Announcement: Office hours every Thurs 4pm. on Zoom (9439098564)

Primal SVM

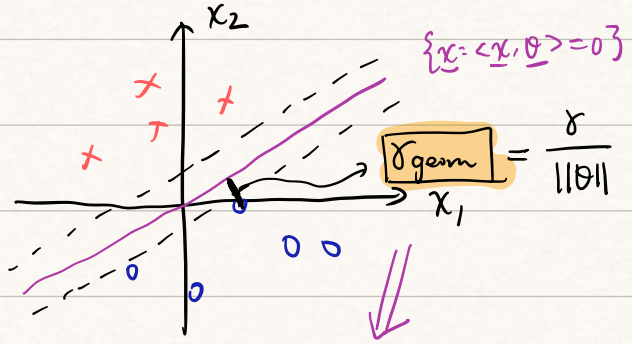
Recall: $\mathcal{D} = \{(\underline{x}_t, y_t) : 1 \le t \le n\}$ $\underline{x}_t \in \mathbb{R}^d$, $y_t \in \{-1, +1\}$.

want to find a linear classifier:

$$\mathcal{F}_{lin} = \{ f : \mathbb{R}^d \to \{-1,+1\} : f(\underline{x}) = \text{sign}(\langle \underline{x}, \underline{\theta} \rangle) \text{ for some } \theta \in \mathbb{R}^d \}.$$

$\Downarrow$

Perceptron Alg. to learn the linear classifier.

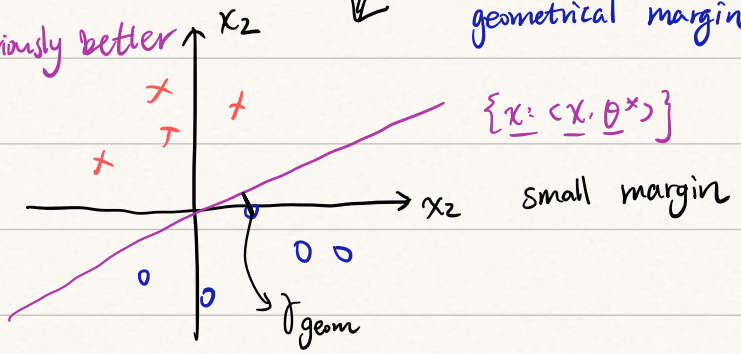


Perceptron Alg. → the decision boundary is not unique.

large margin

this is obviously better

we cannot control the geometrical margin



small margin

$\gamma = \min_{1 \le t \le n} y_t \langle \underline{x}_t, \underline{\theta} \rangle$ → agreement

Rmk: Perceptron alg. don't give us a unique $\underline{\theta} \in \mathbb{R}^d$.

The result $L = \{\underline{x} : \langle \underline{x}, \underline{\theta} \rangle = 0\}$ may have a small margin.

$\underline{\theta}$: parameter for the linear classifier

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

Constraint ① Classifies all samples correctly

$\Leftrightarrow$ $y_t \langle \underline{x}_t, \underline{\theta} \rangle \ge \gamma$ for all $t = 1, 2, \ldots, n$ ⇒ we want to find $\underline{\theta}$ satisfy this

objective ② At the same time, max γ_{geom}

$\Leftrightarrow \max \frac{\gamma}{\|\theta\|}$

optimization problems!

$\max_{\underline{\theta} \in \mathbb{R}^d} \frac{\gamma}{\|\theta\|}$ subject to $y_t \langle \underline{x}_t, \underline{\theta} \rangle \geq \gamma$ $\Rightarrow$ its solution is guaranteed by γ-linearly separable.

maximize min. distance to decision boundary.

correctly classified.

(γ is positive).

$\min_{\underline{\theta} \in \mathbb{R}^d} \left\| \frac{\underline{\theta}}{\gamma} \right\|$ s.t $y_t \langle \frac{\underline{\theta}}{\gamma}, \underline{x}_t \rangle \geq 1$. for every t.

Actually γ determined by $\underline{\theta}$

let $\underline{\theta}'$ be $\frac{\underline{\theta}}{\gamma}$

$\min_{\underline{\theta}' \in \mathbb{R}^d} \|\underline{\theta}'\|$ s.t $y_t \langle \underline{\theta}', \underline{x}_t \rangle \geq 1$ for every t.

