Learning, Linear Separability and Linear Programming

Lecture 22

November 12, 2013

Labeling...

1. given examples:a database of cars.
2. like to determine which cars are sport cars..
3. Each car record: interpreted as point in high dimensions.
4. Example: sport car with 4 doors, manufactured in 1997, by Quaky (with manufacturer ID 6): $(4, 1997, 6)$.
   Labeled as a sport car.
5. Tractor by General Mess (manufacturer ID 3) in 1998: $(0, 1997, 3)$
   Labeled as not a sport car.
6. Real world: hundreds of attributes. In some cases even millions of attributes!
7. Automate this ***classification*** process: label sports/regular car automatically.

Labeling...

1. given examples:a database of cars.
2. like to determine which cars are sport cars..
3. Each car record: interpreted as point in high dimensions.
4. Example: sport car with 4 doors, manufactured in 1997, by Quaky (with manufacturer ID 6): $(4, 1997, 6)$.
   Labeled as a sport car.
5. Tractor by General Mess (manufacturer ID 3) in 1998: $(0, 1997, 3)$
   Labeled as not a sport car.
6. Real world: hundreds of attributes. In some cases even millions of attributes!
7. Automate this ***classification*** process: label sports/regular car automatically.

Labeling...

1. given examples:a database of cars.
2. like to determine which cars are sport cars..
3. Each car record: interpreted as point in high dimensions.
4. Example: sport car with 4 doors, manufactured in 1997, by Quaky (with manufacturer ID 6): $(4, 1997, 6)$.
   Labeled as a sport car.
5. Tractor by General Mess (manufacturer ID 3) in 1998: $(0, 1997, 3)$
   Labeled as not a sport car.
6. Real world: hundreds of attributes. In some cases even millions of attributes!
7. Automate this ***classification*** process: label sports/regular car automatically.

Labeling...

1. given examples:a database of cars.
2. like to determine which cars are sport cars..
3. Each car record: interpreted as point in high dimensions.
4. Example: sport car with 4 doors, manufactured in 1997, by Quaky (with manufacturer ID 6): $(4, 1997, 6)$.
   Labeled as a sport car.
5. Tractor by General Mess (manufacturer ID 3) in 1998: $(0, 1997, 3)$
   Labeled as not a sport car.
6. Real world: hundreds of attributes. In some cases even millions of attributes!
7. Automate this ***classification*** process: label sports/regular car automatically.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

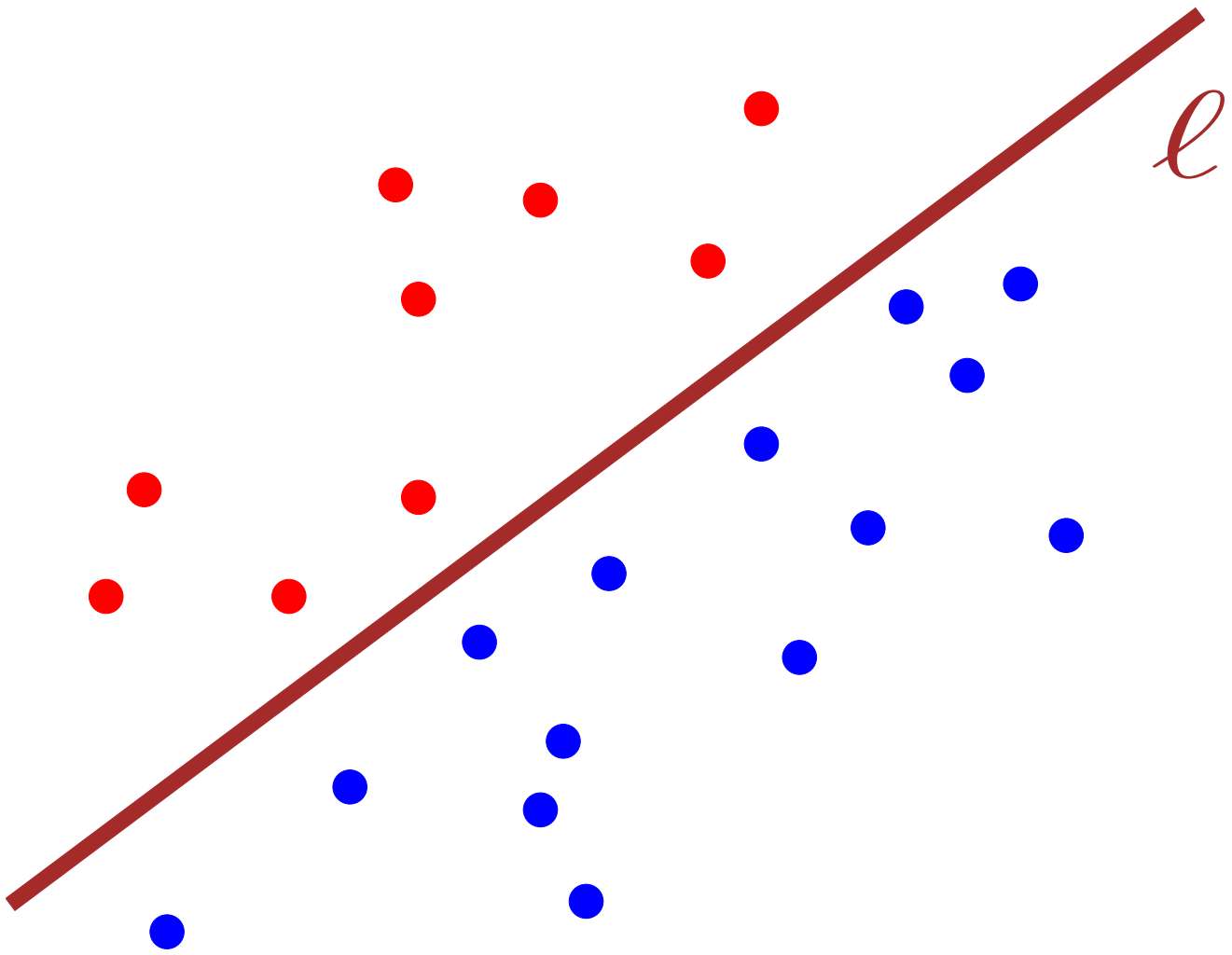
Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Automatic classification...

1. learning algorithm:
   1. given several (or many) classified examples...
   2. ...develop its own conjecture for rule of classification.
   3. ... can use it for classifying new data.
2. ***learning***: ***training*** + ***classifying***.
3. Learn a function: $f : \mathbb{R}^d \to \{-1, 1\}$.
4. challenge: f might have infinite complexity...
5. ...rare situation in real world. Assume learnable functions.
6. red and blue points that are linearly separable.
7. Trying to learn a line ℓ that separates the red points from the blue points.

Linear separability example...



Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Learning linear separation

1. Given red and blue points – how to compute the separating line ℓ?
2. line/plane/hyperplane is the zero set of a linear function.
3. Form: $\forall x \in \mathbb{R}^d \qquad f(x) = \langle a, x \rangle + b,$
   where $a = (a_1, \ldots, a_d)\,, b = (b_1, \ldots, b_d) \in \mathbb{R}^2$.
   $\langle a, x \rangle = \sum_i a_i x_i$ is the ***dot product*** of a and x.
4. classification done by computing sign of $f(x)$: $\mathrm{sign}(f(x))$.
5. If $\mathrm{sign}(f(x))$ is negative: x is not in class.
   If positive: inside.
6. A set of ***training examples***:

$$S = \left\{ (x_1, y_1)\,, \ldots, (x_n, y_n) \right\},$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, for $i = 1, \ldots, n$.

Classification...

1. ***linear classifier*** h: (w, b) where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
2. classification of $x \in \mathbb{R}^d$ is $\mathrm{sign}(\langle w, x \rangle + b)$.
3. ***labeled*** example (x, y), h classifies (x, y) <u>correctly</u> if $\mathrm{sign}(\langle w, x \rangle + b) = y$.
4. Assume a linear classifier exists.
5. Given n labeled example. How to compute the linear classifier for these examples?
6. Use linear programming....
7. looking for $(\mathbf{w}, b)$, such that for an $(\mathbf{x}_i, y_i)$ we have $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = y_i$, which is

$$\begin{aligned} & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 0 && \text{if } y_i = 1, \\ \text{and} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq 0 && \text{if } y_i = -1. \end{aligned}$$

Classification...

