Def$^{\underline{n}}$: Functional margin of a hyperplane (w, b) w.r.t. $(x^{(i)}, y^{(i)})$ is:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

If $y^{(i)} = 1$. want $w^T x^{(i)} + b \gg 0$

If $y^{(i)} = -1$. want $w^T x^{(i)} + b \ll 0$

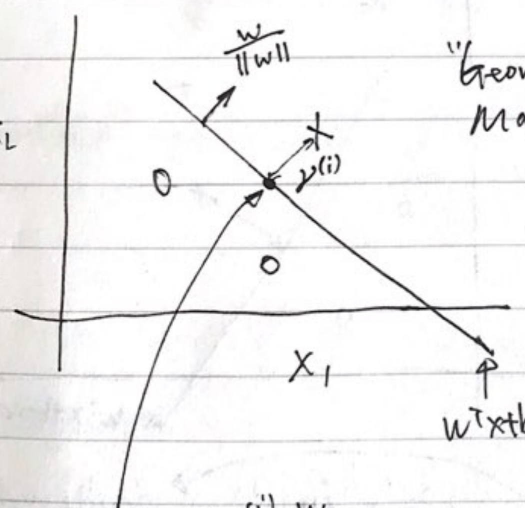
If $y^{(i)}(w^T x^{(i)} + b) > 0$. then classified $(x^{(i)}, y^{(i)})$ correctly.

$$\hat{\gamma} = \min_i \hat{\gamma}^{(i)} \quad \text{(worst case)}$$

$w \to 2w$

$b \to 2b$.

then $\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$ doubles. so: $\|w\| \stackrel{set}{=} 1$.



"Geometric Margin"

$x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|}$

$$w^T\left(x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|}\right) + b = 0.$$

$$w^T x^{(i)} + b = \gamma^{(i)} \frac{w^T w}{\|w\|} = \gamma^{(i)} \|w\|$$

$$\gamma^{(i)} = \left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}$$

More generally, geometric margin

$$\gamma^{(i)} = y^{(i)}\left(\frac{w^T}{\|w\|} x + \frac{b}{\|w\|}\right).$$

If $\|w\| = 1$:

$$\hat{\gamma}^{(i)} = \gamma^{(i)}$$

$$\boxed{\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|w\|}}$$

Geometric margin:

$\gamma = \min_i \gamma^{(i)}$

Maximum classifier:

$$\boxed{\begin{aligned} &\max_{\gamma, w, b} \ \gamma \\ &\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma \quad (\forall i) \\ &\qquad \|w\| = 1 \end{aligned}}$$

$w \to 10w$

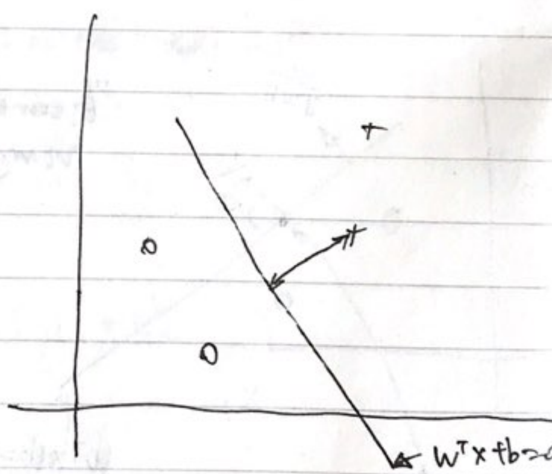
$b \to 10b$

does not change Geometric margin

Lecture 7

- Optimal Margin Classifier
- Primal/Dual optimization problem (KKT)
- SVM dual
- Kernels

$$\left[\begin{aligned} &h_{w,b}(x) = g(w^T x + b) \\ &g(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ &y \in \{-1, 1\} \end{aligned}\right.$$



Func. Margin: $\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$

Geo. Margin: $\gamma^{(i)} = y^{(i)}\left(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|}\right)$

$\gamma = \min_i \gamma^{(i)}$ → double w, b will not change the position of the hyperplane

$\hat{\gamma} = \min_i \hat{\gamma}^{(i)}$

$\|w\| = 1$

$|w_1| = 1$

$w_1^2 + |w_1| = 17$

equivalent edition of the optimal problem:

#1: $\max_{\gamma, w, b} \gamma$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq \gamma \quad (i=1,\dots,m)$

$\|w\| = 1$ ← Non-convex.