$\Updownarrow$

$\min_{\underline{\theta} \in \mathbb{R}^d} \|\underline{\theta}\|$ s.t $y_t \langle \underline{\theta}, \underline{x}_t \rangle \geq 1$. for every t.

Primal – SVM (without offset & slack)

$\min_{\underline{\theta} \in \mathbb{R}^d} \frac{1}{2}\|\underline{\theta}\|^2$ s.t $y_t \langle \underline{\theta}, \underline{x}_t \rangle \geq 1$ for $\forall t$

Uniqueness

Claim: The solution to Primal – SVM is unique. → Contradiction

Pf: $\exists$ 2 different solution vectors $\underline{\theta}_1, \underline{\theta}_2$. $(\underline{\theta}_1 \neq \underline{\theta}_2)$

It must hold $\|\underline{\theta}_1\| = \|\underline{\theta}_2\|$ (obviously)

Consider $\underline{\theta} = \underline{\theta}_1 + \underline{\theta}_2$

$y_t \langle \underline{\theta}, \underline{x}_t \rangle = \frac{1}{2} y_t \langle \underline{\theta}_1, \underline{x}_t \rangle + \frac{1}{2} y_t \langle \underline{\theta}_2, \underline{x}_t \rangle$

$\geq \frac{1}{2} + \frac{1}{2} = 1 \Rightarrow \underline{\theta}$ is feasible

$\|\underline{\theta}\| = \|\frac{1}{2}(\underline{\theta}_1 + \underline{\theta}_2)\| \leq \frac{1}{2}\|\underline{\theta}_1\| + \frac{1}{2}\|\underline{\theta}_2\| = \|\underline{\theta}_1\| = \|\underline{\theta}_2\|$

i) $\|\underline{\theta}\| < \|\underline{\theta}_1\| = \|\underline{\theta}_2\|$ → contradicts to the choice of $\underline{\theta}_1$ & $\underline{\theta}_2$

ii) $\|\theta_1\| = \|\theta_2\| = \|\theta\| \Rightarrow$ triangle inequality holds with equality

$\hookrightarrow \|a+b\| = \|a\| + \|b\| \Leftrightarrow a = \lambda b$

$\|\frac{1}{2}\theta_1 + \frac{1}{2}\theta_2\| = \|\frac{1}{2}\theta_1\| + \|\frac{1}{2}\theta_2\| \Leftrightarrow \frac{1}{2}\theta_1 = \frac{1}{2}\theta_2 \cdot \lambda.$

$\updownarrow$

$\|\theta_1\| = \|\theta_2\| \Rightarrow \lambda = 1 \Rightarrow \theta_1 = \theta_2.$

contradiction!

