① Perceptron Convergence Theorem (PCT)

② Geometric Margin (γ_{geom})

Learning linear classifiers through the origin

all linear classifiers

$\mathcal{F} = \{ f : \mathbb{R}^d \to \{-1, +1\},\ f_\theta(\underline{x}) = \text{sgn}(\langle \underline{\theta}, \underline{x} \rangle) \}$

↑ feature space ↑ for some $\underline{\theta} \in \mathbb{R}^d$

θ → parametrize the function function $f \Rightarrow f_\theta$ or $f(\underline{x}; \underline{\theta})$

↑ learn!

We have a labelled dataset $\mathcal{D} = \{ (\underline{x}_t, y_t) \in \mathbb{R}^d \times \{+1, -1\} : 1 \le t \le n \}$.

Perceptron Algorithm: 1) $k = 0$ initialize $\underline{\theta}^{(k)}$ arbitrarily

(for the proof, set $\underline{\theta}^{(0)} = \underline{0}$)

2) cycle through the dataset $\mathcal{D}$

If $y_t \langle \underline{x}_t, \underline{\theta}^{(k)} \rangle \le 0$, then

Update $\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} + y_t \underline{x}_t$

If $y_t \langle \underline{x}_t, \underline{\theta}^{(k)} \rangle > 0$, do nothing!

Online Algorithm ⟵

⇵

we don't need to cycle through the whole dataset when we have new data!

3) Out!

The disadvantage of Perceptron ⇒ only find one line to separate the 2 class!

↓

not good!

① cannot find a "good" line!

② not unique → depends on the sequence of data

→ finite dataset

Assumptions: ① (R-bounded) → finite dataset / → infinite (countable) dataset

② γ-linearly separable $\Rightarrow \exists \underline{\theta}^* \in \mathbb{R}^d,\ \exists \gamma > 0$

$$y_t \langle \underline{\theta}^*, \underline{x}_t \rangle \geq \gamma \quad \forall t = 1, 2, \dots, n$$

$\rightarrow$ γ is defined by

$$\gamma = \min_t y_t \langle \underline{\theta}^*, \underline{x}_t \rangle$$

Thm: [Novikoff (1962)] Start the perceptron alg. at $\underline{\theta}^{(0)} = \underline{0}$

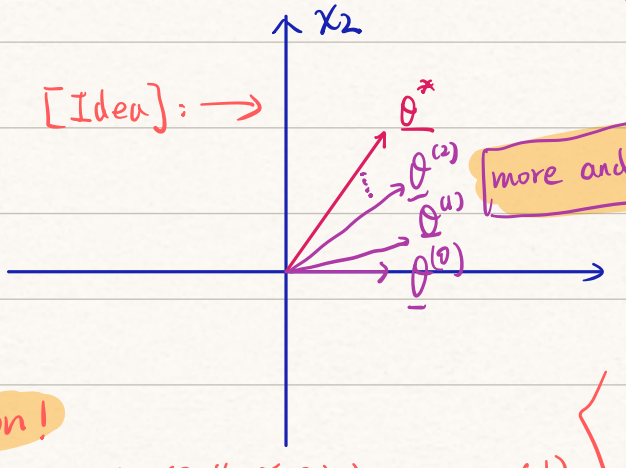
then maximum # of update required

s.t $\{\underline{x}_t\}_{t=1}^n$ are classified correctly k.

satisfies $k \leq R^2 \cdot \boxed{\frac{\|\underline{\theta}^*\|^2}{\gamma^2}}$ Not so good!

→ Many choice of $\underline{\theta}^*$

equality can be achieved?



[Idea]: → more and more "aligned" with $\underline{\theta}^*$

How to measure?

1) $\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle = 0 \Rightarrow$ orthogonal

$\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle$ "large" $\Rightarrow$ more aligned!

2) this "large" require to control $\|\underline{\theta}^{(k)}\|$

3) lower bound $\dfrac{\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle}{\|\underline{\theta}^{(k)}\| \, \|\underline{\theta}^*\|}$

Motivation!

中文意义

$\dfrac{\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle}{\|\underline{\theta}^{(k)}\| \, \|\underline{\theta}^*\|} \leq A$

分子：增长很快（关于k）

分母：增长比较慢（关于k）

$\Rightarrow k$不能太大！

want to show:

1) $\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle$ increases at least linearly (fast) in k

2) $\|\underline{\theta}^{(k)}\|^2$ increases at most linearly fast in k

3) Cauchy-Schwarz.