1. ***linear classifier*** h: (w, b) where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
2. classification of $x \in \mathbb{R}^d$ is $\mathrm{sign}(\langle w, x \rangle + b)$.
3. ***labeled*** example (x, y), h classifies (x, y) <u>correctly</u> if $\mathrm{sign}(\langle w, x \rangle + b) = y$.
4. Assume a linear classifier exists.
5. Given n labeled example. How to compute the linear classifier for these examples?
6. Use linear programming....
7. looking for (w, b), such that for an (x_i, y_i) we have $\mathrm{sign}(\langle \mathrm{w}, \mathrm{x}_i \rangle + b) = y_i$, which is

$$\begin{aligned} & \langle \mathrm{w}, \mathrm{x}_i \rangle + b \geq 0 && \text{if } y_i = 1, \\ \text{and} \quad & \langle \mathrm{w}, \mathrm{x}_i \rangle + b \leq 0 && \text{if } y_i = -1. \end{aligned}$$

Classification...

1. ***linear classifier*** h: (w, b) where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
2. classification of $x \in \mathbb{R}^d$ is $\mathrm{sign}(\langle w, x \rangle + b)$.
3. ***labeled*** example (x, y), h classifies (x, y) <u>correctly</u> if $\mathrm{sign}(\langle w, x \rangle + b) = y$.
4. Assume a linear classifier exists.
5. Given n labeled example. How to compute the linear classifier for these examples?
6. Use linear programming....
7. looking for $(\mathbf{w}, b)$, such that for an $(\mathbf{x}_i, y_i)$ we have $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = y_i$, which is

$$\begin{aligned} & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 0 && \text{if } y_i = 1, \\ \text{and} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq 0 && \text{if } y_i = -1. \end{aligned}$$

Classification...

1. ***linear classifier*** h: (w, b) where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
2. classification of $x \in \mathbb{R}^d$ is $\mathrm{sign}(\langle w, x \rangle + b)$.
3. ***labeled*** example (x, y), h classifies (x, y) <u>correctly</u> if $\mathrm{sign}(\langle w, x \rangle + b) = y$.
4. Assume a linear classifier exists.
5. Given n labeled example. How to compute the linear classifier for these examples?
6. Use linear programming....
7. looking for $(\mathbf{w}, b)$, such that for an $(\mathbf{x}_i, y_i)$ we have $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = y_i$, which is

$$\begin{aligned} & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 0 && \text{if } y_i = 1, \\ \text{and} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq 0 && \text{if } y_i = -1. \end{aligned}$$

Classification...

1. ***linear classifier*** h: (w, b) where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
2. classification of $x \in \mathbb{R}^d$ is $\mathrm{sign}(\langle w, x \rangle + b)$.
3. ***labeled*** example (x, y), h classifies (x, y) <u>correctly</u> if $\mathrm{sign}(\langle w, x \rangle + b) = y$.
4. Assume a linear classifier exists.
5. Given n labeled example. How to compute the linear classifier for these examples?
6. Use linear programming....
7. looking for $(\mathbf{w}, b)$, such that for an $(\mathbf{x}_i, y_i)$ we have $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = y_i$, which is

$$\begin{aligned} & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 0 && \text{if } y_i = 1, \\ \text{and} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq 0 && \text{if } y_i = -1. \end{aligned}$$

Classification...

1. Or equivalently, let $\mathbf{x}_i = \left(x_i^1, \ldots, x_i^d\right) \in \mathbb{R}^d$, for $i = 1, \ldots, m$, and let $\mathbf{w} = \left(w^1, \ldots, w^d\right)$, then we get the linear constraint

$$\sum_{k=1}^{d} w^k x_i^k + b \geq 0 \qquad \text{if } y_i = 1,$$

$$\text{and} \qquad \sum_{k=1}^{d} w^k x_i^k + b \leq 0 \qquad \text{if } y_i = -1.$$

Thus, we get a set of linear constraints, one for each training example, and we need to solve the resulting linear program.

Linear programming for learning?

1. Stumbling block: is that linear programming is very sensitive to noise.
2. If points are misclassified $\implies$ no solution.
3. use an iterative algorithm that converges to the optimal solution if it exists...

Linear programming for learning?

1. Stumbling block: is that linear programming is very sensitive to noise.
2. If points are misclassified $\Longrightarrow$ no solution.
3. use an iterative algorithm that converges to the optimal solution if it exists...

Linear programming for learning?

1. Stumbling block: is that linear programming is very sensitive to noise.
2. If points are misclassified $\implies$ no solution.
3. use an iterative algorithm that converges to the optimal solution if it exists...

Perceptron algorithm...

```
perceptron(S: a set of l examples)
    w_0 ← 0, k ← 0
    R = max_{(x,y)∈S} ||x||.
    repeat
        for (x, y) ∈ S do
            if sign(⟨w_k, x⟩) ≠ y then
                w_{k+1} ← w_k + y * x
                k ← k + 1
    until no mistakes are made in the classification
    return w_k and k
```

Perceptron algorithm

1. Why **perceptron** algorithm converges?
2. Assume made a mistake on a sample $(\mathbf{x}, y)$ and $y = 1$. Then, $\langle \mathbf{w}_k, \mathbf{x} \rangle < 0$, and

$$\begin{aligned} \langle \mathbf{w}_{k+1}, \mathbf{x} \rangle &= \langle \mathbf{w}_k + y * \mathbf{x}, x \rangle = \langle \mathbf{w}_k, \mathbf{x} \rangle + y \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \langle \mathbf{w}_k, \mathbf{x} \rangle + y \|\mathbf{x}\| > \langle \mathbf{w}_k, \mathbf{x} \rangle . \end{aligned}$$

3. "walking" in the right direction..
4. ... new value assigned to $\mathbf{x}$ by $\mathbf{w}_{k+1}$ is larger ("more positive") than the old value assigned to $\mathbf{x}$ by $\mathbf{w}_k$.
5. After enough iterations of such fix-ups, label would change...

Perceptron algorithm

1. Why **perceptron** algorithm converges?
2. Assume made a mistake on a sample $(\mathbf{x}, y)$ and $y = 1$. Then, $\langle \mathbf{w}_k, \mathbf{x} \rangle < 0$, and

$$\begin{aligned} \langle \mathbf{w}_{k+1}, \mathbf{x} \rangle &= \langle \mathbf{w}_k + y * \mathbf{x}, x \rangle = \langle \mathbf{w}_k, \mathbf{x} \rangle + y \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \langle \mathbf{w}_k, \mathbf{x} \rangle + y \|\mathbf{x}\| > \langle \mathbf{w}_k, \mathbf{x} \rangle . \end{aligned}$$

3. "walking" in the right direction..
4. ... new value assigned to $\mathbf{x}$ by $\mathbf{w}_{k+1}$ is larger ("more positive") than the old value assigned to $\mathbf{x}$ by $\mathbf{w}_k$.
5. After enough iterations of such fix-ups, label would change...

Perceptron algorithm

1. Why **perceptron** algorithm converges?
2. Assume made a mistake on a sample $(\mathbf{x}, y)$ and $y = 1$. Then, $\langle \mathbf{w}_k, \mathbf{x} \rangle < 0$, and

$$\begin{aligned} \langle \mathbf{w}_{k+1}, \mathbf{x} \rangle &= \langle \mathbf{w}_k + y * \mathbf{x}, x \rangle = \langle \mathbf{w}_k, \mathbf{x} \rangle + y \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \langle \mathbf{w}_k, \mathbf{x} \rangle + y \|\mathbf{x}\| > \langle \mathbf{w}_k, \mathbf{x} \rangle . \end{aligned}$$

3. "walking" in the right direction..
4. ... new value assigned to $\mathbf{x}$ by $\mathbf{w}_{k+1}$ is larger ("more positive") than the old value assigned to $\mathbf{x}$ by $\mathbf{w}_k$.
5. After enough iterations of such fix-ups, label would change...