$(\theta_1 \neq \theta_2)$

$\Rightarrow$ The solution is unique #

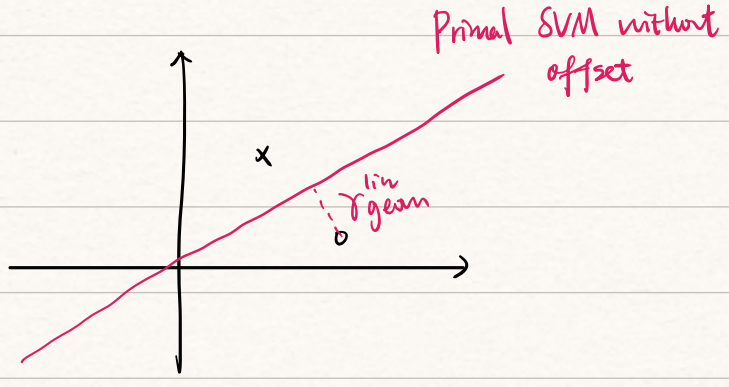
General Primal SVM with offset

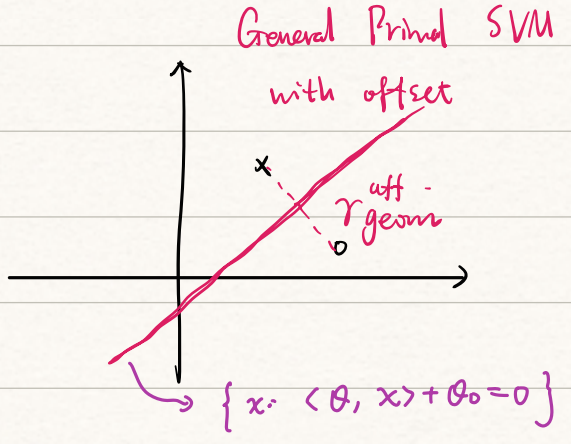
① $\mathcal{F}_{aff} := \{ f : \mathbb{R}^d \to \{+1, -1\} : f(\underline{x}) = \text{sgn}(\langle \underline{x}, \underline{\theta} \rangle + \theta_0)$ for some $\underline{\theta} \in \mathbb{R}^d$ & $\theta_0 \in \mathbb{R} \}$

② Each $f \in \mathcal{F}_{aff}$ is parametrized & $(\underline{\theta}, \theta_0) \in \mathbb{R}^{d+1}$

③ Decision boundary is set of all $\underline{x}$ satisfying $\{ \underline{x} : \langle \underline{x}, \underline{\theta} \rangle + \theta_0 = 0 \}$.

Rmk: offset parameter can lead to larger γ_{geom}





$\rightarrow \gamma_{geom}^{lin} \leq \gamma_{geom}^{aff}$.

Primal SVM with offset problem:

$$\min_{(\underline{\theta},\theta_0)\in\mathbb{R}^{d+1}} \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{s.t} \quad y_t(\langle \underline{\theta}, \underline{x}_t\rangle + \theta_0) \geq 1 \quad \text{for all } t.$$

By writing the constraints, we are assuming that:

dataset is affinely-separable

$\Rightarrow \exists\, (\underline{\theta},\theta_0)\in\mathbb{R}^{d+1}$ s.t $y_t(\langle \underline{\theta}, \underline{x}_t\rangle + \underline{\theta}_0) \geq 0$ for all t.

Question: Why not convert objective function to $\frac{1}{2}\left\|\begin{bmatrix}\underline{\theta}\\ \theta_0\end{bmatrix}\right\|^2$?

Rmk: i) θ_0 only appears in constraints of obj. function

ii) Different from modeling each feature vector $\underline{x}_t$ as $\underline{x}'_t = \begin{bmatrix}\underline{x}_t\\ 1\end{bmatrix}$. (not the right thing to do!)

Ex: $\mathcal{D}$: γ-affinely separable

$L := \{\underline{x} : \langle \underline{x}, \underline{\theta}\rangle + \theta_0 = 0\}$

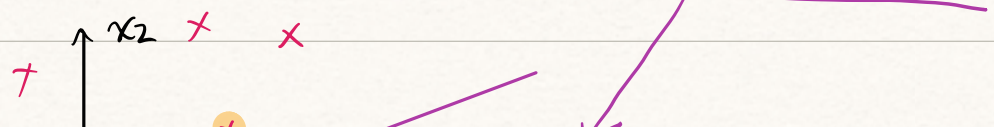
With the offset θ_0, the min distance over all training examples to L is still $\frac{\gamma}{\|\underline{\theta}\|}$ (not θ_0).

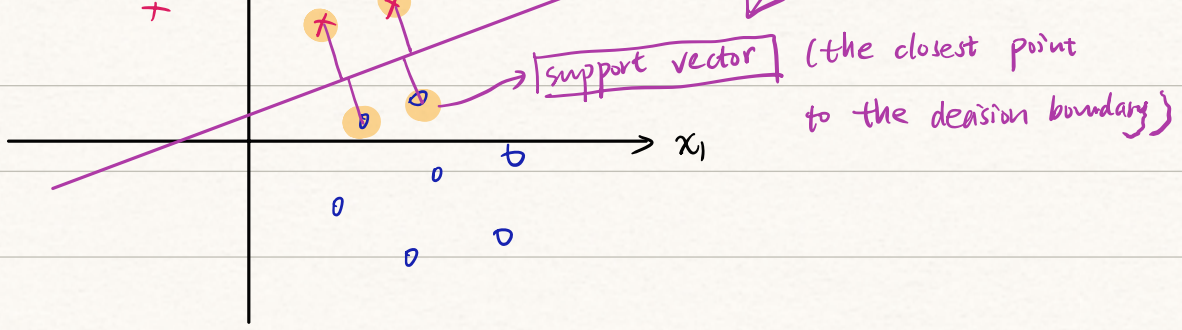
Robustness of Primal SVM

The primal-SVM with offset prob. is quadratic program

- objective quadratic
- constraints linear

Solution to QP depends only on a small # of training samples





Quantify the quality of Primal-SVM with offset

↳ use Leave-one-out Cross-Validation error

$$\text{LOOCV} := \frac{1}{n}\sum_{t=1}^{n} \text{Loss}\left(y_t, f(x_t; \hat{\theta}^{-t}, \hat{\theta}_0^{-t})\right)$$