#2: $\max_{\hat{\gamma}, w, b} \frac{\hat{\gamma}}{\|w\|}$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}$

$\frac{\hat{\gamma}}{\|w\|} = \gamma$

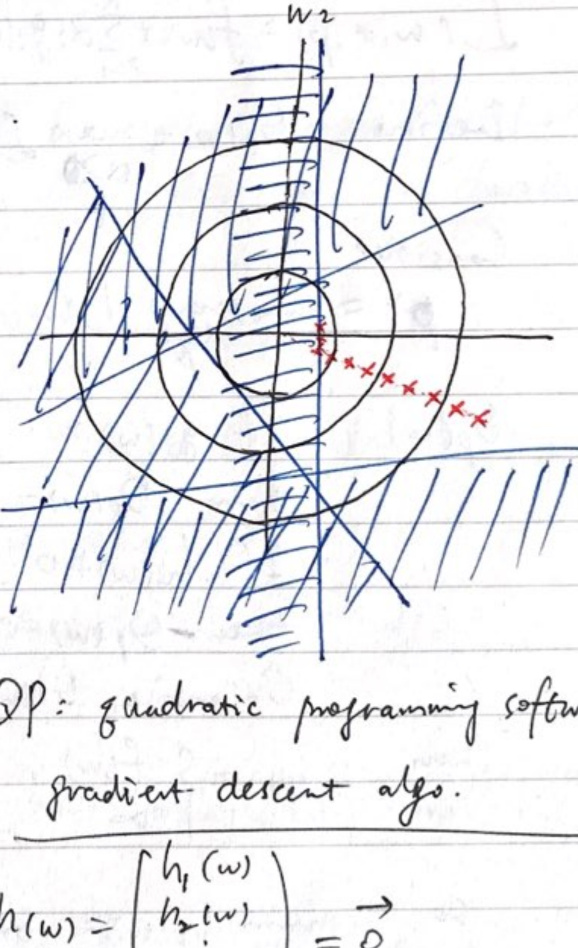
$\boxed{\hat{\gamma} = 1}$ ← impose this constraint

then $\min_i y^{(i)}(w^T x^{(i)} + b) = 1$.

#3:

$\min_{w,b} \|w\|^2$ ← $\max \frac{1}{\|w\|}$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq 1$



Convex optimization:

$\min_w f(w)$

s.t. $h_i(w) = 0, \ i = 1, \dots, l.$

QP: quadratic programming software

gradient descent algo.

$$h(w) = \begin{pmatrix} h_1(w) \\ h_2(w) \\ \vdots \\ h_l(w) \end{pmatrix} = \vec{0}$$

Lagrangian:

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Lagrange Multipliers

$$\frac{\partial \mathcal{L}}{\partial w} \overset{\text{set}}{=} 0 \qquad \frac{\partial \mathcal{L}}{\partial \beta} \overset{\text{set}}{=} 0.$$

for w^* to be a solution, it is necessary that

$\exists \beta^*$ s.t. $\dfrac{\partial f(w^*, \beta^*)}{\partial w} = 0$. $\dfrac{\partial f(w^*, \beta^*)}{\partial \beta} = 0$.

Primal Problem

$\min f(w)$

s.t. $g_i(w) \le 0$, $i = 1 \dots k$ ("$g(w) \le \vec{0}$")

$h_j(w) = 0$. $j = 1 \dots l$ ("$h(w) = \vec{0}$")

Lagrangian:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{l} \beta_j h_j(w)$$

Define: $\Theta_p(w) = \max_{\substack{\alpha \ge 0 \\ \beta}} \mathcal{L}(w, \alpha, \beta)$

Consider:

$$p^* = \min_w \max_{\substack{\alpha \ge 0 \\ \beta}} \mathcal{L}(w, \alpha, \beta) = \min_w \Theta_p(w)$$

$\Theta_p(w)$:
- If $g_i(w) > 0$
  then $\Theta_p(w) = \infty$
- If $h_i(w) \ne 0$
  then $\Theta_p(w) = \infty$
- Otherwise, $\Theta_p(w) = f(w)$

Then $\Theta_p(w) = \begin{cases} f(w), & \text{if constraints satisfied } (g, h) \\ \infty, & \text{otherwise} \end{cases}$

so. $\min_w \Theta_p(w) =$ original problem

Dual Problem

$\theta_D(\alpha,\beta) = \min_{w} \mathcal{L}(w,\alpha,\beta)$

$d^* = \max_{\substack{\alpha \geq 0 \\ \beta}} \min_{w} \mathcal{L}(w,\alpha,\beta) = \max_{\substack{\alpha \geq 0 \\ \beta}} \theta_D(\alpha,\beta)$

$d^* \leq p^*$ $\boxed{\max \min(\dots) \leq \min \max(\dots)}$

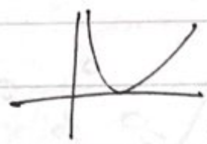
↑ general result for any functions

$$\max_{y\in\{0,1\}} \underbrace{\left(\min_{x\in\{0,1\}} \mathbb{1}\{x=y\}\right)}_{0} \leq \min_{x\in\{0,1\}} \underbrace{\left(\max_{y\in\{0,1\}} \mathbb{1}\{x=y\}\right)}_{1}$$

Some conditions: $d^* = p^*$ | for optimal margin classifier. dual problem has better character.