Perceptron algorithm

1. Why **perceptron** algorithm converges?
2. Assume made a mistake on a sample $(\mathbf{x}, y)$ and $y = 1$. Then, $\langle \mathbf{w}_k, \mathbf{x} \rangle < 0$, and

$$\begin{aligned} \langle \mathbf{w}_{k+1}, \mathbf{x} \rangle &= \langle \mathbf{w}_k + y * \mathbf{x}, x \rangle = \langle \mathbf{w}_k, \mathbf{x} \rangle + y \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \langle \mathbf{w}_k, \mathbf{x} \rangle + y \|\mathbf{x}\| > \langle \mathbf{w}_k, \mathbf{x} \rangle . \end{aligned}$$

3. "walking" in the right direction..
4. ... new value assigned to $\mathbf{x}$ by $\mathbf{w}_{k+1}$ is larger ("more positive") than the old value assigned to $\mathbf{x}$ by $\mathbf{w}_k$.
5. After enough iterations of such fix-ups, label would change...

Perceptron algorithm converges

Theorem

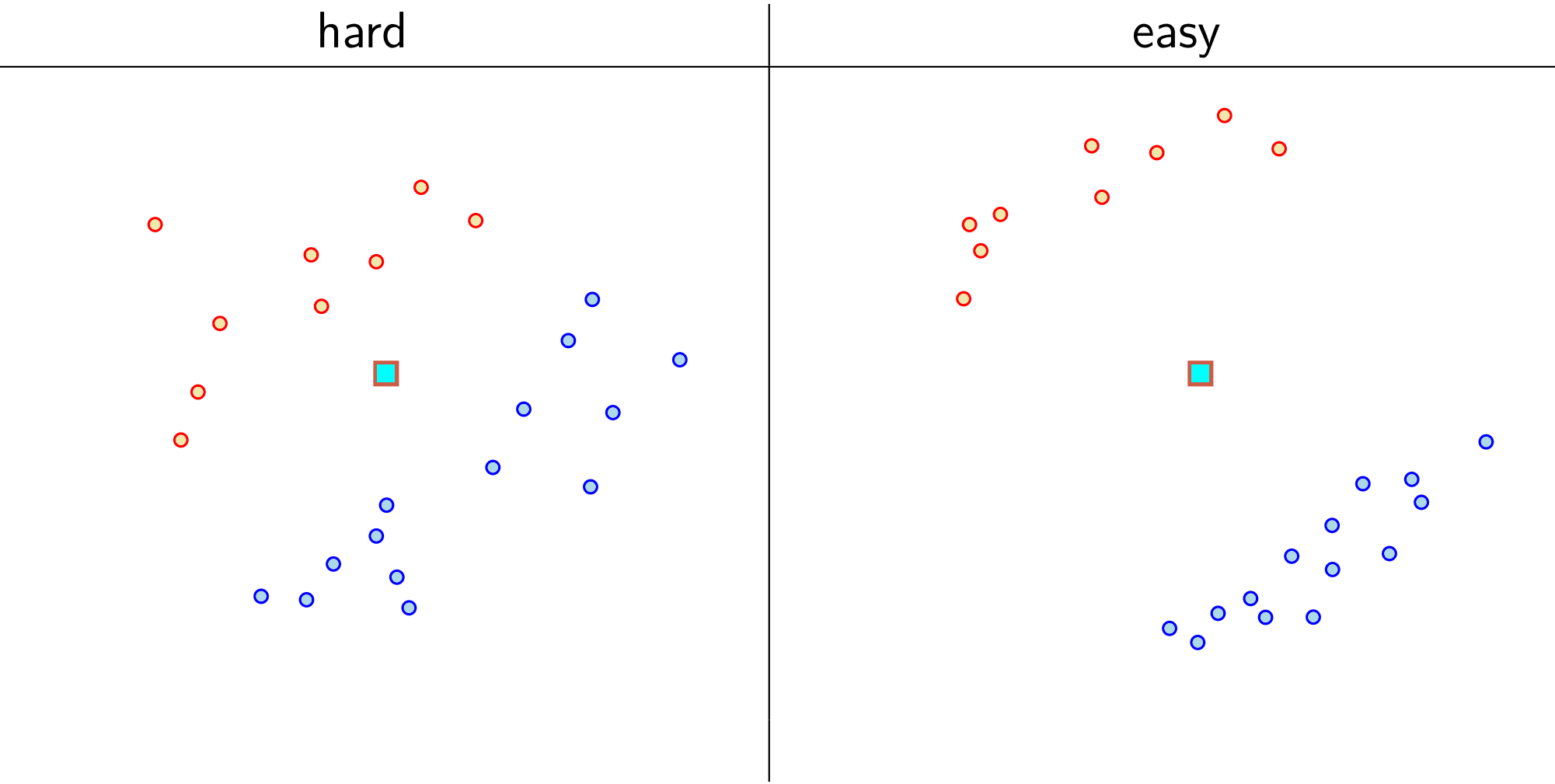
Let S *be a training set of examples, and let* $R = \max_{(x,y)\in S} \left\| x \right\|$. *Suppose that there exists a vector* w_{opt} *such that* $\left\| w_{opt} \right\| = 1$, *and a number* $\gamma > 0$, *such that*

$$y \langle w_{opt},\ x \rangle \geq \gamma \qquad \forall (x, y) \in S.$$

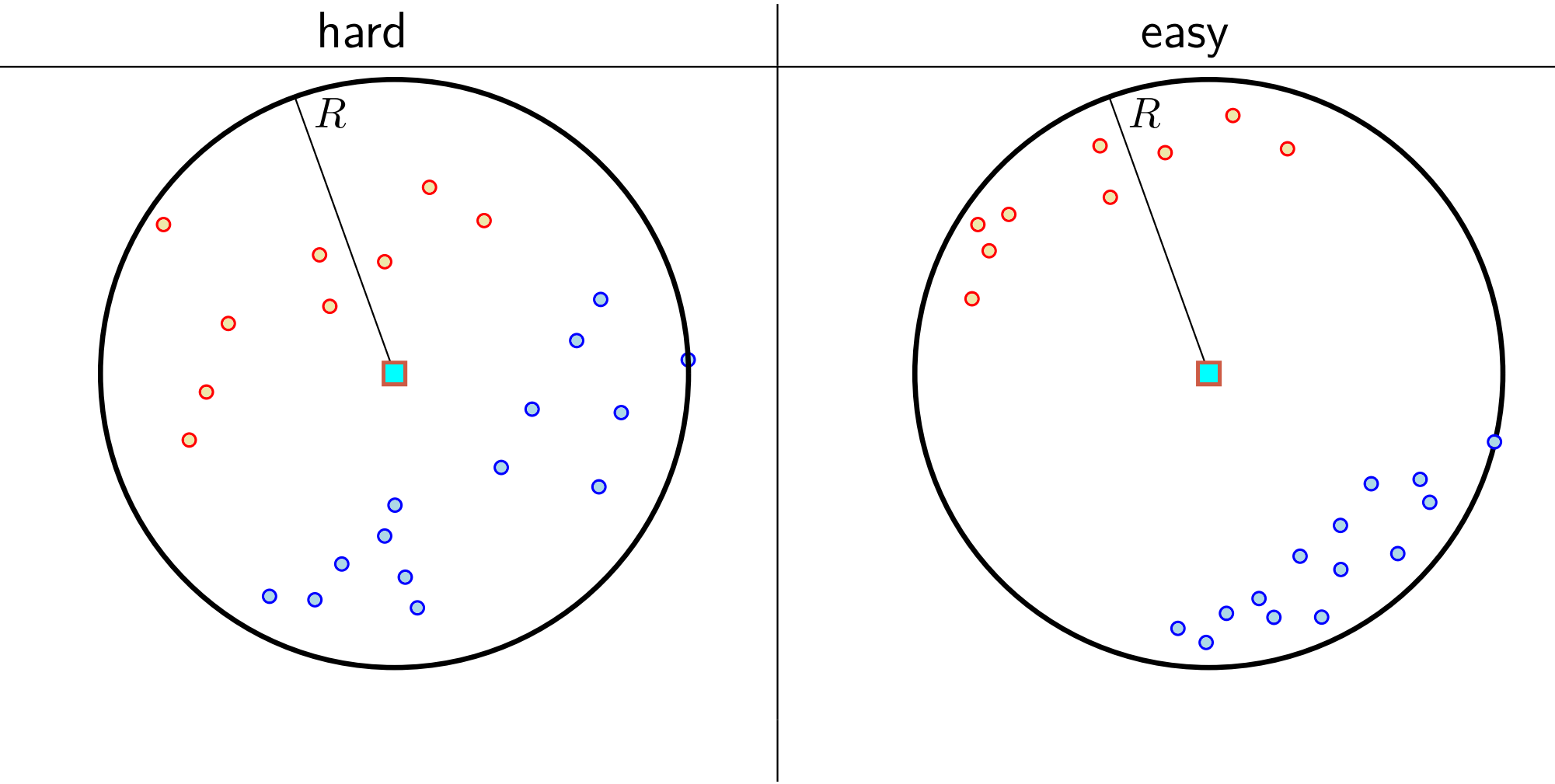
Then, the number of mistakes made by the online **perceptron** *algorithm on* S *is at most*