Loss: 0-1 loss

$(\hat{\theta}^{-t}, \hat{\theta}_0^{-t})$: parameters learned when (x_t, y_t) is removed from $\mathcal{D}$

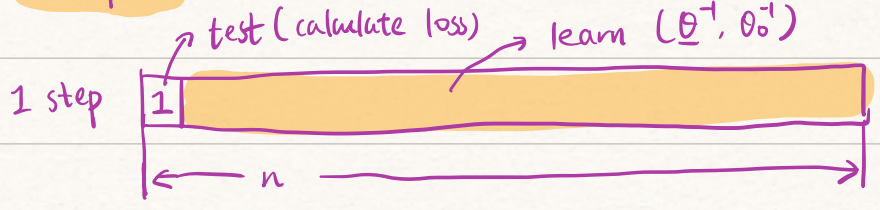
Actually there are n SVMs!

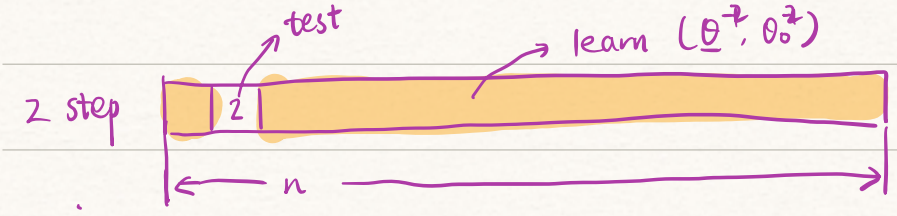
$$(\hat{\theta}^{-t}, \hat{\theta}_0^{-t}) = \underset{(\theta,\theta_0)\in\mathbb{R}^{d+1}}{\text{argmin}} \ \frac{1}{2}\|\theta\|^2 \quad \text{s.t} \ \ y_s(\langle \theta, x_s\rangle + \theta_0) \geq 1$$

for $s \in \{1,2,\dots,n\}/\{t\}$

$$\text{Loss}(y, \hat{y}) = \begin{cases} 1, & y \neq \hat{y} \\ 0, & \text{else} \end{cases}$$

The process of LOOCV:





⋮

☆ → try to mimic the scenario that you are training samples and test on an independent set!

Rmk: If LOOCV is small, model generalization well!

Prop: Let the # of support vectors be $N \in \{0,1,\dots,n\}$.

Then $\text{LOOCV} \leq \frac{N}{n}$ (obviously)

Rmk: If # of support vectors is small,

then PRIMAL-SVM has a good quality of generalization!

Allowing Misclassified Samples

↳ dataset is not affinely separable

Real datasets are often not affinely separable

↓

PRIMAL-SVM with offset & slack variables

$$\min_{(\underline{\theta}, \theta_0, \underline{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^n} \quad \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n} \xi_t \qquad C > 0.$$

$$\text{s.t} \quad y_t(\langle \underline{\theta}, x_t \rangle + \theta_0) \geq 1 - \xi_t \qquad t = 1, 2, \ldots, n$$

① if dataset is affinely separable, then $\underline{\xi} = \underline{0}$ can be achieved

② if dataset is not affinely separable, $\underline{\xi}$ can not be $\underline{0}$