---

let f be convex (Hessian $H \geq 0$)



suppose h_i is affine ($h_i(w) = a_i^T w + b$)

and suppose g_i is (strictly) feasible. ($\exists w$ st. $\forall i$: $g_i(w) < 0$)

Then $\exists w^*, \alpha^*, \beta^*$ st. w^* solve primal

α^*, β^* solve dual, and $p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$

$$\begin{cases} \frac{\partial}{\partial w}\mathcal{L}(w^*,\alpha^*,\beta^*) = 0 \\ \boxed{\alpha_i^* g_i(w^*) = 0} \leftarrow \text{KKT complementary condition} \\ \frac{\partial}{\partial \beta}\mathcal{L}(w^*,\alpha^*,\beta^*) = 0 \\ g_i(w^*) \leq 0 \\ \alpha_i^* \geq 0 \end{cases}$$

KKT: short for Karush-Kuhn-Tucker

KKT condition:

If $\alpha_i > 0 \Rightarrow g_i(w^*) = 0$ ($g_i(w)$ is an "active" constraint).

Lagrange multiplier: $\alpha_i, \beta_i \xrightarrow{\text{in SVM problem}} \alpha_i$

Parameters : $w \xrightarrow{\text{in SVM problem}} w, b$ (notation change

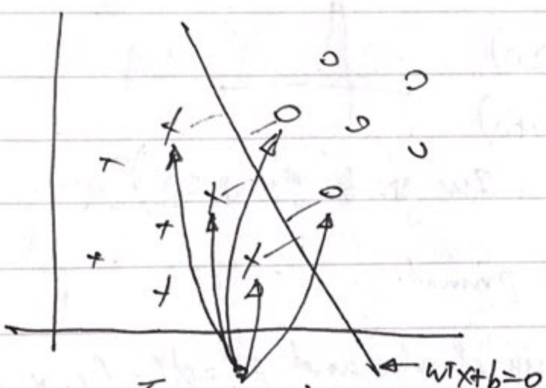
$$\min \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad \underbrace{y^{(i)}(w^T x^{(i)} + b) \geq 1}, \quad i = 1 \ldots m.$$

$$g_i(w, b) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0$$

$\alpha_i > 0 \Rightarrow g_i(w, b) = 0$ (active constraint)

$\Leftrightarrow (x^{(i)}, y^{(i)})$ has functional margin 1.



Func. Margin=1
(usually $\alpha \neq 0$)
"Support Vectors"

Support vectors : $\alpha_i > 0$

$\alpha_i = 0$ for most non-support vectors

$$\min \mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 \right)$$

Dual problem:

$$\theta_D(\alpha) = \min_{w,b} \mathcal{L}(w,b,\alpha)$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \overset{\text{set}}{=} 0 \Rightarrow \boxed{w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}}$$

$$\frac{\partial}{\partial b} \mathcal{L}(w,b,\alpha) = - \sum_{i=1}^{m} \alpha_i y^{(i)} \overset{\text{set}}{=} 0 .$$

$$\mathcal{L} = \frac{1}{2} \underbrace{w^T w}_{} - \sum_{i=1}^{m} \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 \right)$$

$$\hookrightarrow \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)} \right)$$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$- \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle + \sum_{i=1}^{m} \alpha_i$$

$$= \underbrace{\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle}_{W(\alpha)}$$

$$\max_{\alpha} \theta_D(\alpha)$$

- If $\sum_i y^{(i)} \alpha_i \neq 0$
  $\theta_D(\alpha) = -\infty$
- If $\sum_i y^{(i)} \alpha_i = 0$
  then $\theta_D(\alpha) = W(\alpha)$

Dual problem:

$\max W(\alpha)$

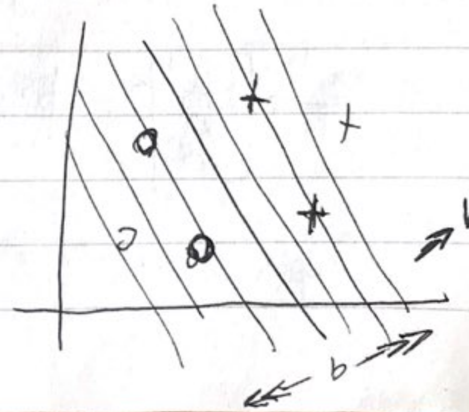
s.t. $\alpha_i \geq 0$

$\sum_i y^{(i)} \alpha_i = 0$

Solve for α. (later)

