The Perceptron Mistake Bound

Machine Learning

THE UNIVERSITY OF UTAH

Some slides based on lectures from Dan Roth, Avrim Blum and others

Where are we?

- The Perceptron Algorithm
- Variants of Perceptron
- Perceptron Mistake Bound

Convergence

Convergence theorem

- If there exist a set of weights that are consistent with the data (i.e. the data is linearly separable), the perceptron algorithm will converge.

Convergence

Convergence theorem

– If there exist a set of weights that are consistent with the data (i.e. the data is linearly separable), the perceptron algorithm will converge.

Cycling theorem

– If the training data is *not* linearly separable, then the learning algorithm will eventually repeat the same set of weights and enter an infinite loop

Perceptron Learnability

- Obviously Perceptron cannot learn what it cannot represent
  - Only linearly separable functions

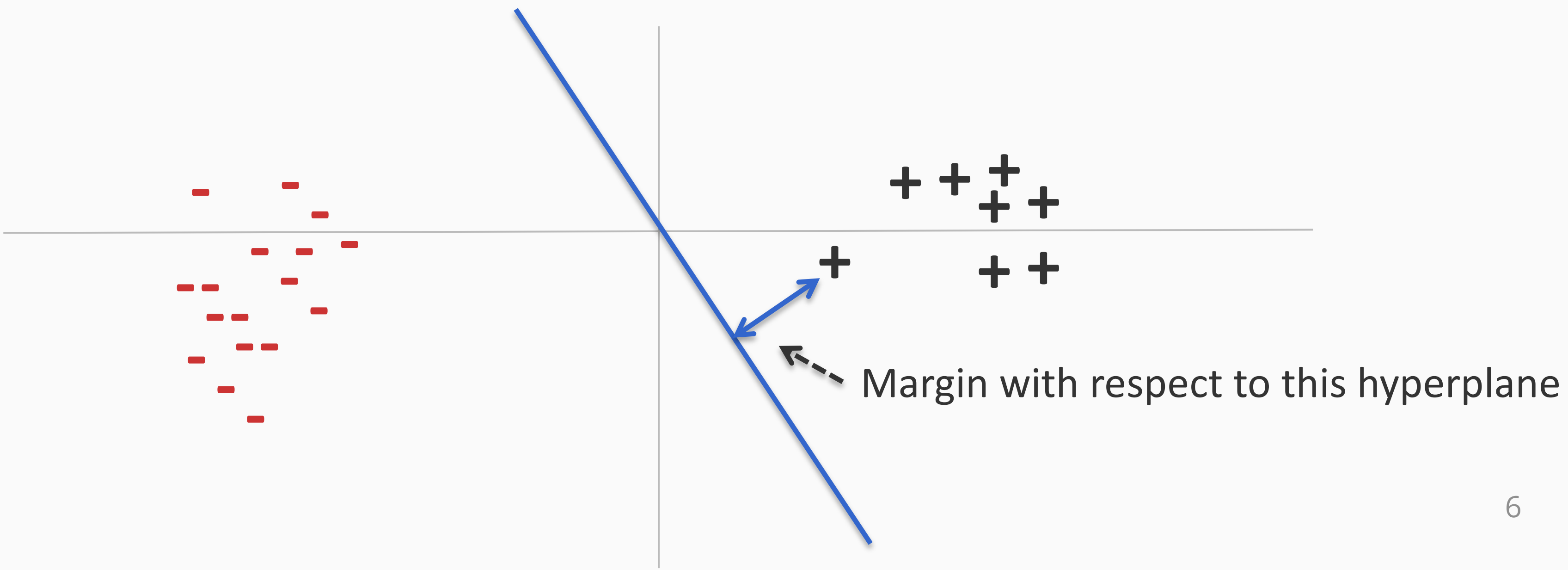
- Minsky and Papert (1969) wrote an influential book demonstrating Perceptron's representational limitations

  - Parity functions can't be learned (XOR)
    - We have already seen that XOR is not linearly separable

  - In vision, if patterns are represented with local features, can't represent symmetry, connectivity

