Perceptron Learning Algorithm Lecture Supplement

Perceptron Learning Algorithm Convergence

In this section we prove that, when a linearly separable set of training examples $(\overline{x}_1, y_1), \ldots, (\overline{x}_n, y_n)$ is provided as input to the Perceptron Learning algorithm, then the algorithm will eventually terminate, meaning that values $\overline{w}$ and b have been found for which $y_i(\overline{w} \cdot \overline{x}_i - b) > 0$, for all $i = 1, \ldots, n$.

Positive vector sets

To simplify the analysis, notice that, if we add an extra component equal to 1 to vector $\overline{x}_i$, $i = 1, \ldots, n$, then we may think of b as an extra component to $\overline{w}$. Then, after absorbing b into $\overline{w}$, we have $y_i(\overline{w} \cdot \overline{x}_i) > 0$, for all $i = 1, \ldots, n$. Finally, if we replace each $\overline{x}_i$ with the vector $y_i\overline{x}_i$, then after absorbing y_i into $\overline{x}_i$, we have $\overline{w} \cdot \overline{x}_i > 0$, for all $i = 1, \ldots, n$. We say that a set of vectors $\overline{x}_1, \ldots, \overline{x}_n$ is **positive** iff there exists a vector $\overline{w}$ for which $\overline{w} \cdot \overline{x}_i > 0$, for all $i = 1, \ldots, n$.

Example 1. Re-write the linearly separable set of training examples

$$((-1, 1), 1), ((-2, 3), 1), ((1, 3), 1), ((3, -1), -1), ((4, 5), -1),$$

as a set of three-dimensional positive vectors.

Cauchy-Schwarz inequality

Theorem 1 (Cauchy-Schwarz-Bunyakovsky Inequality). If $\overline{u}$ and $\overline{v}$ are vectors in a dot-product vector space, then

$$(\overline{u} \cdot \overline{v})^2 \leq |\overline{u}|^2 |\overline{v}|^2,$$

which implies that

$$|\overline{u} \cdot \overline{v}| \leq |\overline{u}||\overline{v}|.$$

Proof of Theorem 1. Theorem 1 is intuitively true if we recall that $|\overline{u} \cdot \overline{v}|$ is the length of the projection of $\overline{u}$ on to $\overline{v}$, times the length of $\overline{v}$. Then the result is true if we believe that the length of a projection of $\overline{u}$ on to $\overline{v}$ should not exceed the length of $\overline{u}$. The following is a more formal proof.

For any scalar t, by several applications of the four properties of inner products, we have

$$0 \leq (t\overline{u} + \overline{v}) \cdot (t\overline{u} + \overline{v}) = t^2(\overline{u} \cdot \overline{u}) + 2t(\overline{u} \cdot \overline{v}) + \overline{v} \cdot \overline{v} =$$

$$t^2|\overline{u}|^2 + 2t(\overline{u} \cdot \overline{v}) + |\overline{v}|^2,$$

which may be written as $at^2+bt+c \geq 0$, where $a = |\overline{u}|^2$, $b = 2(\overline{u} \cdot \overline{v})$, and $c = |\overline{v}|^2$. But $at^2+bt+c \geq 0$ implies that the equation $at^2 + bt + c = 0$ either has no roots, or exactly one root. In other words, we must have

$$b^2 - 4ac \leq 0,$$

which implies

$$4(\overline{u} \cdot \overline{v})^2 \leq 4|\overline{u}|^2 |\overline{v}|^2,$$

or

$$(\overline{u} \cdot \overline{v})^2 \leq |\overline{u}|^2 |\overline{v}|^2.$$

Convergence Theorem

Theorem 2. Let $x_1, \ldots, x_n$ be a set of positive vectors. Then the Perceptron Learning algorithm determines a weight vector $\overline{w}$ for which $\overline{w} \cdot \overline{x}_i > 0$, for all $i = 1, \ldots, n$.

Proof of Theorem 2. Since the set of input vectors is positive, there is a weight vector $\overline{w^*}$ for which $|\overline{w^*}| = 1$, and there exists a $\delta > 0$ for which, for $i = 1, 2, \ldots, n$,

$$|\overline{w^*} \cdot \overline{x}_i| > \delta.$$

Furthermore, let $r > 0$ be such that $|\overline{x}_i| \leq r$, for all $i = 1, \ldots, n$. Let k be the number of times the vector $\overline{w}$ in the perceptron learning algorithm has been updated, and let $\overline{w}_k$ denote the value of the weight vector after the k th update. We assume $\overline{w}_0 = \overline{0}$; i.e. the algorithm begins with a zero weight vector. The objective is to show that k must be bounded. Suppose $\overline{x}_i$ is used for the k th update in the algorithm. Then $\overline{w}_k$ can be recursively written as

$$\overline{w}_k = \overline{w}_{k-1} + x_i,$$

where $\overline{w}_{k-1} \cdot \overline{x}_i \leq 0$.

Claim. $|\overline{w}_k|^2 \leq kr^2$.

The proof of this claim is by induction on k. For $k = 0$, $\overline{w}_0 = \overline{0}$, and so $|\overline{w}_0|^2 = 0 \leq 0(r^2) = 0$.