$$\left(\frac{R}{\gamma} \right)^2.$$

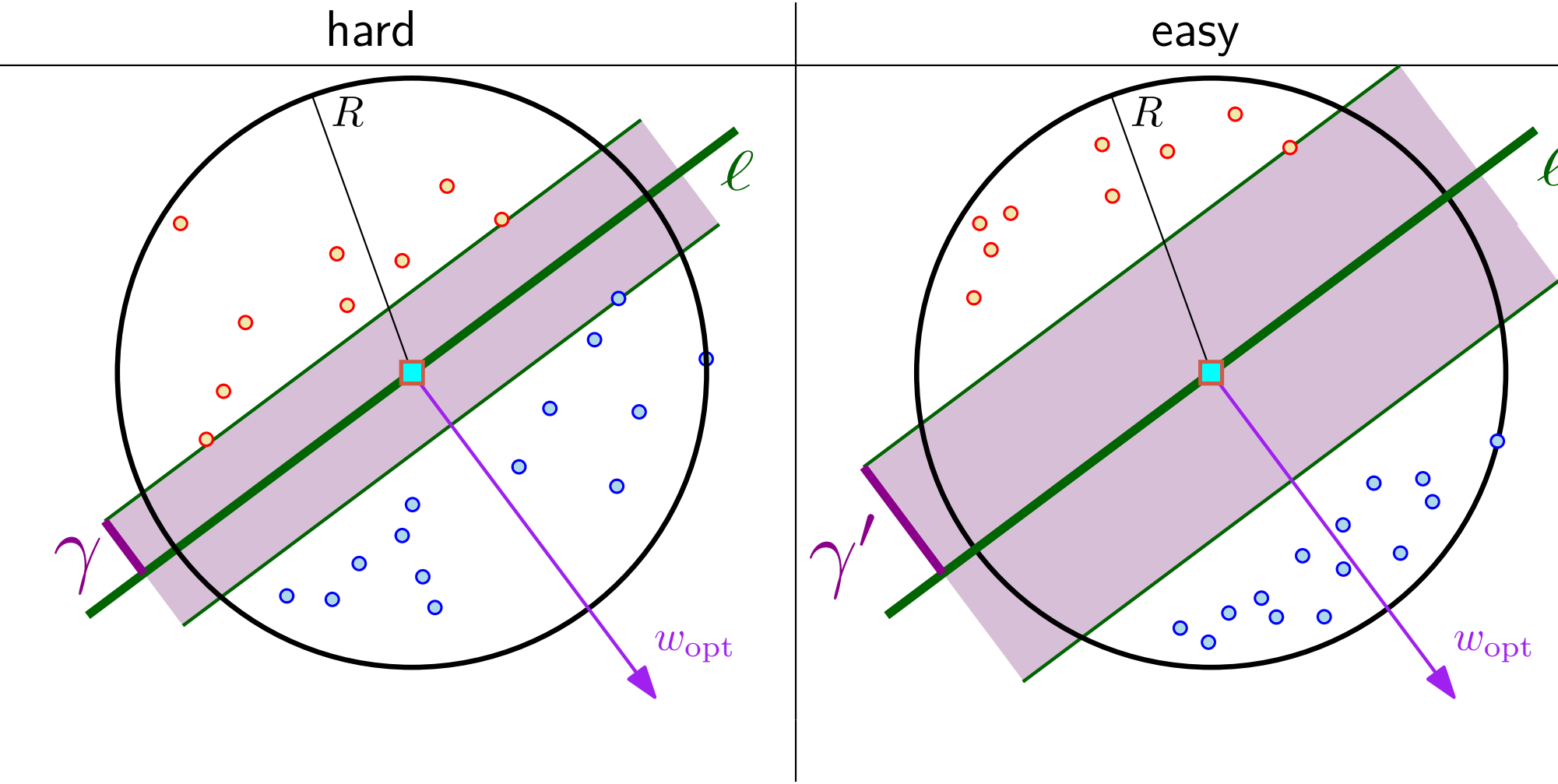
Claim by figure...



Claim by figure...

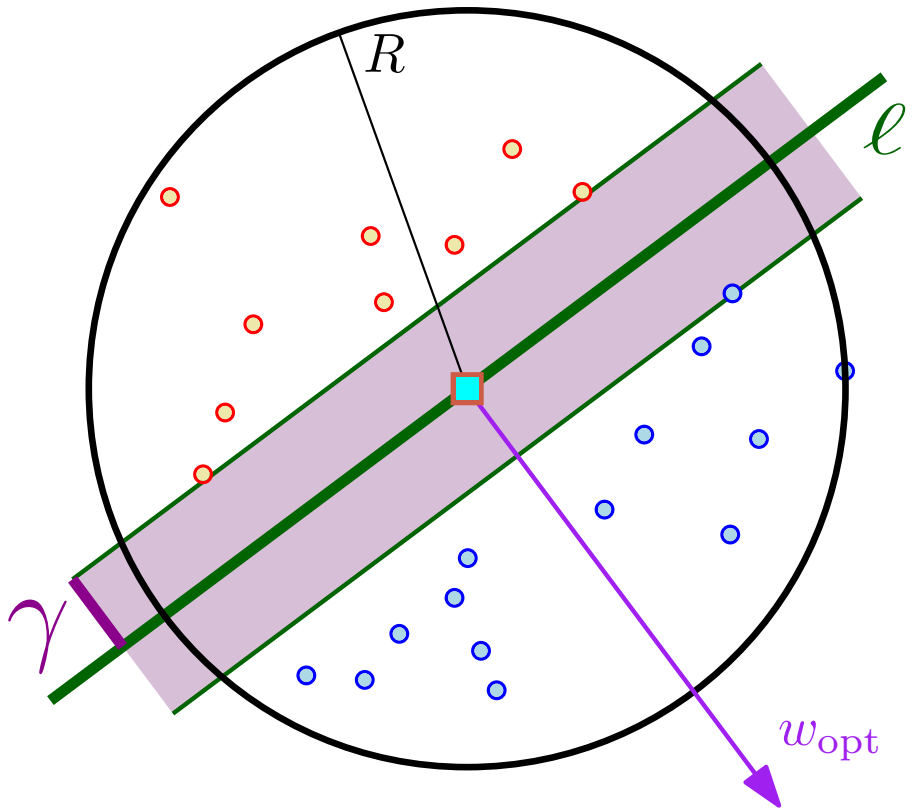


Claim by figure...