Margin

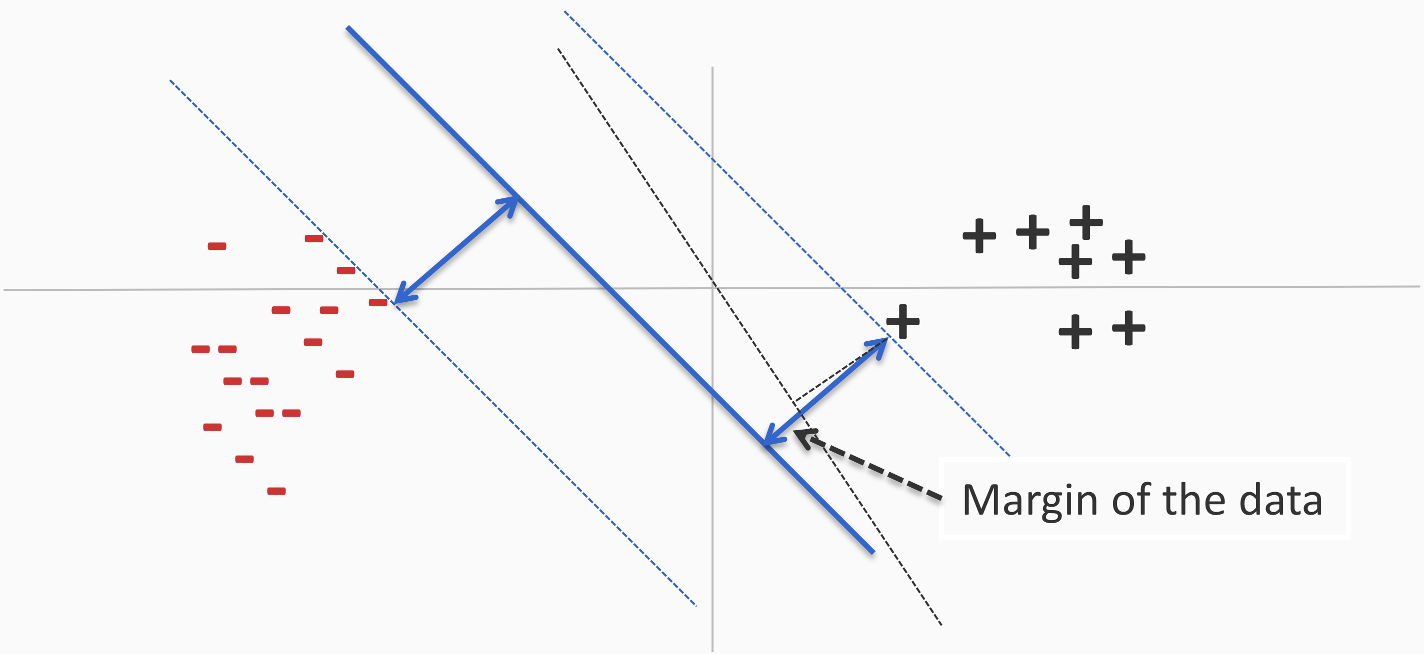
The margin of a hyperplane for a dataset is the distance between the hyperplane and the data point nearest to it.



Margin

The margin of a hyperplane for a dataset is the distance between the hyperplane and the data point nearest to it.

The margin of a data set (γ) is the maximum margin possible for that dataset using any weight vector.



Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

We can always find such an R. Just look for the farthest data point from the origin.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

The data has a margin γ.
Importantly, the data is *separable*.
γ is the complexity parameter that defines the separability of data.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

Then, the perceptron algorithm will make no more than R^2/γ^2 mistakes on the training sequence.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

Then, the perceptron algorithm will make no more than R^2/γ^2 mistakes on the training sequence.

If **u** hadn't been a unit vector, then we could scale it in the mistake bound. This will change the final mistake bound to $\left(\frac{||\mathrm{u}||\mathrm{R}}{\gamma}\right)^2$.

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose we have a binary classification dataset with n dimensional inputs.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

If the data is separable,…

Then, the perceptron algorithm will make no more than R^2/γ^2 mistakes on the training sequence.