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$b = \frac{\max_{i: y^{(i)}=1} w^T x^{(i)} + \min_{i: y^{(i)}=0} w^T x^{(i)}}{2}.$$



31

$w = \sum_i \alpha_i y^{(i)} x^{(i)}$

$h_{w,b}(x) = g(w^T x + b)$

$\uparrow$ threshold function

$w^T x + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$

Kernels $x^{(i)}$ — very high dimentional ($x^{(i)} \in \mathbb{R}^{\infty}$)

$\langle x^{(i)}, x^{(j)} \rangle$ efficiently computed

$\boxed{\text{Lecture 8}}$

SVM
- kernels
- soft margin
- SMO

$\min \frac{1}{2} \|w\|^2$

s.t. $y^{(i)} (w^T x^{(i)} + b) \geq 1$

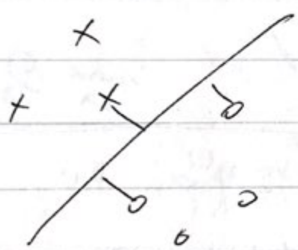
Dual Problem:

$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \boxed{\langle x^{(i)}, x^{(j)} \rangle}$

s.t. $\alpha_i \geq 0$

$\sum_i y_i \alpha_i = 0$

$w = \sum_i \alpha_i y^{(i)} x^{(i)}$



$h_{w,b}(x) = g(w^T x + b)$

$= g\left(\sum_i \alpha_i y^{(i)} \boxed{\langle x^{(i)}, x \rangle} + b \right)$

Have $x \in \mathbb{R}$ living area

$x \xrightarrow{\phi} \begin{bmatrix} x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix} = \phi(x)$

Replace $\langle x^{(i)}, x^{(j)} \rangle$ with $\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$

$\phi(x)$ - very high dim. cannot compute efficiently

$\underbrace{K(x^{(i)}, x^{(j)})} = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$: sometimes compute efficiently.

$x, z \in \mathbb{R}^n$

$$K(x,z) = (x^T z)^2 = \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_j z_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j)(z_i z_j)$$

$$= (\phi(x))^T(\phi(z))$$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \\ \sqrt{2} x_3 \\ c \end{bmatrix}$$

Need $O(n^2)$ to compute $\phi(x)$.

But need $O(n)$ time to compute $K(x, z)$

$K(x,z) = (x^T z + c)^2$

generally: $K(x,z) = (x^T z + c)^d$

$\rightarrow \binom{n+d}{d}$ features of all monomials up to degree d.

$x \mapsto \phi(x) \qquad z \mapsto \phi(z)$

$\langle \phi(x), \phi(z) \rangle$

x, z similar : $\langle \phi(x), \phi(z) \rangle$ maybe large;
- - - dissimilar : — — - - - small.

How to choose kernel function?

$K(x,z)$ - large if x, z similar

small if x, z dissimilar

$$K(x,z) = e^{-\frac{\|x-z\|^2}{2\sigma^2}} \rightarrow \text{invalid?}$$

$\exists\, \phi$ s.t. $K(x,z) = \langle \phi(x), \phi(z) \rangle$?

Suppose K is a kernel. let $\{x^{(1)}, \ldots, x^{(m)}\}$ be given

let $K \in \mathbb{R}^{m \times n}$

$$K_{ij} = K(x^{(i)}, x^{(j)})$$

Then for any vector $z \in \mathbb{R}^m$. $z^T K z$

$$\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \, \phi(x^{(i)})^T \phi(x^{(j)}) \, z_j \\
&= \sum_i \sum_j z_i \sum_k (\phi(x^{(i)}))_k (\phi(x^{(j)}))_k \, z_j \\
&= \sum_k \sum_i \sum_j z_i (\phi(x^{(i)}))_k (\phi(x^{(j)}))_k \, z_j \\
&= \sum_k \left(\sum_i z_i \, \phi(x^{(i)})_k \right)^2 \geq 0
\end{aligned}$$

$K \geq 0$. (positive semi-definite). And the inverse judgement is true

Theorem (Mercer): Let $K(x,z)$ be given, then K is valid (is Mercer) **Kernel** (i.e. $\exists \phi$ st. $K(x,z)=\phi(x)^T\phi(z)$) iff. for all $\{x^{(1)} \dots x^{(m)}\}$ ($m<\infty$) the kernel matrix $K \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite.