↳ $\underline{\xi} \geq 0$

Rmk: i) Margin constraint violated for t^{th} example if $\xi_t > 0$

ii) Penalty to objective → $C\xi_t$

iii) $C \to +\infty$, then $\xi_t \to 0$ for all t

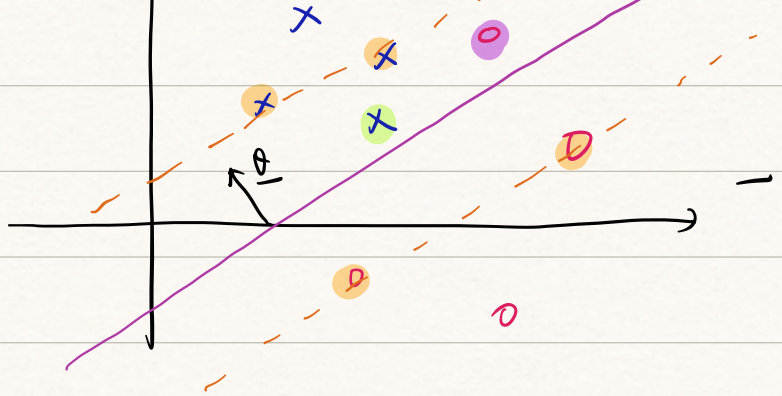
⇒ we obtain original SVM (without ξ_t slack)

iv) for general case $C > 0$, some margin violation is allowed.

v) trade-off between max-margin objective $\frac{1}{2}\|\underline{\theta}\|^2$

and margin violation $C\sum_t \xi_t$

Q: Which are the support vectors for SVM with slack & offset?

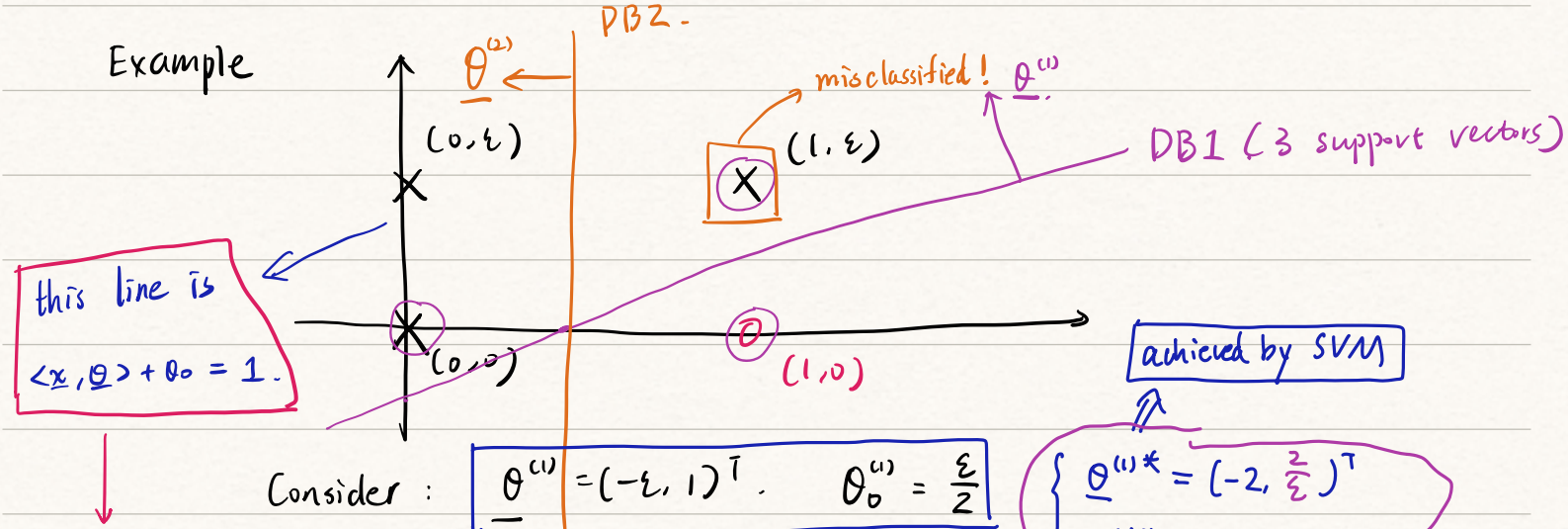


Answer: i) Points that lie on margin ⇒ Usual case

ii) Those that violate margin constraints but not enough to be misclassified

iii) Misclassified points

Example



achieved by SVM

Consider: $\underline{\theta}^{(1)} = (-\varepsilon, 1)^T$. $\theta_0^{(1)} = \frac{\varepsilon}{2}$

$$\begin{cases} \underline{\theta}^{(1)*} = (-2, \frac{2}{\varepsilon})^T \\ \theta_0^{(1)*} = 1 \end{cases}$$

this is because these two points are support vectors on the line.