Pf: Part I:

update whenever make error

$$\Rightarrow \langle \underline{\theta}^*, \underline{\theta}^{(k)} \rangle = \langle \underline{\theta}^*, \underline{\theta}^{(k-1)} + y_t \underline{x}_t \rangle$$

$$= \langle \underline{\theta}^*, \underline{\theta}^{(k-1)} \rangle + y_t \langle \underline{\theta}^*, \underline{x}_t \rangle$$

$\geq \gamma$ - linealy separable.

$$\geqslant \langle \underline{\theta}^*, \underline{\theta}^{(k-1)} \rangle + \gamma .$$

since $\underline{\theta}^{(0)} = \underline{0}$, $\langle \underline{\theta}^*, \underline{\theta}^{(k)} \rangle \geqslant k\gamma$.

↑ "unnormalized alignment"

Part II:

$$\|\underline{\theta}^{(k)}\|^2 = \langle \underline{\theta}^{(k-1)} + y_t \underline{x}_t, \underline{\theta}^{(k-1)} + y_t \underline{x}_t \rangle$$

$$= \|\underline{\theta}^{(k-1)}\|^2 + \|\underline{x}_t\|^2 + 2 y_t \langle \underline{x}_t, \underline{\theta}^{(k-1)} \rangle$$

$$\leqslant \|\theta^{(k-1)}\|^2 + R^2 .$$

($\mathcal{D}$ is R-bounded) + (Update Property)

$$\leqslant kR^2$$

Part III

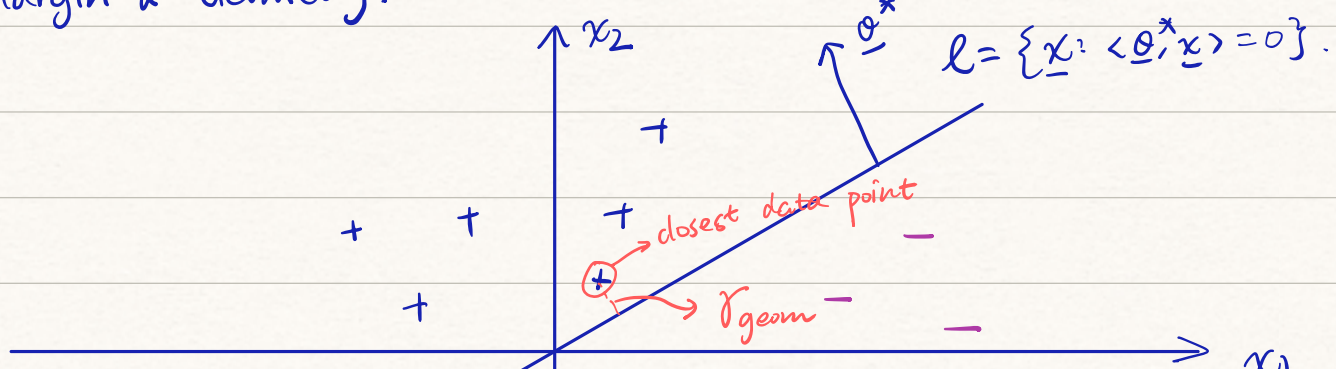
$$\frac{\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle}{\|\underline{\theta}^{(k)}\| \|\underline{\theta}^*\|} \leq 1 \Rightarrow 1 \geqslant \frac{\langle \underline{\theta}^{(k)}, \underline{\theta}^* \rangle}{\|\underline{\theta}^{(k)}\| \|\underline{\theta}^*\|} \geqslant \frac{k\gamma}{\sqrt{kR^2}\|\underline{\theta}^*\|}$$

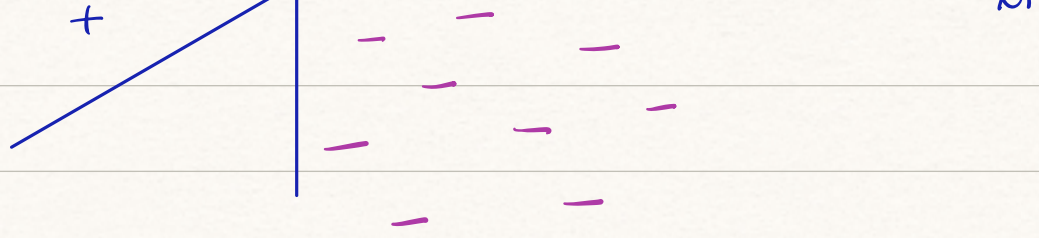
{ PART I, PART II.

$$\Rightarrow k \leqslant R^2 \frac{\|\underline{\theta}^*\|^2}{\gamma^2}$$

□.

