Perceptrons

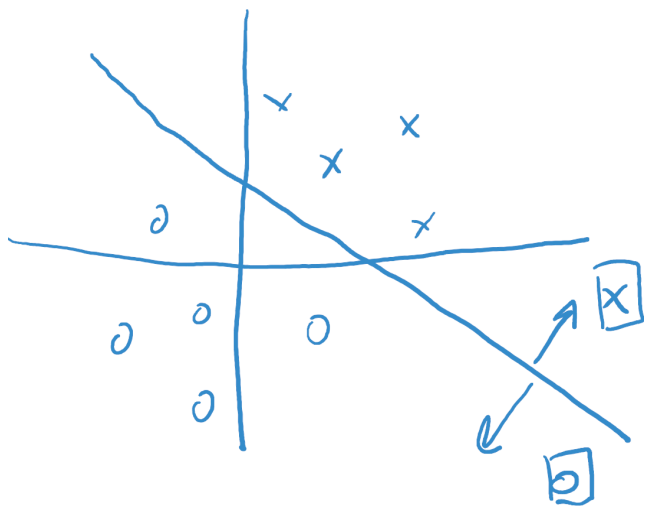
(/edit_new.html#/pages/groups/18)

MODERATE

1 ¶
2 ¶
3 ¶
4 ¶
5 ¶
6 ¶
7 ¶

1. Linear discriminant

A linear discriminant is one of the simplest classifiers in machine learning. It looks like this:



Given a point $(x, y) \in \mathrm{R}^2$, the linear discriminant predicts that it is positive (marked as x) if

$$ax + by + k > 0$$

It predicts the point is negative (marked as o) if $ax + by + k < 0$, and offers no opinion if $ax + by + k = 0$. Here a, b, and k are adjustable parameters, which let us fit the perceptron to observed data.

For convenience, we will write

$$w = \begin{pmatrix} a \\ b \\ k \end{pmatrix}$$

for our vector of parameters, and

$$u = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

for a training example; then the classification rule is to predict positive (x) when

$$w \cdot u > 0\,.$$

---

2. Streaming data

We'd like to learn from *streaming data*: that is, we observe a sequence of data points $(x_t, y_t) \in \mathrm{R}^2$ one at a time. After each point, we predict a label $\hat{\ell}_t \in \{-1, +1\}$; then we get to see the true label ℓ_t before we need to predict the next point.

---

3. The perceptron algorithm

To learn from streaming data, we can use any of several algorithms. One of the simplest is the *perceptron algorithm*: we initialize our parameters as $w_1 = (0, 0, 0)^T$. As long as we predict correctly, we leave our parameters unchanged, $w_{t+1} = w_t$. Whenever we make a mistake, we update our parameters: if we should have predicted x we set

$$w_{t+1} = w_t + u_t$$

and if we should have predicted o we set

$$w_{t+1} = w_t - u_t$$

This rule makes sense since

$$w_{t+1} \cdot u_t > w_t \cdot u_t$$

in the first case, and

$$w_{t+1} \cdot u_t < w_t \cdot u_t$$

in the second. (To see why, expand out the dot product on the LHS and use $u_t \cdot u_t > 0$.) So, we are more likely to predict the correct label if we see the same example again.

---

4. Mistake bound

The perceptron algorithm satisfies many nice properties. Here we'll prove a simple one, called a *mistake bound*: if there exists an optimal parameter vector w^* that can classify all of our examples correctly, then the perceptron algorithm will make at most a small number of mistakes before discovering an optimal parameter vector.

In more detail, suppose that $w^* \cdot u_t > \epsilon$ for all positive examples, and $w^* \cdot u_t < -\epsilon$ for all negative ones. Also assume that our examples are bounded: there is a constant U such that $\|u_t\| \leq U$ for all t.

Then, the perceptron algorithm will make at most

$$\frac{U^2 \|w^*\|^2}{\epsilon^2}$$

mistakes in total. For example, if our examples have norm at most 2, if our optimal parameter vector has norm $\|w^*\| = 3$, and if $\epsilon = \frac{1}{2}$, then the number of mistakes M satisfies

$$M \leq 144\,.$$

---

5. Proof of mistake bound, part I

First we show a lower bound on $w_t \cdot w^*$.

After a mistake on a positive example, we have

$$\begin{aligned} w_{t+1} \cdot w^* &= w_t \cdot w^* + u_t \cdot w^* \\ &\geq w_t \cdot w^* + \epsilon \end{aligned}$$

since, by assumption, $u_t \cdot w^* \geq \epsilon$. Similarly, on a negative example, we have

$$\begin{aligned} w_{t+1} \cdot w^* &= w_t \cdot w^* - u_t \cdot w^* \\ &\geq w_t \cdot w^* + \epsilon \end{aligned}$$

since, by assumption, $u_t \cdot w^* \leq -\epsilon$. So, after M mistakes, we have

$$w_t \cdot w^* \geq \epsilon M$$

by induction: the LHS starts at 0, and increases by at least ϵ with each mistake.

---

6. Proof, part II

Next we show an upper bound on $\|w^t\|$. After a mistake on a positive example, we have

$$\begin{aligned} w_{t+1} \cdot w_{t+1} &= w_t \cdot w_t + 2w_t \cdot u_t + u_t \cdot u_t \\ &\leq w_t \cdot w_t + 0 + U^2 \end{aligned}$$

To see why, note that $w_t \cdot u_t \leq 0$, since we (mistakenly) classified this example as negative. And, $u_t \cdot u_t = \|u_t\|^2 \leq U^2$ by assumption.

Similarly, after a mistake on a negative example, we have

$$\begin{aligned} w_{t+1} \cdot w_{t+1} &= w_t \cdot w_t - 2w_t \cdot u_t + u_t \cdot u_t \\ &\leq w_t \cdot w_t + 0 + U^2 \end{aligned}$$

In this case, $w_t \cdot u_t \geq 0$, since we (mistakenly) classified this example as positive.

So, after M mistakes, we have

$$w_t \cdot w_t \leq MU^2$$

since $w_t \cdot w_t$ starts at zero, doesn't change unless we make a mistake, and increases by at most U^2 on each mistake. Rewriting, we have

$$\|w_t\| \leq U\sqrt{M}$$

for all t.

7. Proof, part III

From part I we have

$$M \leq \frac{w_t \cdot w^*}{\epsilon}$$

By Hölder's inequality, we therefore have

$$M \leq \frac{\|w_t\| \|w^*\|}{\epsilon}$$

Substituting in the conclusion of part II, we get

$$M \leq \frac{U\sqrt{M}\|w^*\|}{\epsilon}$$

and rearranging we get

$$M \leq \frac{U^2\|w^*\|^2}{\epsilon^2}$$

as desired.