Use this to test a function is/not a kernel

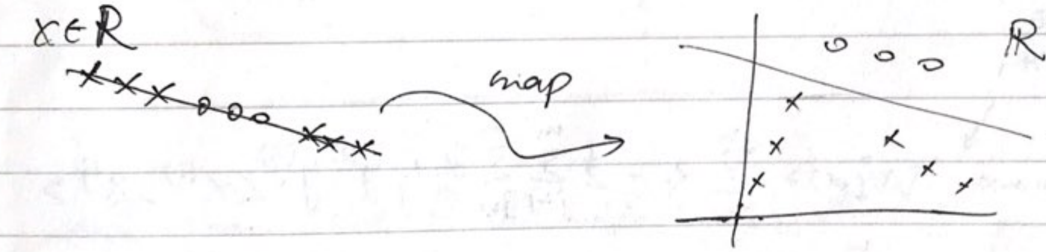
eg: $K(x,x)=-1$ not valid. $\because -1 \neq \phi(x)^T\phi(x)$

Choose $K(x,z)=e^{-\frac{\|x-z\|^2}{2\sigma^2}}$ (Gaussian Kernel)

or $(x^Tz+c)^d$, or etc.

Replace $\langle x^{(i)}, x^{(j)}\rangle$ with $\underline{K(x^{(i)}, x^{(j)})}$ ✓

$x^{(i)} \longrightarrow \underline{\phi(x^{(i)})}$ ✗

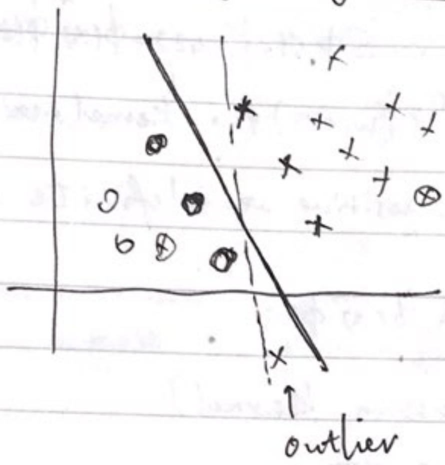


have linear classifier in high-dim. space

~~$\langle x^{(i)}, x^{(j)}\rangle \to K(x^{(i)}, x^{(j)})$ for any algo. which can be written as inner-product form~~

$\langle x^{(i)}, x^{(j)}\rangle \to K(x^{(i)}, x^{(j)})$ for any algo. **which** can be written as inner-product form

L_1 norm soft margin SVM



penalty

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1, \dots, m.$$

If $y^{(i)}(w^T x^{(i)} + b) > 0 \Rightarrow$ classified correctly.

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_{i=1}^{m}\alpha_i\left(y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i\right) - \sum_{i=1}^{m} r_i \xi_i$$

SOME MATH ...

$$\max \ W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle x^{(i)}, x^{(j)}\rangle.$$

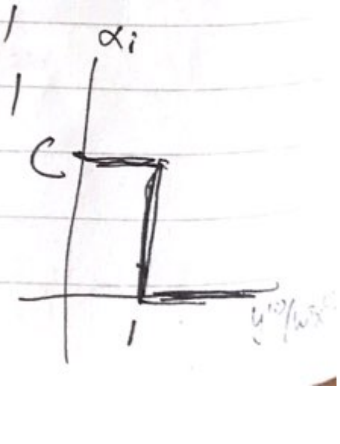
$$\text{s.t. } \sum_{i=1}^{m} y^{(i)}\alpha_i = 0$$

$$0 \leq \alpha_i \leq C. \quad i = 1, \dots, m$$

---

convergence criteria (derive from KKT condition):

$$\begin{cases} \alpha_i = 0 \Rightarrow y^{(i)}(w^T x^{(i)} + b) \geq 1 \\ \alpha_i = C \Rightarrow y^{(i)}(w^T x^{(i)} + b) \leq 1 \\ 0 < \alpha_i < C \Rightarrow y^{(i)}(w^T x^{(i)} + b) = 1 \end{cases}$$



Digression:

<u>coordinate ascent</u> (another opt. question)

max $W(\alpha_1, \dots \alpha_m)$ (no constraints on α_i's)