Margin & Geometry.





Claim: $\gamma_{geom} = \frac{\gamma}{\|\theta^*\|}$ is the **smallest distance** from any training sample to the decision boundary ℓ.

perceptron alg. $k \le R^2 \frac{\|\theta^*\|}{\gamma^2} = \frac{R^2}{\gamma_{geom}^2}$

It shows that: the closer those training samples are to the decision boundary, the worst the upper bound of k is!

(γ_{geom} is smaller in this sense)

Rmk: (If dataset is normalized) $\Rightarrow k \le \frac{1}{\gamma_{geom}^2}$

$\Updownarrow$

$R=1$

So γ_{geom} represents the difficulty of classification problem!

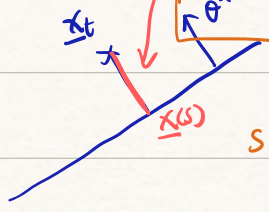
Target: ↓ → Actually we just prove when $\underline{x}_t$ attains γ. its distance is $\frac{\gamma}{\|\theta^*\|} = \gamma_{geom}$

Pf: To calculate the distance of closest point $\underline{x}_t$ to the decision boundary $\ell = \{x : \langle \underline{\theta}^*, \underline{x} \rangle = 0\}$.

Note that $\underline{x}_t$ satisfies $y_t \langle \underline{x}_t, \underline{\theta}^* \rangle = \gamma$.

Method → define a line segment $\underline{x}(t)$ with $\underline{x}(0) = \underline{x}_t$ parallel to $\underline{\theta}^*$ towards the boundary.

$$\underline{x}(s) = \underline{x}(0) - s \cdot y_t \cdot \frac{\theta^*}{\|\theta^*\|} \Rightarrow \langle \underline{x}(s), \underline{\theta}^* \rangle = 0 \text{ to solve } s.$$



s: length of segment from $\underline{x}(s)$ to $\underline{x}(0)$

$$\langle \underline{x}_t - s \cdot y_t \frac{\theta^*}{\|\theta^*\|}, \theta^* \rangle = 0$$

$$\rightarrow \langle \underline{x}_t, \underline{\theta}^* \rangle = s \cdot y_t \|\theta^*\|$$

$$\rightarrow s = \frac{y_t \langle \underline{x}_t, \underline{\theta}^* \rangle}{\|\theta^*\|} = \frac{\gamma}{\|\theta^*\|}$$

Another Pf:

$$s = \left| \left\langle x_t, \frac{\theta^*}{\|\theta^*\|} \right\rangle \right|$$

$$= \frac{1}{\|\theta^*\|} \left| \langle x_t, \theta^* \rangle \right|$$

$$= \frac{1}{\|\theta^*\|} \cdot y_t \langle x_t, \theta^* \rangle$$

$$= \frac{\gamma}{\|\theta^*\|}$$

#

↳ the distance of x_t to ℓ.

#

⇓ for the definition

$$x(t) = x(0) - s \boxed{y_t} \frac{\theta^*}{\|\theta^*\|}$$

↙ to guarantee that $s > 0$.

Rmk: i). Bound on # of updates of perceptron alg. is

$$k \le \left(\frac{R}{\gamma_{geom}} \right)^2 - (*)$$

ii) $\frac{R}{\gamma_{geom}}$ can be viewed as the difficulty of learning linear classification

→ related to VC dimension

iii) $x_t \in \mathbb{R}^d$

The bound $(*)$ is almost independent of d.

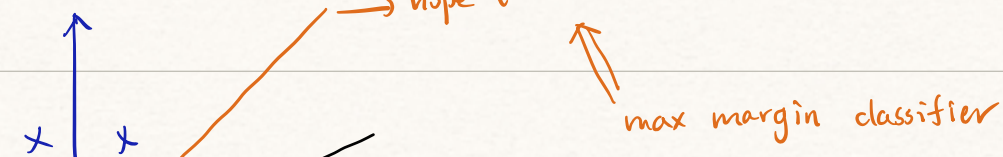
$R = \max_i \|x_i\|$ ↳ increase with dimension d! (R gets larger)