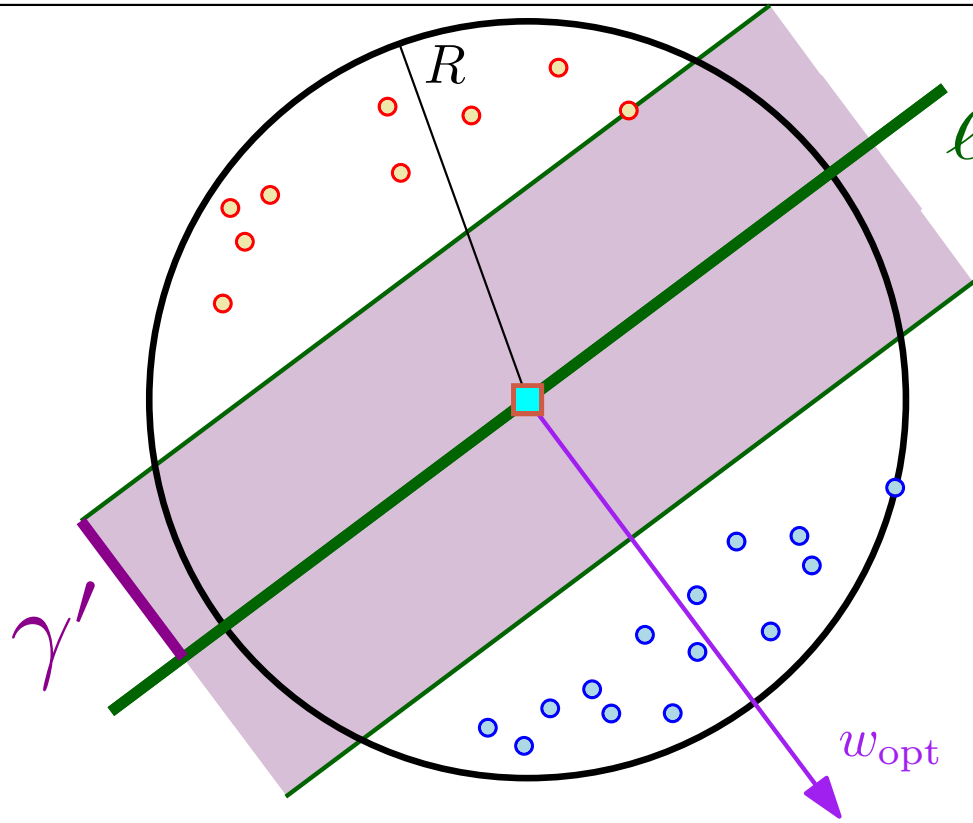


Claim by figure...

hard



easy



\# errors: $(R/\gamma)^2$

\# errors: $(R/\gamma')^2$

Proof of Perceptron convergence...

1. Idea of proof: **perceptron** weight vector converges to w_{opt}.
2. Distance between w_{opt} and kth update vector:

$$\alpha_k = \left\| w_k - \frac{R^2}{\gamma} w_{opt} \right\|^2 .$$

3. Quantify the change between α_k and α_{k+1}
4. Example being misclassified is (x, y).

Proof of Perceptron convergence...

1. Idea of proof: **perceptron** weight vector converges to w_{opt}.
2. Distance between w_{opt} and kth update vector:

$$\alpha_k = \left\| w_k - \frac{R^2}{\gamma} w_{opt} \right\|^2 .$$

3. Quantify the change between α_k and α_{k+1}
4. Example being misclassified is (x, y).

Proof of Perceptron convergence...

1. Idea of proof: **perceptron** weight vector converges to w_{opt}.
2. Distance between w_{opt} and kth update vector:

$$\alpha_k = \left\| w_k - \frac{R^2}{\gamma} w_{opt} \right\|^2 .$$

3. Quantify the change between α_k and α_{k+1}
4. Example being misclassified is (x, y).

Proof of Perceptron convergence...

1. Idea of proof: **perceptron** weight vector converges to w_{opt}.
2. Distance between w_{opt} and kth update vector:

$$\alpha_k = \left\| w_k - \frac{R^2}{\gamma} w_{opt} \right\|^2 .$$

3. Quantify the change between α_k and α_{k+1}
4. Example being misclassified is (x, y).

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $$\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$$
$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$
$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$
$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $$\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$$
$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$
$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$
$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$

$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$

$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$

$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$

$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$

$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$

$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$

$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$

$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$

$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$

$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$

$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$
$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$
$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. Example being misclassified is (x, y) (both are constants).
2. $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y * \mathbf{x}$
3. $\alpha_{k+1} = \left\| w_{k+1} - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \left\| w_k + y\mathbf{x} - \frac{R^2}{\gamma} w_{opt} \right\|^2$

$$= \left\| \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\|^2$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x}, \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) + y\mathbf{x} \right\rangle$$
$$= \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \left(w_k - \frac{R^2}{\gamma} w_{opt} \right) \right\rangle$$
$$+ 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), \mathbf{x} \right\rangle + \langle \mathbf{x}, \mathbf{x} \rangle$$
$$= \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \| x \|^2 .$$

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. (x, y) is misclassified: $\mathrm{sign}(\langle w_k, \mathrm{x} \rangle) \neq y$
3. $\implies \mathrm{sign}(y \langle \mathrm{w}_k, \mathrm{x} \rangle) = -1$
4. $\implies y \langle \mathrm{w}_k, \mathrm{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathrm{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathrm{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathrm{x} \rangle < 0$.

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. $(\mathbf{x}, y)$ is misclassified: $\mathbf{sign}(\langle w_k, \mathbf{x} \rangle) \neq y$
3. $\implies \mathbf{sign}(y \langle \mathbf{w}_k, \mathbf{x} \rangle) = -1$
4. $\implies y \langle \mathbf{w}_k, \mathbf{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathbf{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathbf{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathbf{x} \rangle < 0$.

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. (x, y) is misclassified: $\mathrm{sign}(\langle w_k, \mathrm{x} \rangle) \neq y$
3. $\implies \mathrm{sign}(y \langle \mathrm{w}_k, \mathrm{x} \rangle) = -1$
4. $\implies y \langle \mathrm{w}_k, \mathrm{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathrm{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathrm{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathrm{x} \rangle < 0$.

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. (x, y) is misclassified: $\mathrm{sign}(\langle w_k, \mathrm{x} \rangle) \neq y$
3. $\implies \mathrm{sign}(y \langle \mathrm{w}_k, \mathrm{x} \rangle) = -1$
4. $\implies y \langle \mathrm{w}_k, \mathrm{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathrm{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathrm{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathrm{x} \rangle < 0$.

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. (x, y) is misclassified: $\mathrm{sign}(\langle w_k, \mathrm{x} \rangle) \neq y$
3. $\implies \mathrm{sign}(y \langle \mathrm{w}_k, \mathrm{x} \rangle) = -1$
4. $\implies y \langle \mathrm{w}_k, \mathrm{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathrm{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathrm{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathrm{x} \rangle < 0$.

Proof of Perceptron convergence...

1. We proved: $\alpha_{k+1} = \alpha_k + 2y \left\langle \left(w_k - \frac{R^2}{\gamma} w_{opt} \right), x \right\rangle + \left\| x \right\|^2$.
2. (x, y) is misclassified: $\mathrm{sign}(\langle w_k, \mathrm{x} \rangle) \neq y$
3. $\implies \mathrm{sign}(y \langle \mathrm{w}_k, \mathrm{x} \rangle) = -1$
4. $\implies y \langle \mathrm{w}_k, \mathrm{x} \rangle < 0$.
5. $\left\| x \right\| \leq R \implies$

$$\begin{aligned} \alpha_{k+1} &\leq \alpha_k + R^2 + 2y \langle w_k, \mathrm{x} \rangle - 2y \left\langle \frac{R^2}{\gamma} w_{opt}, \mathrm{x} \right\rangle \\ &\leq \alpha_k + R^2 + \qquad -2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle . \end{aligned}$$

6. ... since $2y \langle w_k, \mathrm{x} \rangle < 0$.

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt}\ ,\ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2$.

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt}\ ,\ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2.$

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt} \ , \ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2.$

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt} \ , \ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$

$$\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$$
$$\leq \alpha_k + R^2 - 2R^2$$
$$\leq \alpha_k - R^2.$$

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt} \ , \ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2.$

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt} \ , \ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2.$

Proof of Perceptron convergence...