…then the Perceptron algorithm will find a separating hyperplane after making a finite number of mistakes

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\text{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

Proof (preliminaries)

The setting

- Initial weight vector $\mathbf{w}$ is all zeros
- Learning rate = 1
  - Effectively scales inputs, but does not change the behavior
- All training examples are contained in a ball of size R.
  - That is, for every example $(\mathbf{x}_i, y_i)$, we have
$$||\mathbf{x}_i|| \leq R$$
- The training data is separable by margin γ using a unit vector $\mathbf{u}$.
  - That is, for every example $(\mathbf{x}_i, y_i)$, we have
$$y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$$

Proof (1/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\operatorname{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

1. Claim: After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$

Proof (1/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\text{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

1. Claim: After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$

$$\mathbf{u}^T \mathbf{w}_{t+1} \quad = \quad \mathbf{u}^T \mathbf{w}_t + y_i \mathbf{u}^T \mathbf{x}_i$$

Proof (1/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\operatorname{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

1. Claim: After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$

$$\begin{aligned} \mathbf{u}^T \mathbf{w}_{t+1} &= \mathbf{u}^T \mathbf{w}_t + y_i \mathbf{u}^T \mathbf{x}_i \\ &\geq \mathbf{u}^T \mathbf{w}_t + \gamma \end{aligned}$$

Because the data is separable by a margin γ

Proof (1/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\mathrm{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

1. Claim: After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$

$$\begin{aligned} \mathbf{u}^T \mathbf{w}_{t+1} &= \mathbf{u}^T \mathbf{w}_t + y_i \mathbf{u}^T \mathbf{x}_i \\ &\geq \mathbf{u}^T \mathbf{w}_t + \gamma \end{aligned}$$

Because the data is separable by a margin γ

Because $\mathbf{w}_0 = \mathbf{0}$ (that is, $\mathbf{u}^T \mathbf{w}_0 = \mathbf{0}$), straightforward induction gives us $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$

Proof (2/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\mathrm{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

2. Claim: After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

Proof (2/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\operatorname{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

2. Claim: After t mistakes, $\left|\left|\mathbf{w}_t\right|\right|^2 \leq tR^2$

$$
\begin{aligned}
\left|\left|\mathbf{w}_{t+1}\right|\right|^2 &= \left|\left|\mathbf{w}_t + y_i \mathbf{x}_i\right|\right|^2 \\
&= \left|\left|\mathbf{w}_t\right|\right|^2 + 2y_i(\mathbf{w}_t^T \mathbf{x}_i) + \left|\left|\mathbf{x}_i\right|\right|^2
\end{aligned}
$$

Proof (2/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\text{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

2. Claim: After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$\begin{aligned} ||\mathbf{w}_{t+1}||^2 &= ||\mathbf{w}_t + y_i \mathbf{x}_i||^2 \\ &= ||\mathbf{w}_t||^2 + 2y_i(\mathbf{w}_t^T \mathbf{x}_i) + ||\mathbf{x}_i||^2 \end{aligned}$$

The weight is updated only when there is a mistake. That is when $y_i \mathbf{w}_t^T \mathbf{x}_i < 0$.

$||\mathbf{x}_i|| \leq R$, by definition of R

Proof (2/3)

- Receive an input $(\mathbf{x}_i, y_i)$
- if $\operatorname{sgn}(\mathbf{w}_t^T \mathbf{x}_i) \neq y_i$:
  Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$

2. Claim: After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$
\begin{aligned}
||\mathbf{w}_{t+1}||^2 &= ||\mathbf{w}_t + y_i \mathbf{x}_i||^2 \\
&= ||\mathbf{w}_t||^2 + 2y_i(\mathbf{w}_t^T \mathbf{x}_i) + ||\mathbf{x}_i||^2 \\
&\leq ||\mathbf{w}_t||^2 + R^2
\end{aligned}
$$

Because $\mathbf{w}_0 = \mathbf{0}$ (that is, $\mathbf{u}^T \mathbf{w}_0 = \mathbf{0}$),
straightforward induction gives us $||\mathbf{w}_t||^2 \leq tR^2$