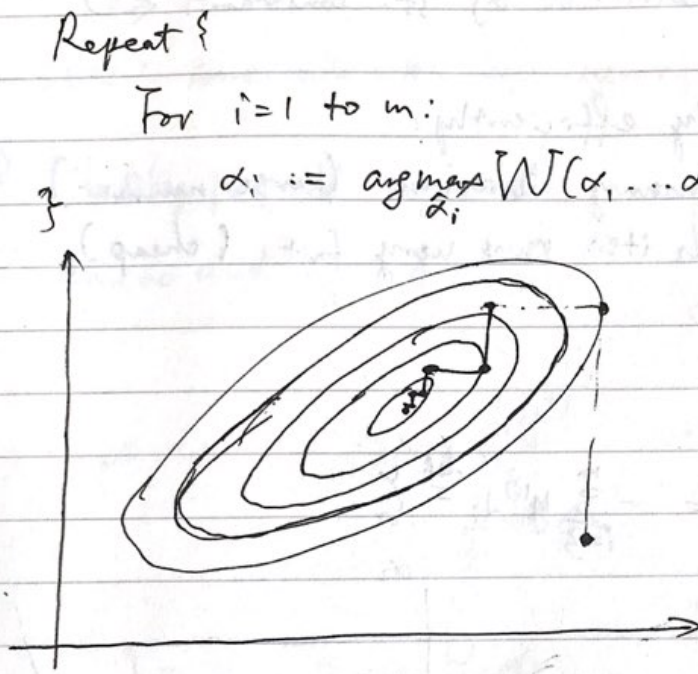
Repeat {

For $i=1$ to m:

Hold everything except α_i fixed.

$\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \dots \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \dots \alpha_m)$

}



high dim: not fixed order $\alpha_1, \dots \alpha_m$

heuristic value function decide which is chosen to vary (change)

* more steps to converge (compared to Newton's method)
* inner loops executes very quick.

<u>SMO</u>

Coordinate ascent cannot work directly on SVM dual opt. problem.

∵ constraints : $\sum_{i=1}^{m} y^{(i)} \alpha_i = 0$.

∴ Change 2 α_i's at a time.

SMO Algo. is due to

John Platt @ Microsoft

Outline:

Select α_i α_j (heuristic)

Hold all α_i's fixed except α_i α_j

(*) key step

Optimize $W(\alpha)$ w.r.t. α_i α_j s.t. constraints

SMO do extremely efficiently:

Although many iterations (large number)

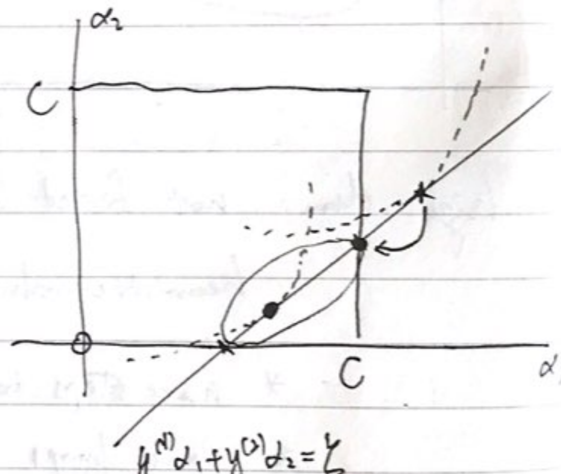
But each iter. runs very fast (cheap)

Update α_1 . α_2

Know $\sum_i y^{(i)}\alpha_i = 0$.

$$y^{(1)}\alpha_1 + y^{(2)}\alpha_2 = -\sum_{i=3}^{m} y^{(i)}\alpha_i \stackrel{\text{def}}{=} \zeta$$

$$0 \le \alpha_i \le C.$$



$$W(\alpha_1, \alpha_2, \alpha_3 \dots \alpha_m)$$

$$= W\left(\frac{\zeta - y^{(2)}\alpha_2}{y^{(1)}}, \alpha_2, \alpha_3, \dots \alpha_m\right)$$

$$= a\alpha_2^2 + b\alpha_2 + c$$

(quadratic func.)

$y^{(1)}\alpha_1 + y^{(2)}\alpha_2 = \zeta$

$$\Rightarrow \alpha_1 = \frac{\zeta - y^{(2)}\alpha_2}{y^{(1)}}$$

optimize quadratic func.

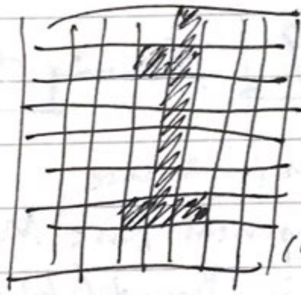
in inner loop operates very efficiently.

Compute b is not hard. Do it after class.