1. Proved: $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$.
2. $\mathrm{sign}(\langle w_{\mathrm{opt}}, x \rangle) = y$.
3. By margin assumption: $y \langle w_{opt} \ , \ x \rangle \geq \gamma, \quad \forall (x, y) \in S$.
4. $\alpha_{k+1} \leq \alpha_k + R^2 - 2\frac{R^2}{\gamma} y \langle w_{opt}, x \rangle$
   $\leq \alpha_k + R^2 - 2\frac{R^2}{\gamma}\gamma$
   $\leq \alpha_k + R^2 - 2R^2$
   $\leq \alpha_k - R^2.$

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}$.
3. $\forall i \qquad \alpha_i \geq 0$.
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}$. ∎

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}.$
3. $\forall i \qquad \alpha_i \geq 0.$
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}.$ ∎

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}.$
3. $\forall i \qquad \alpha_i \geq 0.$
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}.$ ∎

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}.$
3. $\forall i \qquad \alpha_i \geq 0.$
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}.$ ∎

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}.$
3. $\forall i \qquad \alpha_i \geq 0.$
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}.$ ∎

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}.$
3. $\forall i \qquad \alpha_i \geq 0.$
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}.$

Proof of Perceptron convergence...

1. We have: $\alpha_{k+1} \leq \alpha_k - R^2$
2. $\alpha_0 = \left\| 0 - \frac{R^2}{\gamma} w_{opt} \right\|^2 = \frac{R^4}{\gamma^2} \left\| w_{opt} \right\|^2 = \frac{R^4}{\gamma^2}$.
3. $\forall i \qquad \alpha_i \geq 0$.
4. Q: max # classification errors can make?
5. ... # of updates
6. .. # of updates $\leq \alpha_0 / R^2$...
7. A: $\leq \frac{R^2}{\gamma^2}$. ■

Concluding comment...

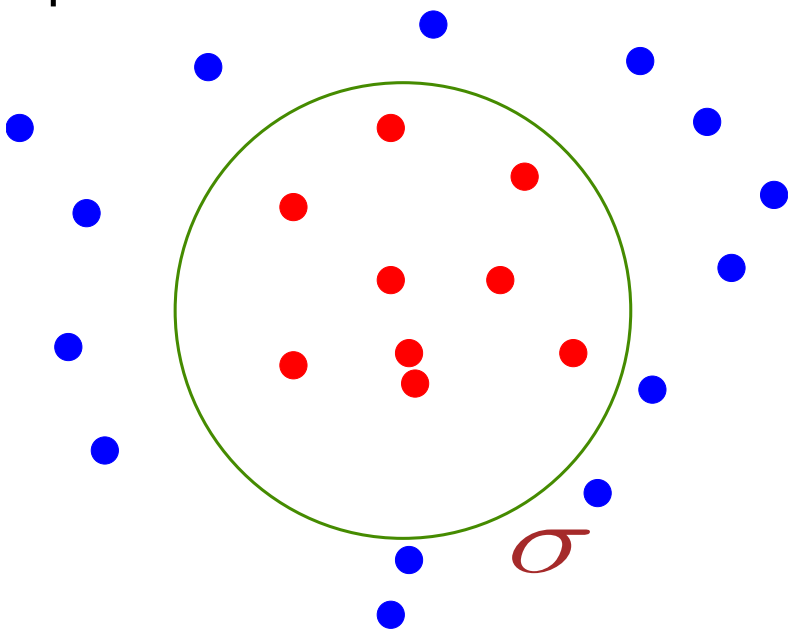
Any linear program can be written as the problem of separating red points from blue points. As such, the **perceptron** algorithm can be used to solve linear programs.

Learning a circle...

1. Given a set of red points, and blue points in the plane, we want to learn a circle that contains all the red points, and does not contain the blue points.
2. Q: How to compute the circle $\boldsymbol{\sigma}$?
3. ***Lifting***: $\ell : (x, y) \rightarrow (x,\ y,\ x^2 + y^2)$.
4. $z(P) = \left\{ \ell(x, y) = (x,\ y,\ x^2 + y^2) \ \middle| (x, y) \in P \right\}$

Learning a circle...

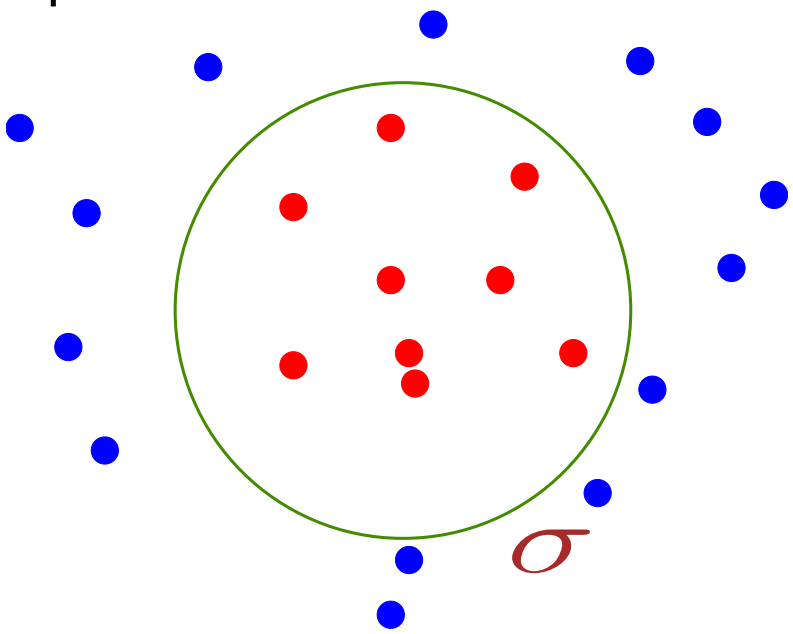
1. Given a set of red points, and blue points in the plane, we want to learn a circle that contains all the red points, and does not contain the blue points.



2. Q: How to compute the circle σ ?
3. ***Lifting***: $\ell : (x, y) \to (x,\ y,\ x^2 + y^2)$.
4. $z(P) = \left\{ \ell(x, y) = (x,\ y,\ x^2 + y^2) \ \middle|\ (x, y) \in P \right\}$

Learning a circle...

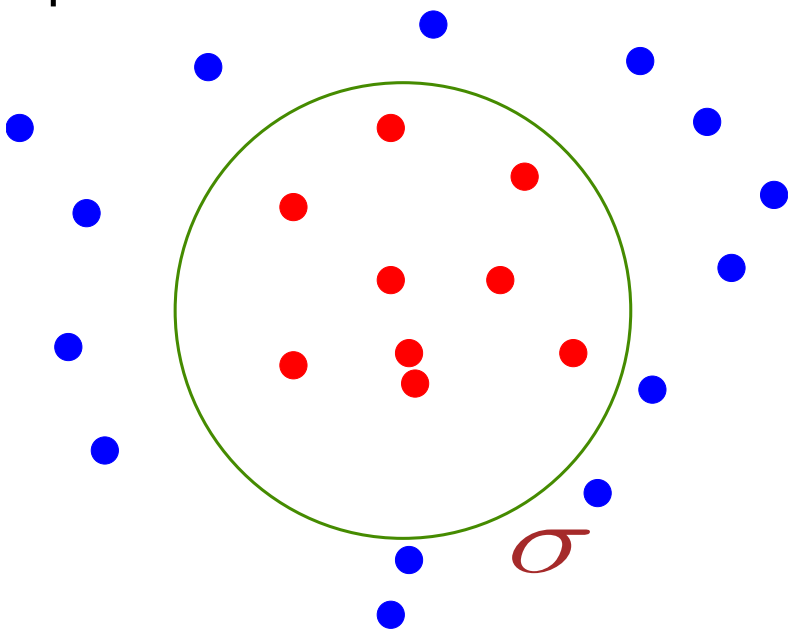
1. Given a set of red points, and blue points in the plane, we want to learn a circle that contains all the red points, and does not contain the blue points.



2. Q: How to compute the circle σ ?
3. *Lifting*: $\ell : (x, y) \to (x,\ y,\ x^2 + y^2)$.
4. $z(P) = \left\{ \ell(x, y) = (x,\ y,\ x^2 + y^2) \ \middle|\ (x, y) \in P \right\}$

Learning a circle...