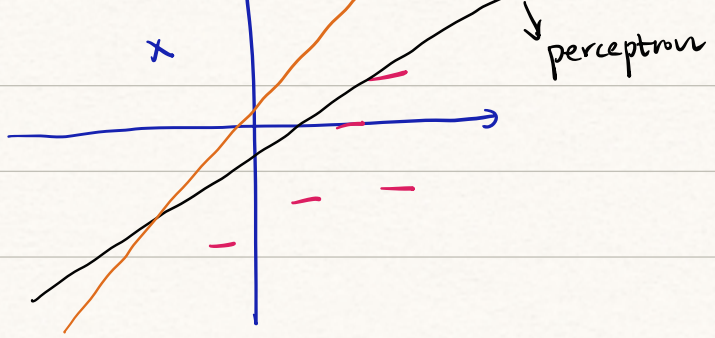
iv) the bound $(*)$ is almost independent of n.

Recall: $\gamma = \min_{1 \le t \le n} y_t \langle x_t, \theta^* \rangle$. ↓ when $n \uparrow$

$\Rightarrow \gamma_{geom} \downarrow$

Variants





Standard perceptron : Update if $y_t \langle x_t, \theta^{(k)} \rangle \le 0 \Leftrightarrow$ mis label

↳ leads to not robust

① variation

margin perceptron alg: Fix $\beta \in (0, \gamma)$ e.s. $\beta = \frac{\gamma}{2}$

Update if ~~$y_t \langle x_t, \theta^{(k)} \rangle \le \beta$~~

① $0 \frac{\beta}{2} \le \frac{\langle x_t, \theta^{(k)} \rangle}{\|\theta^{(k)}\|} \le \frac{\beta}{2}$

② $y_t \langle x_t, \theta^{(k)} \rangle \le 0$

Theoretical ⇒ cannot be programmed

Since we don't know β (γ).

Assumptions: $\|\theta^*\| = 1$ $R = 1$ $\boxed{\beta = \frac{\gamma}{2}}$ → $(\beta = (1-\varepsilon)\gamma)$

prove that $K \le \frac{\boxed{12}}{\gamma^2}$ → $f(\varepsilon)$ (Can be improved)

② **Batch Perceptron Alg:**

合并成 $y_t \frac{\langle x_t, \theta^{(k)} \rangle}{\|\theta^{(k)}\|} \le \frac{\beta}{2}$

Update: $y_k := \{ t = 1, 2, \dots, n : y_t \langle x_t, \theta^{(k)} \rangle \le 0 \}$ → All the wrong samples

$$\theta^{(k+1)} = \theta^{(k)} + \eta_k \sum_{t \in y_k} y_t x_t \quad -(*)$$

Trajectory of $\{\theta^{(k)}\}_{k \in \mathbb{N}}$ is "smoother". (in some sense)

Run $(*)$ until $y_k = \phi$.

Thm: (i) $\eta_k \ge 0$, (ii) $\lim_{m \to \infty} \sum_{k=1}^{m} \eta_k = \infty$, (iii) $\lim_{m \to \infty} \frac{\sum_{k=1}^{m} \eta_k^2}{(\sum_{k=1}^{m} \eta_k)^2} = 0$,

E.s. $\eta_k = \frac{1}{k}$.

then Batch Perceptron Alg. converges!

Outline:

① Perceptron (linear classifier)

⇓

PCT → $k \le R^2 \cdot \frac{\|\theta^*\|^2}{\gamma^2} := R^2 \cdot \frac{1}{\gamma_{geom}}$ and $\gamma_{geom} = \frac{\gamma}{\|\theta^*\|}$

Geometric Meaning: the closest distance from every data set to the decision Boundary.

② Variants { Margin Perceptron Alg.
Batch Perceptron Alg.

need { ① $\|x_t\| = 1$
② $\gamma \to \|\theta^*\| = 1$

i) Margin Perceptron [change the Update]

{ Perceptron: Update if $y_t \langle x_t, \theta^{(k)} \rangle \le 0$.
M-Perceptron: Update if $y_t \langle x_t, \theta^{(k)} \rangle \le \beta(\gamma) \|\theta^{(k)}\|$

geometric boundary

A[thm] like PCT

we are UNKNOWN

⇓

误分类+分类不清 theoretically!

ii) Batch Perceptron Alg.

⇓

① when we have $\theta^{(k)}$. find and set I_k.

$$I_k = \{t : y_t < \underline{x}_t, \underline{\theta}^{(k)} > \leq 0\}$$

② Update: $\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} + \eta_k \sum_{i \in I_k} y_i \underline{x}_i$