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$
2. After t mistakes, $\left|\left|\mathbf{w}_t\right|\right|^2 \leq tR^2$

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T\mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\|$$

From (2)

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T\mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\| \geq \mathbf{u}^T\mathbf{w}_t$$

From (2)

$\boldsymbol{u}^T\mathbf{w}_t = ||\mathbf{u}||||\mathbf{w}_t||cos(\text{angle between them})$

But $||\mathbf{u}|| = 1$ and cosine is less than 1

So $\mathbf{u}^T\mathbf{w}_t \leq ||\mathbf{w}_t||$

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T\mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\| \geq \mathbf{u}^T\mathbf{w}_t$$

From (2)

$$\boldsymbol{u}^T\mathbf{w}_t = ||\mathbf{u}||||\mathbf{w}_t||cos(\text{angle between them})$$

But $||\mathbf{u}|| = 1$ and cosine is less than 1

So $\mathbf{u}^T\mathbf{w}_t \leq ||\mathbf{w}_t||$

(alternatively, using the Cauchy-Schwarz inequality)

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T\mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\| \geq \mathbf{u}^T\mathbf{w}_t \geq t\gamma$$

From (2)

From (1)

$\boldsymbol{u}^T\mathbf{w}_t = ||\mathbf{u}||||\mathbf{w}_t||cos(\text{angle between them})$

But $||\mathbf{u}|| = 1$ and cosine is less than 1

So $\mathbf{u}^T\mathbf{w}_t \leq ||\mathbf{w}_t||$

(alternatively, using the Cauchy-Schwarz inequality)

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T\mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\| \geq \mathbf{u}^T\mathbf{w}_t \geq t\gamma$$

Number of mistakes $t \leq \frac{R^2}{\gamma^2}$

Proof (3/3)

What we know:

1. After t mistakes, $\mathbf{u}^T \mathbf{w}_t \geq t\gamma$
2. After t mistakes, $||\mathbf{w}_t||^2 \leq tR^2$

$$R\sqrt{t} \geq \|\mathbf{w}_t\| \geq \mathbf{u}^T \mathbf{w}_t \geq t\gamma$$

Number of mistakes $t \leq \frac{R^2}{\gamma^2}$

Bounds the total number of mistakes!

Mistake Bound Theorem [Novikoff 1962, Block 1962]

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots$ be a sequence of training examples such that every feature vector $\mathbf{x}_i \in \Re^n$ with $||\mathbf{x}_i|| \leq R$ and the label $y_i \in \{-1, 1\}$.

Suppose there is a unit vector $\mathbf{u} \in \Re^n$ (i.e., $||\mathbf{u}|| = 1$) such that for some positive number $\gamma \in \Re, \gamma > 0$, we have $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for every example $(\mathbf{x}_i, y_i)$.

Then, the perceptron algorithm will make no more than R^2/γ^2 mistakes on the training sequence.

The Perceptron Mistake bound

Number of mistakes $\leq \frac{R^2}{\gamma^2}$

- R is a property of the dimensionality. How?
  - For Boolean functions with n attributes, show that $R^2 = n$.

- γ is a property of the data

- Exercises:
  - How many mistakes will the Perceptron algorithm make for disjunctions with n attributes?
    - What are R and γ?
  - How many mistakes will the Perceptron algorithm make for k-disjunctions with n attributes?
  - Find a sequence of examples that will force the Perceptron algorithm to make $O(n)$ mistakes for a concept that is a k-disjunction.

Beyond the separable case

- Good news
  - Perceptron makes no assumption about data distribution, could be even adversarial
  - After a fixed number of mistakes, you are done. Don't even need to see any more data
- Bad news: Real world is not linearly separable
  - Can't expect to *never* make mistakes again
  - What can we do: more features, try to be linearly separable if you can, use averaging

What you need to know

- What is the perceptron mistake bound?
- How to prove it

Summary: Perceptron

- Online learning algorithm, very widely used, easy to implement
- Additive updates to weights
- Geometric interpretation
- Mistake bound
- Practical variants abound
- You should be able to implement the Perceptron algorithm and its variants, and also prove the mistake bound theorem