DB1: $\langle \underline{\theta}^{(1)}, \underline{x} \rangle + \theta_0 = 0$

$\Leftrightarrow -\varepsilon x_1 + x_2 + \frac{\varepsilon}{2} = 0$

$\Rightarrow \langle \underline{\theta}^{(1)*}, \underline{x} \rangle + \theta_0^{(1)*} = 0$

$\Rightarrow \langle \underline{\theta}^{(1)*}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rangle + \theta_0^{(1)*} = 1$

Classify all samples correctly, but margin small.

for SVM algorithm, we must guarantee that $y_t(\langle \underline{\theta}^*, \underline{x}_t \rangle + \theta_0^*) = 1$ for those $(\underline{x}_t, y_t)$ lie in the boundary

$\underline{\theta}^{(2)*} = (-2, 0)$ $\theta_0^{(2)*} = 1$

$\underline{\theta}^{(2)} = (-1, 0)$ $\theta_0^{(2)} = \frac{1}{2}$

PB2: $-x_1 + \frac{1}{2} = 0 \Rightarrow x_1 = \frac{1}{2}$

Top right point is misclassified ⇒ 1 tag error

margin is larger $r_2 \gg r_1$

$\Rightarrow$ training error costs $2C$! ($\xi_t = 2$) ↘ explain why $\xi_t = 2$

since $\langle \underline{\theta}^{(2)*}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \rangle + \theta_0^{(2)*}$

$= -2 + 1$

$= -1$

$= 1 - (2)$ → ξ_t

To conclude, we prefer DB1 to DB2 if

$$\frac{1}{2\gamma_1^2} \leq \frac{1}{2\gamma_2^2} + 2C$$

↳ this B actually from:

objective: $\frac{1}{2}\|\theta\|^2 + C\sum_t \xi_t$ → $\frac{1}{2\gamma_{geom}^2}$

$\gamma_1 = \frac{1}{\|\underline{\theta}_1^*\|} = \frac{1}{\sqrt{4 + \frac{4}{2^2}}}$

$\gamma_2 = \frac{1}{\|\underline{\theta}_2^*\|} = \frac{1}{2}$

{ If $C \to \infty$, this inequality is satisfied and all points are classified correctly

If $C \approx 0$, this means we are allowed to make some mistakes!

Outline:

1. Start from Perceptron → disadvantage { not unique; not robust

⇓ still linear classifier first.

SVM · (HARD) → USE FOR LINEARLY SEPARABLE DATASET!

Model: $\max \frac{\gamma}{\|\theta\|}$

$$\min_{\theta} \quad \|\theta\|$$

$$\text{s.t} \quad y_t \langle x_t, \underline{\theta} \rangle \geq \gamma \quad \text{for all } t$$

$$\Updownarrow$$

$$\min_{\theta} \quad \frac{\|\theta\|}{\gamma} := \|\theta\|$$

$$\text{s.t} \quad y_t \langle x_t, \frac{\underline{\theta}}{\gamma} \rangle \geq 1. \quad \text{for all } t$$

$$\Updownarrow$$

$$\min_{\theta} \quad \|\theta\| \iff \min_{\underline{\theta}} \frac{1}{2}\|\theta\|^2$$

$$\text{s.t} \quad y_t \langle x_t, \underline{\theta} \rangle \geq 1 \quad \text{for all } t$$

↳ Support vector (x_t, y_t)

$\Downarrow$

Satisfy that $y_t \langle x_t, \hat{\underline{\theta}} \rangle = 1$.

$\Downarrow$

★ this implies MARGIN $\gamma_{geom} = \frac{1}{\|\hat{\theta}\|}$

2. SVM with offsets $\longrightarrow$ Affine Classifier.

↳ $f_{\underline{\theta}, \theta_0}(x) = \langle \underline{x}, \underline{\theta} \rangle + \theta_0$

Model: $\min_{\underline{\theta}, \theta_0} \frac{1}{2}\|\theta\|^2$

$\text{s.t} \quad y_t\left(\langle x_t, \underline{\theta} \rangle + \theta_0\right) \leq 1$

Note, $\gamma_{geom} = \frac{1}{\|\hat{\theta}\|}$ also holds!