Applications of SVM

① Handler's Digit Recognition

$K(x,y) = (x^T y)^d$ or $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$



10×10

SVM is comparable with best neural networks $\quad x \in \mathbb{R}^{100}$

② Classify protein seq.'s

amino acid seq. (A ⋯ Z)

BAJTSIAIBAJTAU

$\phi(x) = ?$ (hard problem ⋯)

$\phi(x) \in \mathbb{R}^{(20^4)}$

$= \mathbb{R}^{160000}$

$$\begin{array}{l} \text{AAAA} \\ \text{AAAB} \\ \text{AAAC} \\ \vdots \\ \text{AAAZ} \\ \text{AABA} \\ \vdots \\ \text{BAJT} \\ \vdots \\ \text{TSIA} \\ \vdots \\ \text{ZZZZ} \end{array} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 2 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \phi(x)$$

2 ← occur 2 times

use Dynamic Programming (DP):

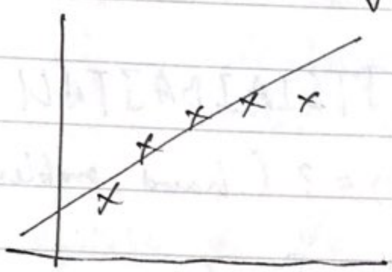
compute $\phi(x)^T \phi(z)$.

Basic (Applicative) Part of this course is over.

NEXT: Understanding these algorithms

Lecture 9 learning theory

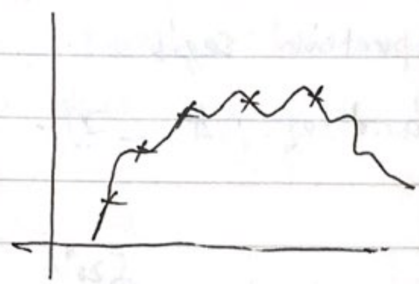
- Bias / Variance
- Empirical Risk Minimization
- Union Bound / Hoeffding Inequality
- Uniform convergence.



$\theta_0 + \theta_1 x$

"underfit"

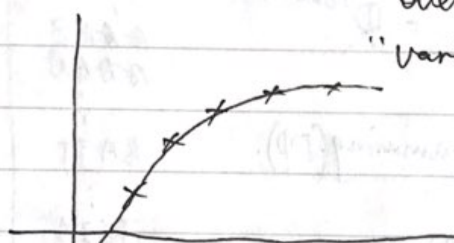
"bias"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"overfit"

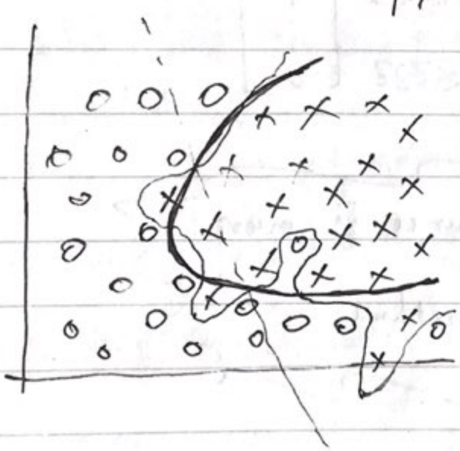
"variance"



$\theta_0 + \theta_1 x + \theta_2 x^2$

"fit"

"balanced"



$h_\theta(x) = g(\theta_0 + \theta_1 x + \cdots + \theta_n x^n)$

Linear Classification:

$h_\theta(x) = g(\theta^T x)$

$g(z) = 1\{z \geq 0\}$ (note $y \in \{0,1\}$)

$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ (training set) $\quad (x^{(i)}, y^{(i)}) \sim \mathcal{D}$

Training error of h_θ:

$$\hat{\varepsilon}(h_\theta) = \hat{\varepsilon}_S(h_\theta) = \frac{1}{m} \sum_{i=1}^{m} 1\{h_\theta(x^{(i)}) \neq y^{(i)}\}$$

ERM: $\hat{\theta} = \arg\min_\theta \hat{\varepsilon}_S(h_\theta)$ ← SVM. logistic regression. is approx. of ERM.
(ERM is more general)

Hypothesis class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^{n+1}\}$.

$h_\theta : X \mapsto \{0,1\}$.

ERM: $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}_S(h)$

Generalization error:

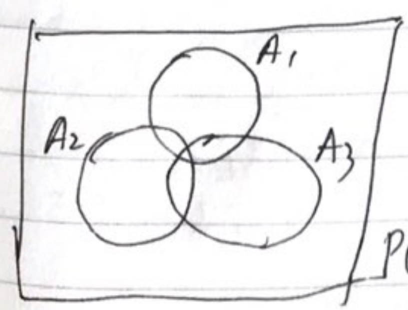
$$\varepsilon(h) = P_{(x,y) \in \mathcal{D}}(h(x) \neq y).$$

Union Bound: let $A_1\ A_2 \dots A_k$ be k event.
(not necessarily independent)

Then

$$P(A_1 \cup A_2 \cup \dots \cup A_k) \leq P(A_1) + P(A_2) + \dots + P(A_k)$$

"or"