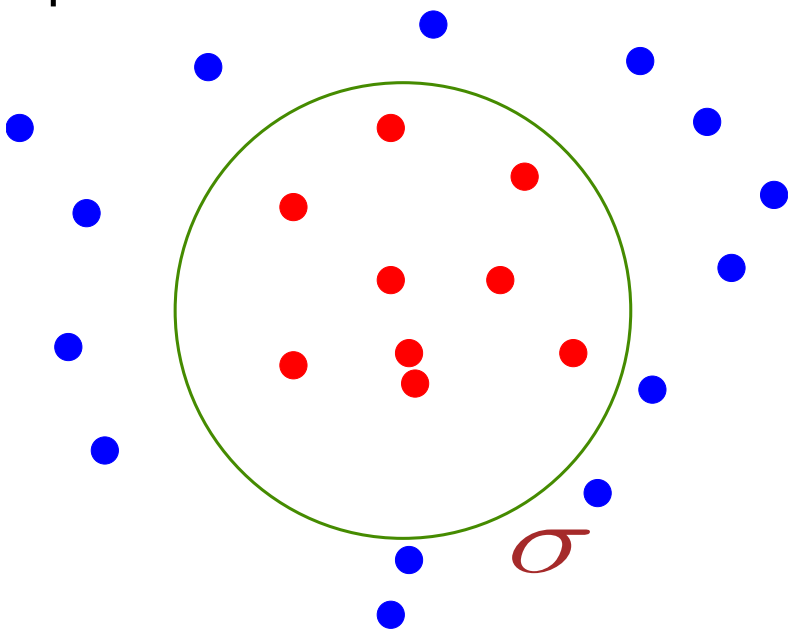
1. Given a set of red points, and blue points in the plane, we want to learn a circle that contains all the red points, and does not contain the blue points.



2. Q: How to compute the circle σ ?
3. ***Lifting***: $\ell : (x, y) \to (x,\ y,\ x^2 + y^2)$.
4. $z(P) = \left\{ \ell(x, y) = (x,\ y,\ x^2 + y^2) \;\middle|\; (x, y) \in P \right\}$

Learning a circle...

1. Given a set of red points, and blue points in the plane, we want to learn a circle that contains all the red points, and does not contain the blue points.



2. Q: How to compute the circle σ ?
3. ***Lifting***: $\ell : (x, y) \to (x,\ y,\ x^2 + y^2)$.
4. $z(P) = \left\{ \ell(x, y) = (x,\ y,\ x^2 + y^2) \ \middle|\ (x, y) \in P \right\}$

Learning a circle...

Theorem

Two sets of points $\boldsymbol{R}$ and $\boldsymbol{B}$ are separable by a circle in two dimensions, if and only if $\ell(\boldsymbol{R})$ and $\ell(\boldsymbol{B})$ are separable by a plane in three dimensions.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \quad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \quad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof

1. $\sigma \equiv (x-a)^2 + (y-b)^2 = r^2$: circle containing R, and all points of B outside.
2. $\forall (x,y) \in R \qquad (x-a)^2 + (y-b)^2 \leq r^2$
   $\forall (x,y) \in B \qquad (x-a)^2 + (y-b)^2 > r^2$.
3. $\forall (x,y) \in R \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 \leq 0$.
   $\forall (x,y) \in B \quad -2ax - 2by + (x^2+y^2) - r^2 + a^2 + b^2 > 0$.
4. Setting $z = z(x,y) = x^2 + y^2$:
   $h(x,y,z) = -2ax - 2by + z - r^2 + a^2 + b^2$
   $\forall (x,y) \in R \quad h(x,y,z(x,y)) \leq 0$
5. $\iff \forall (x,y) \in R \quad h(\ell(x,y)) \leq 0$
   $\qquad \forall (x,y) \in B \quad h(\ell(x,y)) > 0$
6. $p \in \sigma \iff h(\ell(p)) \leq 0$.
7. Proved: if point set is separable by a circle $\implies$ lifted point set $\ell(R)$ and $\ell(B)$ are separable by a plane.

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d$.
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d$.
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ∎

Proof: Other direction

1. Assume $\ell(R)$ and $\ell(B)$ are linearly separable. Let separating place be:

$$h \equiv ax + by + cz + d = 0$$

2. $\forall (x, y, x^2 + y^2) \in \ell(R)$: $ax + by + c(x^2 + y^2) + d \leq 0$
3. $\forall (x, y, x^2 + y^2) \in \ell(B)$: $ax + by + c(x^2 + y^2) + d \geq 0$.
4. $U(h) = \left\{ (x, y) \;\middle|\; h((x, y, x^2 + y^2)) \leq 0 \right\}$.
5. If $U(h)$ is a circle $\implies$ $R \subset U(h)$ and $B \cap U(h) = \emptyset$.
6. $U(h) \equiv ax + by + c(x^2 + y^2) \leq -d.$
7. $\iff \left(x^2 + \frac{a}{c}x\right) + \left(y^2 + \frac{b}{c}y\right) \leq -\frac{d}{c}$
8. $\iff \left(x + \frac{a}{2c}\right)^2 + \left(y + \frac{b}{2c}\right)^2 \leq \frac{a^2+b^2}{4c^2} - \frac{d}{c}$
9. This is disk in the plane, as claimed. ■

A closing comment...

Linear separability is a powerful technique that can be used to learn complicated concepts that are considerably more complicated than just hyperplane separation. This lifting technique showed above is the ***kernel technique*** or ***linearization***.

A Little Bit On VC Dimension

1. Q: how complex is the function trying to learn?
2. VC-dimension is one way of capturing this notion. (VC = Vapnik, Chervonenkis,1971).
3. A matter of expressivity: What is harder to learn:
    1. A rectangle in the plane.
    2. A halfplane.
    3. A convex polygon with k sides.

A Little Bit On VC Dimension

1. Q: how complex is the function trying to learn?
2. VC-dimension is one way of capturing this notion. (VC = Vapnik, Chervonenkis,1971).
3. A matter of expressivity: What is harder to learn:
    1. A rectangle in the plane.
    2. A halfplane.
    3. A convex polygon with k sides.

A Little Bit On VC Dimension

1. Q: how complex is the function trying to learn?
2. VC-dimension is one way of capturing this notion. (VC = Vapnik, Chervonenkis,1971).
3. A matter of expressivity: What is harder to learn:
   1. A rectangle in the plane.
   2. A halfplane.
   3. A convex polygon with k sides.

Thinking about concepts as binary functions...

1. $X = \{p_1, p_2, \ldots, p_m\}$: points in the plane.
2. $\mathcal{H}$: set of all halfplanes.
3. A half-plane $r \in \mathcal{H}$ defines a binary vector
$$r(X) = (b_1, \ldots, b_m)$$
where $b_i = 1$ if and only if p_i is inside r.
4. Possible binary vectors generated by halfplanes:
$U(X, \mathcal{H}) = \{r(X) \mid r \in \mathcal{H}\}$.
5. A set X of m elements is ***shattered*** by R if
$$|U(X, R)| = 2^m.$$
6. What does this mean?
7. The ***VC-dimension*** of a set of ranges R is the size of the largest set that it can shatter.

Thinking about concepts as binary functions...

1. $X = \{p_1, p_2, \ldots, p_m\}$: points in the plane.
2. $\mathcal{H}$: set of all halfplanes.
3. A half-plane $r \in \mathcal{H}$ defines a binary vector
$$r(X) = (b_1, \ldots, b_m)$$
where $b_i = 1$ if and only if p_i is inside r.
4. Possible binary vectors generated by halfplanes:
$U(X, \mathcal{H}) = \{r(X) \mid r \in \mathcal{H}\}$.
5. A set X of m elements is ***shattered*** by R if
$$|U(X, R)| = 2^m.$$
6. What does this mean?
7. The ***VC-dimension*** of a set of ranges R is the size of the largest set that it can shatter.

Thinking about concepts as binary functions...

1. $X = \{p_1, p_2, \ldots, p_m\}$: points in the plane.
2. $\mathcal{H}$: set of all halfplanes.
3. A half-plane $r \in \mathcal{H}$ defines a binary vector
$$r(X) = (b_1, \ldots, b_m)$$
where $b_i = 1$ if and only if p_i is inside r.
4. Possible binary vectors generated by halfplanes:
$U(X, \mathcal{H}) = \{r(X) \mid r \in \mathcal{H}\}$.
5. A set X of m elements is ***shattered*** by R if
$$|U(X, R)| = 2^m.$$
6. What does this mean?
7. The ***VC-dimension*** of a set of ranges R is the size of the largest set that it can shatter.