For the inductive step, assume that $|\overline{w}_j|^2 \leq jr^2$, for all $j < k$. Then

$$|\overline{w}_k|^2 = |\overline{w}_{k-1} + \overline{x}_i|^2 = (\overline{w}_{k-1} + \overline{x}_i) \cdot (\overline{w}_{k-1} + \overline{x}_i) \leq$$
$$|\overline{w}_{k-1}|^2 + r^2 \leq (k-1)r^2 + r^2 = kr^2,$$

and the claim is proved.

Thus, $|\overline{w}_k| \leq r\sqrt{k}$.

Next, we may use induction a second time to prove a lower bound on $\overline{w^*} \cdot \overline{w}_k$, namely that $\overline{w^*} \cdot \overline{w}_k \geq k\delta$. This is certainly true for $k = 0$. Now if the inductive assumption is that $\overline{w^*} \cdot \overline{w}_{k-1} \geq (k-1)\delta$, then

$$\overline{w^*} \cdot \overline{w}_k = \overline{w^*} \cdot (\overline{w}_{k-1} + \overline{x}_i) =$$
$$\overline{w^*} \cdot \overline{w}_{k-1} + \overline{w^*} \cdot \overline{x}_i \geq \overline{w^*} \cdot \overline{w}_{k-1} + \delta \geq (k-1)\delta + \delta = k\delta,$$

and the lower bound is proved.

Finally, applying the Cauchy-Schwarz inequality, we have

$$|\overline{w^*}| \cdot |\overline{w}_k| \geq \overline{w^*} \cdot \overline{w}_k \geq k\delta.$$

And since $|\overline{w^*}| = 1$, this implies $|\overline{w}_k| \geq k\delta$.

Putting the two inequalities together yields $k\delta \leq r\sqrt{k}$, which yields $k \leq \frac{r^2}{\delta^2}$. Therefore, k is bounded, and the algorithm must terminate.

Exercises

1. Describe five features that could be used for the purpose of classifying a fish as either a salmon or a trout.

2. Plot the training samples $((0,0),+1)$, $((0,1),-1)$, $((1,0),-1)$, $((1,1),+1)$ and verify that the two classes are *not* linearly separable. Then provide an algebraic proof. Hint: assume $\overline{w} = (w_1, w_2)$ and b are the parameters of a separating line, and obtaine a contradiction.

3. If vector $\overline{w} = (-2, 1, 5)$ is normal to plane P and P contains the point $(0, 0, -5)$, then provide an equation for P.

4. Provide an equation of a plane P that is normal to vector $\overline{w} = (1, -1, 3)$ and passes through the point $(0, 1, -2)$.

5. If the vector $\overline{v} = (2, 1, 5)$ makes a 60-degree angle with a unit vector $\overline{u}$, compute $\overline{u} \cdot \overline{v}$.

6. Prove that the Cauchy-Schwarz inequality becomes an equality iff $\overline{v} = k\overline{u}$, for some constant k.

7. Establish that, for any n-dimensional vector v, $|v| = \sqrt{v \cdot v}$.

8. Given the feature vectors from the two classes

$$C_+ = (0.1, -0.2), (0.2, 0.1), (-0.15, 0.1), (1.1, 0.8), (1.2, 1.1),$$

and

$$C_- = (1.1, -0.1), (1.25, 0.15), (0.9, 0.1), (0.1, 1.2), (0.2, 0.9),$$

Compute the centers $\mathbf{c}_+$ and $\mathbf{c}_-$ and provide the equation of the Simple-Learning algoirthm decision surface. Use the decision surface to classify. the vector $(0.5, 0.5)$.

9. Give an example using only three linearly separable training vectors, where the surface obtained from the Simple-Learning algorithm misclassifies at least one of the training vectors.

10. Re-write the linearly separable set of training examples

$$((1,1),1), ((0,2),1), ((3,0),1), ((-2,-1),-1), ((0,-2),-1),$$

as a set of three-dimensional positive vectors.

11. Demonstrate the Perceptron Learning algorithm with $\eta = 1$ using the positive vectors obtained from the previous exercise as input. Start with $\overline{w}_0 = \overline{0}$, and use the order

$$(0,2,-1), (2,1,-1), (3,0,1), (1,1,1), (0,2,1)$$

when checking for misclassifications. Compute the final normal vector $\overline{w^*}$, and verify that the surface $(\overline{w^*}_1, \overline{w^*}_2) \cdot \overline{x} = -\overline{w^*}_3$ separates the original data.

Exercise Solutions

1. Answers may vary. Here are five that come to mind: weight (grams), length from head to tail (cm), girth (cm), number of fins (1-10), primary color.

2. Assume the training samples are separated by the line $\overline{w} \cdot \overline{x} = b$, where $\overline{w} = (w_1, w_2)$. Then i) $\overline{w} \cdot (1,1) = w_1 + w_2 \geq b$, ii) $\overline{w} \cdot (0,0) = 0 \geq b$, iii) $\overline{w} \cdot (1,0) = w_1 < b$, and iv) $\overline{w} \cdot (0,1) = w_2 < b$. Then iii) and iv) yield $w_1 + w_2 < 2b$, and combining this with i) yields $b < 2b$, or $b > 0$, which contradicts ii). Therefore, the training samples are not linearly separable.

