With bias we can adjust it to center in origin	No, with bias the classification line will be shifted not passing origin	yes, With bias we can adjust it to pass origin

3. What is the update rule for the bias term b of a vanilla perceptron, given the learning rate γ and gold label y , when the classifier label doesn't match the gold label during the training?

What is the update rule when the classifier label matches the gold label during training?

Update b with w
 Circle through all examples, until no more errors are made
 Predict the label of instance x to be $y' = \text{sgn}(w \cdot x)$
 If $y' \neq y$, update the weight vector:
 $w = w + \gamma y x$ (γ - a constant, learning rate)
 Otherwise, if $y' = y$, leave weights unchanged.

2.2 Implementation (30 pt)

You need to implement a vanilla perceptron and an averaged perceptron model for this part. Both the models should be implemented with a bias term. This part could be completed by editing only `perceptron.py`, unless you plan to extract any new features for the task. You need to initialize the parameters (`__init__()` method), learn them from training data (`fit()` method) and use the learned parameters to classify new instances (`predict()` method) for each of the models. Take note that `__init__.py`, `man.py` and `classifier.py` may be replaced while grading. Do not make any modification to these files. You need to follow the description of the models discussed in the lecture slides(link). Report the results obtained

3 Naive Bayes 40 pt

3.1 theory

1 5pt) given a text document d which is a sequence of words w_1, w_2, \dots, w_n , we want to compute $P(c+|d)$ and $P(c-|d)$. We use Bayes theorem to estimate the probabilities. Compute the equation for $P(c+|d)$ in terms of $P(d|c+)$ using Bayes theorem

$$P(Y|X) = P(X|Y)P(Y)/P(X)$$

$$= P(X|Y)P(Y) / [P(X|Y=+)P(Y=+) + P(X|Y=-)P(Y=-)]$$

$$P(c+|d) = P(d|c+)P(c+)/P(d)$$

2 5pt) To estimate $P(d|c+)$ using the training data without making any assumptions, we need impractically large amounts of data.

Let us say that the size of the vocabulary is V and length of all documents is exactly l (we can ensure this by padding shorter texts with dummy tokens).

In a binary classification task, how many parameters do we need to learn, in order to correctly estimate $P(d|c+)$ for any given document without making independence assumptions?

$$n=l/v$$
$$2^n$$

3 5pt) If we make the unigram assumption, what is, if we assume that occurrence of each word in the document is independent of other words, then how many parameters do we need to learn to be able to estimate $P(d|c+)$

Independent $n+1$

4 5) in a binary text classification task, mention the equation you would use to estimate $P(c+)$ and $P(c-)$ from the training data ($c+$ and $c-$ are the two classes for the classification problem)

$$P(c+)=\text{size}(c+)/(\text{size}(c+)+\text{size}(c-))$$
$$P(c-)=\text{size}(c-)/(\text{size}(c+)+\text{size}(c-))$$

4 Analysis

You need to perform additional experiments to answer the following questions. You don't need to submit your code for this part. You only need to include your plots and discussions in your report. Make sure that the code you submit does not include any changes you don't want to be included, as that might affect your chances of getting extra credit

- 1) remove bias term from the perceptron and averaged perceptron and report the performances. Is the performance better or worse than the respective models with a bias term? Discuss in one sentence

The model performs less accurate after removing the perception and averaged perceptron compared to the models with a bias term.

- 2) Plot a graph of test accuracy s number of iterations for averaged perceptron for number of iterations= $\{1,2,5,10,20,50\}$. Does Perceptron converge I.e. does the training accuracy become 100? If yes, after how many iterations? If not why

As perceptron does not converge, it would become 0. Ad the data set size goes up, the training accuracy decrease from 75.

- 3) Plot graphs of Naives Bayes test accuracy vs vocabulary size, Averaged Perceptron test-accuracy vs vocabulary size for vocabulary sizes= $\{100,500,1000,5000,10000,20000\}$. How does the performance change? Discuss in two sentences

As the vocal size goes up, the accuracy goes up. The speed of change is decreasing.

Analysis 4.3 Accuracy for Naive Bayers by Xiao Qin

