6. Perceptron Updates

Bookmark this page

Homework due Sep 30, 2020 05:29 IST *Completed*

In this problem, we will try to understand the convergence of perceptron algorithm and its relation to the ordering of the training samples for the following simple example.

Consider a set of $n = d$ labeled $d-$dimensional feature vectors, $\{(x^{(t)}, y^{(t)}), t = 1, \ldots, d\}$ defined as follows:

$$x_i^{(t)} = \cos(\pi t) \quad \text{if} \quad i = t \tag{4.7}$$

$$x_i^{(t)} = 0 \quad \text{otherwise,} \tag{4.8}$$

Recall the no-offset perceptron algorithm, and assume that $\theta \cdot x = 0$ is treated as a mistake, regardless of label. Assume that in all of the following problems, we initialize $\theta = 0$ and when we refer to the perceptron algorithm we only consider the no-offset variant of it.

Working out Perceptron Algorithm

3/3 points (graded)

Consider the $d = 2$ case. Let $y^{(1)} = 1, y^{(2)} = 1$. Assume that the feature vector $x^{(1)}$ is presented to the perceptron algorithm before $x^{(2)}$.

For this particular assignment of labels, work out the perceptron algorithm until convergence.

Let $\hat{\theta}$ be the resulting θ value after convergence. Note that for $d = 2$, $\hat{\theta}$ would be a two-dimensional vector. Let's denote the first and second components of $\hat{\theta}$ by $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively.

Please enter the total number of updates made to θ by perceptron algorithm:

2

✔ **Answer:** 2

Please enter the numerical value of $\hat{\theta}_1$:

-1

✔ **Answer:** -1

Please enter the numerical value of $\hat{\theta}_2$:

1

✔ **Answer:** 1

Solution:

The first iteration of perceptron with data point $(x^{(i)}, y^{(i)})$ will be a mistake due to our initialization of $\theta^{(0)} = \bar{0}$. The first update sets $\theta^{(1)} = y^{(i)} x^{(i)} = x^{(i)}$. The second iteration of perceptron with data point $(x^{(j)}, y^{(j)})$ will also yield a mistake since

$$\theta^{(1)} \cdot x^{(j)} = x^{(i)} \cdot x^{(j)} = 0$$

. Thus, for the $d = 2$ example, after 2 updates $\theta^{(2)} = x^{(i)} + x^{(j)}$. We now check whether the second pass yields mistakes

$$\begin{aligned} y^{(i)} \theta^{(2)} \cdot x^{(i)} &= y^{(i)} \left(y^{(i)} x^{(i)} + y^{(j)} x^{(j)} \right) \cdot x^{(i)} \\ &= \left\| x^{(i)} \right\|^2 + x^{(i)} \cdot x^{(j)} \\ &= \left\| x^{(i)} \right\|^2 + 0 \end{aligned}$$

$$> 0$$

so the first point is classified correctly. Similarly, the second point is also classified correctly.

Therefore, it only takes two updates to classify the $d = 2$ dataset and θ converges to $\hat{\theta} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

General Case - Number of updates

1/1 point (graded)

Now consider any $d > 0$.

For the specific dataset we are considering, please choose the correct answer from the options below:

- (selected) Perceptron algorithm will make exactly d updates to θ regardless of the order and labelings of the feature vectors
- Perceptron algorithm will make at least d updates to θ with the exact number of updates depending on the ordering of the feature vectors presented to it
- Perceptron algorithm will make at least d updates to θ with the exact number of updates depending on the ordering of the feature vectors presented to it and their labeling
- Perceptron algorithm will make at most d updates to θ with the exact number of updates depending on the ordering of the feature vectors presented to it and their labeling

Solution:

Using the intuition that we gained from the $d = 2$ case, we notice that upon the first pass of the data points, every $\theta \cdot x^{(i)} = 0$ for all i. In other words, each data point we consider in sequence lies on the current boundary. Since each point starts as a mistake, there are $\geq d$ updates. Now it remains to be shown that after d updates, all points are classified correctly. After the ith update, we add $y^{(i)} x^{(i)}$ to $\theta^{(i-1)}$. After d updates,

$$\theta^{(d)} = \sum_{i=1}^{d} y^{(i)} x^{(i)}$$

We check whether $y^{(t)} \theta^{(d)} \cdot x^{(t)} > 0$ for all t to ensure there are no mistakes. Notice that the only non-zero term of the dot product occurs when $i = t$. Thus,

$$y^{(t)} \theta^{(d)} \cdot x^{(t)} = \left(y^{(t)}\right)^2 \left\| x^{(t)} \right\|^2 > 0$$

for all $t = 1, 2, \cdots, d$. Therefore, the perceptron algorithm will make exactly d updates regardless of the order and the labelings of the feature vectors.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Sketching convergence

1/1 point (graded)

Consider the case with $d = 3$. Also assume that all the feature vectors are positively labelled. Let P denote the plane through the three points in a 3-d space whose vector representations are given by the feature vectors $x^{(1)}, x^{(2)}, x^{(3)}$.

Let $\hat{\theta}$ denote the value of θ after perceptron algorithm converges for this example. Let v denote the vector connecting the origin and $\hat{\theta}$. Which of the following options is true regarding the vector represented by $\hat{\theta}$.

- ( ) v is parallel to the plane P
- ( ) v is perpendicular to the plane P and pointing away from it
- (•) v is perpendicular to the plane P and pointing towards it
- ( ) $\hat{\theta}$ lies on the plane P

✓

Solution:

Note that from the previous problem

$$\hat{\theta} = \sum_{i=1}^{d} y^{(i)} x^{(i)}$$

Evaluating the above expression, for the current example gives,

$$\hat{\theta} = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

The equation of the plane is given by:

$$\begin{vmatrix} x & -1 & 0 \\ y-1 & -1 & -1 \\ z & 0 & -1 \end{vmatrix} = 0$$

That is,

$$x - y + z = -1$$

Or equivalently,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} = 1$$

Hence, v is perpendicular to P and pointing towards it.

Submit You have used 2 of 3 attempts