3. Advantage of SVM (HARD) { Uniqueness of Solution; Robustness

$\Downarrow$

i.e., support vectors (# is K.) will restrict

$LOOCV \leq \frac{K}{n}$

4. Soft SVM (SVM with offset & slackness)

① Model: $\min_{\underline{\theta}, \theta_0, \underline{\xi}} \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{1}^{n} \xi_t$

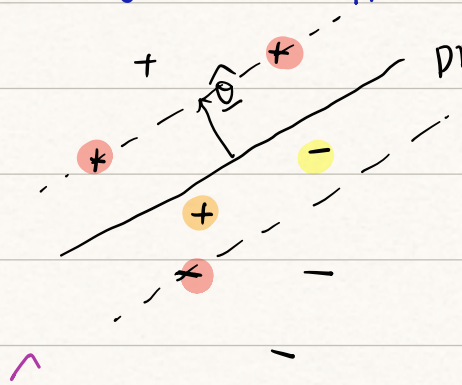
s.t $y_t(\langle \underline{x}_t, \underline{\theta}\rangle + \theta_0) \geq 1 - \xi_t \quad t = 1, 2, \dots, n$

the meaning of $\hat{\underline{\theta}}, \hat{\theta}_0$:

we have the { decision boundary
margin boundary

the distance from DB to MB is $\frac{1}{\|\underline{\theta}\|}$

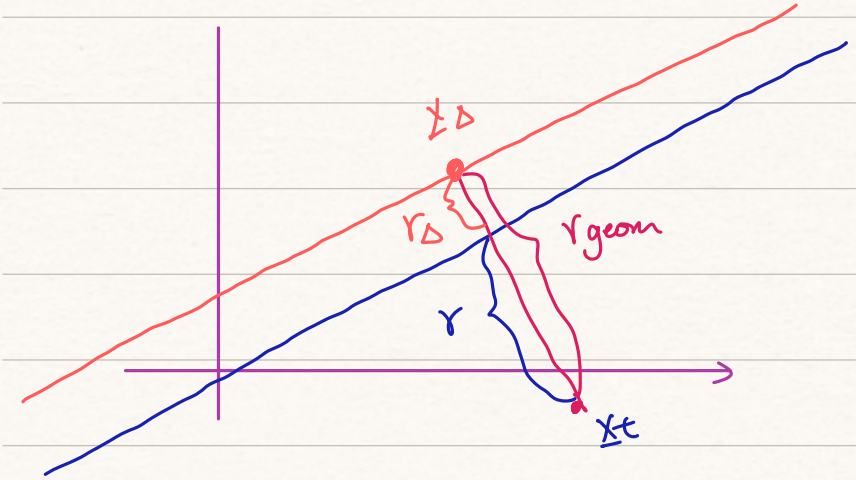
② Three types of Support vectors



- lies in DB.
- misclassified
- have a penalty but classified correctly

$\langle \underline{\theta}, \underline{x}\rangle + \theta_0 = 0$

$\langle \underline{\theta}, \underline{x}\rangle = 0$



- $r = \frac{(\langle \underline{\theta}, \underline{x}_t\rangle) \cdot y_t}{\|\underline{\theta}\|}$

$r_\Delta = \frac{(\langle \underline{\theta}, \underline{x}_\Delta\rangle \cdot y_\Delta}{\|\underline{\theta}\|}$ ($-\theta_0$, $-y_t$)

$\Rightarrow r^g = \frac{(\langle \underline{\theta}, \underline{x}_t\rangle + \theta_0) y_t}{\|\underline{\theta}\|}$

For another case $\begin{cases} y_\Delta = y_t \\ r^g = r - r_\Delta \end{cases}$ ✓

$$\Rightarrow r^g = \frac{(\langle \underline{\theta}, x_t \rangle + \theta_0)\, y_t}{\|\underline{\theta}\|}$$

$\Rightarrow$ for affine-SVM

$$\min \frac{1}{2}\|\underline{\theta}\|^2$$

$$\text{s.t} \quad y_t(\langle \underline{\theta}, x_t \rangle + \theta_0) \geq 1$$

$$\Rightarrow \frac{1}{\|\underline{\theta}^*\|} \rightarrow r^g$$