Thinking about concepts as binary functions...

1. $X = \{p_1, p_2, \ldots, p_m\}$: points in the plane.
2. $\mathcal{H}$: set of all halfplanes.
3. A half-plane $r \in \mathcal{H}$ defines a binary vector
$$r(X) = (b_1, \ldots, b_m)$$
where $b_i = 1$ if and only if p_i is inside r.
4. Possible binary vectors generated by halfplanes:
$U(X, \mathcal{H}) = \{r(X) \mid r \in \mathcal{H}\}$.
5. A set X of m elements is ***shattered*** by R if
$$|U(X, R)| = 2^m.$$
6. What does this mean?
7. The ***VC-dimension*** of a set of ranges R is the size of the largest set that it can shatter.

Thinking about concepts as binary functions...

1. $X = \{p_1, p_2, \ldots, p_m\}$: points in the plane.
2. $\mathcal{H}$: set of all halfplanes.
3. A half-plane $r \in \mathcal{H}$ defines a binary vector
$$r(X) = (b_1, \ldots, b_m)$$
where $b_i = 1$ if and only if p_i is inside r.
4. Possible binary vectors generated by halfplanes:
$U(X, \mathcal{H}) = \{r(X) \mid r \in \mathcal{H}\}$.
5. A set X of m elements is ***shattered*** by R if
$$|U(X, R)| = 2^m.$$
6. What does this mean?
7. The ***VC-dimension*** of a set of ranges R is the size of the largest set that it can shatter.

Examples

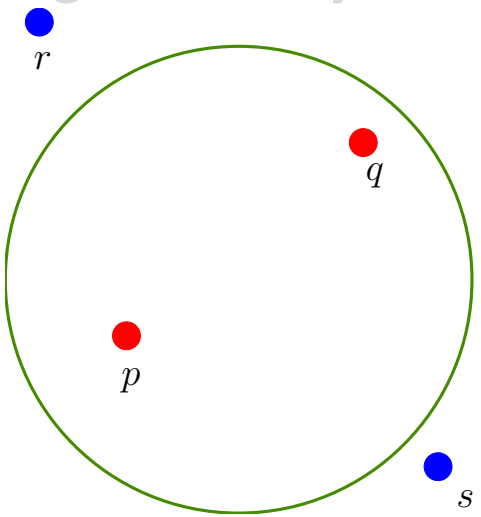
What is the VC dimensions of circles in the plane?

X is set of n points in the plane

$\mathcal{C}$ is a set of all circles.

$X = \{p, q, r, s\}$

What subsets of X can we generate by circle?



Examples

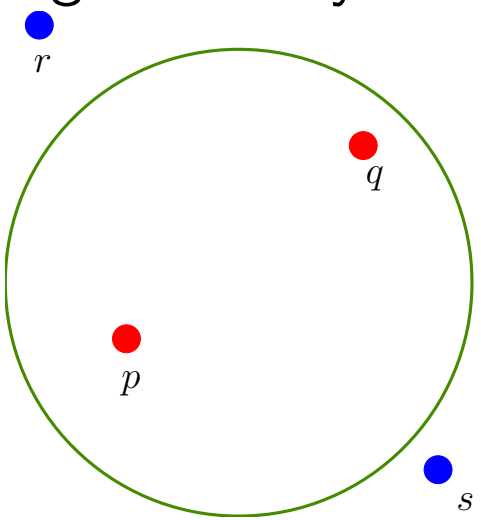
What is the VC dimensions of circles in the plane?

X is set of n points in the plane

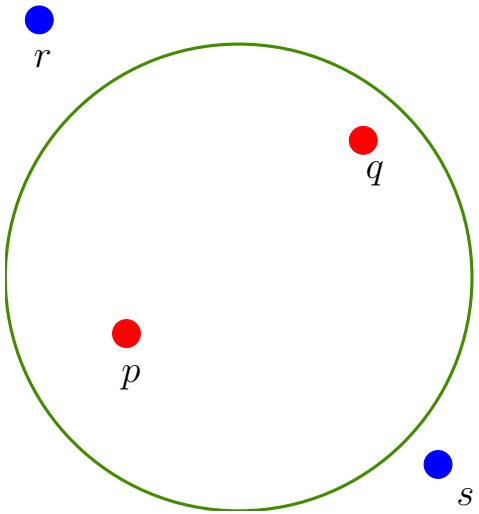
$\mathcal{C}$ is a set of all circles.

$X = \{p, q, r, s\}$

What subsets of X can we generate by circle?



Subsets realized by disks

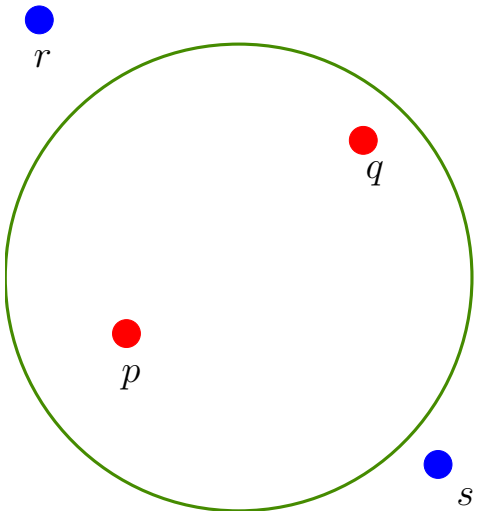


$\{\}, \{r\}, \{p\}, \{q\}, \{s\}, \{p, s\}$, $\{p, q\}$, $\{p, r\}, \{r, q\}\{q, s\}$ and $\{r, p, q\}$, $\{p, r, s\}\{p, s, q\}, \{s, q, r\}$ and $\{r, p, q, s\}$

We got only 15 sets. There is one set which is not there. Which one?

The VC dimension of circles in the plane is 3.

Subsets realized by disks

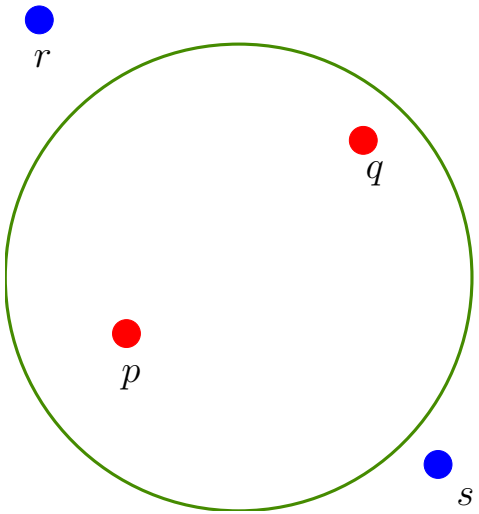


$\{\}, \{r\}, \{p\}, \{q\}, \{s\}, \{p, s\}$, $\{p, q\}$, $\{p, r\}$, $\{r, q\}\{q, s\}$ and $\{r, p, q\}$, $\{p, r, s\}\{p, s, q\}$, $\{s, q, r\}$ and $\{r, p, q, s\}$

We got only 15 sets. There is one set which is not there. Which one?

The VC dimension of circles in the plane is 3.

Subsets realized by disks



$\{\}, \{r\}, \{p\}, \{q\}, \{s\}, \{p, s\}$, $\{p, q\}$, $\{p, r\}$, $\{r, q\}\{q, s\}$ and $\{r, p, q\}$, $\{p, r, s\}\{p, s, q\}$, $\{s, q, r\}$ and $\{r, p, q, s\}$

We got only 15 sets. There is one set which is not there. Which one?

The VC dimension of circles in the plane is 3.

Sauer's Lemma

Lemma (Sauer Lemma)

If $\boldsymbol{R}$ *has VC dimension* $\boldsymbol{d}$ *then* $|\boldsymbol{U}(\boldsymbol{X}, \boldsymbol{R})| = \boldsymbol{O}\left(\boldsymbol{m}^{\boldsymbol{d}}\right)$, *where* $\boldsymbol{m}$ *is the size of* $\boldsymbol{X}$.

Notes

Notes

Notes

Notes