3. Since
$$b = \overline{w} \cdot (0,0,-5) = (-2,1,5) \cdot (0,0,-5) = 25,$$
the equation is $\overline{w} \cdot \overline{x} = 25$.

4. Since
$$b = \overline{w} \cdot (0,1,-2) = (0,1,-2) \cdot (1,-1,3) = -7,$$
the equation is $\overline{w} \cdot \overline{x} = -7$.

5. $$\overline{u} \cdot \overline{v} = |\overline{u}||\overline{v}| \cos 60^\circ = (\sqrt{30})(1)(1/2) = \frac{\sqrt{30}}{2}.$$

6. If $\overline{v} = k\overline{u}$, for some constant k, then
$$|\overline{u} \cdot \overline{v}| = |\overline{u} \cdot (k\overline{u})| = |k|(\overline{u} \cdot \overline{u}) = |k||\overline{u}||\overline{u}| = |\overline{v}||\overline{u}| = |\overline{u}||\overline{v}|.$$
Now assume that $|\overline{u} \cdot \overline{v}| = |\overline{u}||\overline{v}|$. Without loss of generality, assume that $|\overline{u}| = 1$. Then
$$\overline{w} = \mathrm{proj}(\overline{v}, \overline{u}) = (\overline{u} \cdot \overline{v})\overline{u}.$$
Now consider $\overline{v} - \overline{w}$. Then
$$|\overline{v} - \overline{w}|^2 = (\overline{v} - \overline{w}) \cdot (\overline{v} - \overline{w}) = |\overline{v}|^2 + |\overline{w}|^2 - 2(\overline{v} \cdot \overline{w}) =$$
$$|\overline{v}|^2 + (\overline{u} \cdot \overline{v})^2 - 2(\overline{u} \cdot \overline{v})(\overline{u} \cdot \overline{v}) = 0,$$
since $|\overline{v}|^2 = (1)|\overline{v}|^2 = |\overline{u}|^2|\overline{v}|^2 = (\overline{u} \cdot \overline{v})^2$. Hence, $|\overline{v} - \overline{w}| = 0$, which implies $\overline{v}$ is a multiple of $\overline{w}$, which in turn is a multiple of $\overline{u}$.

7. $$\sqrt{\overline{v} \cdot \overline{v}} = \sqrt{v_1^2 + \cdots + v_n^2} = |\overline{v}|.$$

8. $\overline{c}_+ = (0.49, 0.38)$, while $\overline{c}_- = (0.71, 0.45)$. Then $\overline{w} = \overline{c}_+ - \overline{c}_- = (-0.22, -0.07)$, and $\overline{c} = 1/2(1.2, 0.83) = (0.6, 0.415)$. Finally, $b = \overline{w} \cdot \overline{c} = -0.16105$. The equation of the decision surface is thus
$$(-0.22, -0.07) \cdot \overline{x} = -0.16105.$$
Then
$$(-0.22, -0.07) \cdot (0.5, 0.5) = -0.145 > -0.16105,$$
which implies that $(0.5, 0.5)$ is classified as being in Class $+1$.

9. Consider $((0,-1),-1)$, $((0,0),1)$, and $((0,4),1)$. Then $\overline{c}_+ = (0,2)$, while $\overline{c}_- = (0,-1)$. Then $\overline{w} = \overline{c}_+ - \overline{c}_- = (0,3)$, and $\overline{c} = 1/2(0,1) = (0,0.5)$. Finally, $b = \overline{w} \cdot \overline{c} = 1.5$. The equation of the decision surface is thus
$$(0,3) \cdot \overline{x} = 1.5.$$
Then
$$(0,3) \cdot (0,0) = 0 < 1.5,$$
which implies that $(0,0)$ is misclassified as being in Class -1.

10. Adding a $+1$ component to each vector yields
$$(1,1,1),(0,2,1),(3,0,1),(-2,-1,1),(0,-2,1).$$
Then scaling each vector with its class label yields
$$(1,1,1),(0,2,1),(3,0,1),(2,1,-1),(0,2,-1).$$

11.
$$\overline{w}_0 \cdot (1,1,1) = 0 \Rightarrow \overline{w}_1 = \overline{w}_0 + (0,2,-1) = (0,2,-1).$$
$$\overline{w}_1 \cdot (3,0,1) = -1 \Rightarrow \overline{w}_2 = \overline{w}_1 + (3,0,1) = (3,2,0).$$
$\overline{w}_2 \cdot \overline{x} > 0$ for each training vector $\overline{x}$. Therefore, $\overline{w^*} = \overline{w}_2 = (3,2,0)$. Finally, the decision surface to the original set of training vectors (see previous exercise) has equation $(3,2) \cdot \overline{x} = 0$. Notice that $(3,2) \cdot \overline{x} > 0$ for every $\overline{x}$ in Class $+1$, and $(3,2) \cdot \overline{x} < 0$ for every $\overline{x}$ in Class -1, which is the desired result.