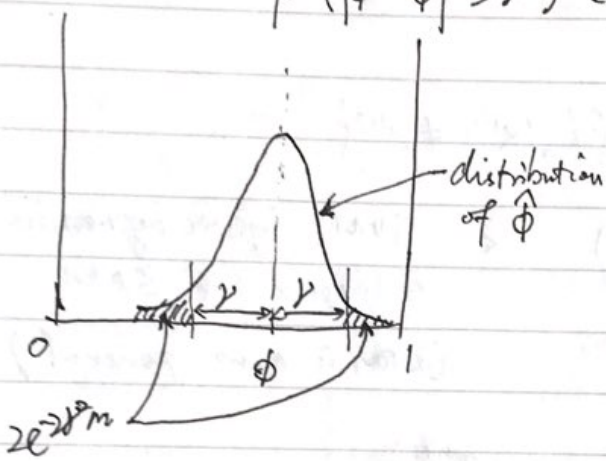
$P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(A_2) + P(A_3)$

Hoeffding Inequality: Let $z_1 \dots z_m$ be ~ I.I.D

Bernoulli(ϕ) random variable ($P(z_i = 1) = \phi$).

Let $\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} z_i$ and let any $\gamma > 0$ be fixed.

Then $P(|\hat{\phi} - \phi| > \gamma) \le 2e^{-2\gamma^2 m}$



Central Limit Theorem: works only m is large.

The case of finite $\mathcal{H}$

$\mathcal{H} = \{h_1, h_2, \dots h_K\}$ K hypothesis

$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}_S(h_i)$

Strategy: (1) $\hat{\varepsilon} \approx \varepsilon$

(2) Show bound on $\varepsilon(\hat{h})$.

(1) Fix any $h_j \in \mathcal{H}$.

Define $Z_i = 1\{h_j(x^{(i)}) \neq y^{(i)}\} \in \{0,1\} \sim$ Bernoulli

$P(Z_i = 1) = \varepsilon(h_j)$. (All $Z_i \sim$ I.I.D)

$\hat{\varepsilon}(h_j) = \underbrace{\frac{1}{m}\sum_{i=1}^{m} Z_i}_{\text{mean } \varepsilon(h_j)} = \frac{1}{m}\sum_{i=1}^{m} 1\{h_j(x^{(i)}) \neq y^{(i)}\}$

By Hoeffding Ineq:

$$P(|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| \geq \gamma) = 2e^{-2\gamma^2 m}.$$

A_j = event that $|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma$.

$P(A_j) \leq 2e^{-2\gamma^2 m}$.

$P(\exists\, h_j \in \mathcal{H} \text{ st. } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma)$

$= P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq \sum_{i=1}^{k} P(A_i) \leq \sum_{i=1}^{k} 2e^{-2\gamma^2 m} = 2ke^{-2\gamma^2 m}$.

(1 − both sides).

$P(\nexists\, h_j \in \mathcal{H} \text{ st. } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma)$

$= P(\forall h_j \in \mathcal{H} \text{ st. } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| < \gamma) \geq 1 - 2ke^{-2\gamma^2 m}$

So w.p. $1 - 2ke^{-2\gamma^2 m}$, $\hat{\varepsilon}(h)$ will be within γ of $\varepsilon(h)$ for all $h \in \mathcal{H}$.

"with probability"

"Uniform Convergence"

Given γ and δ, what is m?

$\delta = 2K e^{-2\gamma^2 m}$, solve for m.

So long as $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$,

then w.p. $1-\delta$, we have $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ for all $h \in \mathcal{H}$

"Sample complexity" bound.

" $\forall k$, $\log k \leq 30$
in practical

"Error bound"

Solve for γ for fixed m, δ.

w.p. $1-\delta$, we have that $\forall h \in \mathcal{H}$.

$$|\hat{\varepsilon}(h) - \varepsilon(h)| \leq \underbrace{\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}}_{\text{"}\gamma\text{"}}$$

---

(2) Let assume $\forall h \in \mathcal{H}$. $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$. ①

Can we prove something about $\varepsilon(\hat{h})$.

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h) \qquad ②$$

$$h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h) \qquad ③$$

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma && \text{--- by ①} \\ &\leq \hat{\varepsilon}(h^*) + \gamma && \text{--- by ②} \\ &\leq \varepsilon(h^*) + \gamma + \gamma \\ &= \varepsilon(h^*) + 2\gamma. \end{aligned}$$

$\hat{\varepsilon}(h)$ - training error of h

$\varepsilon(h)$ - generalization error of h.

Theorem. Let $|\mathcal{H}| = k$ and let m, δ be fixed, then w.p. $1-\delta$

$$\boxed{\varepsilon(\hat{h}) \le \underbrace{\min_{h \in \mathcal{H}} \varepsilon(h)}_{\varepsilon(h^*)} + 2\underbrace{\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}}_{\gamma}} \quad (*)$$

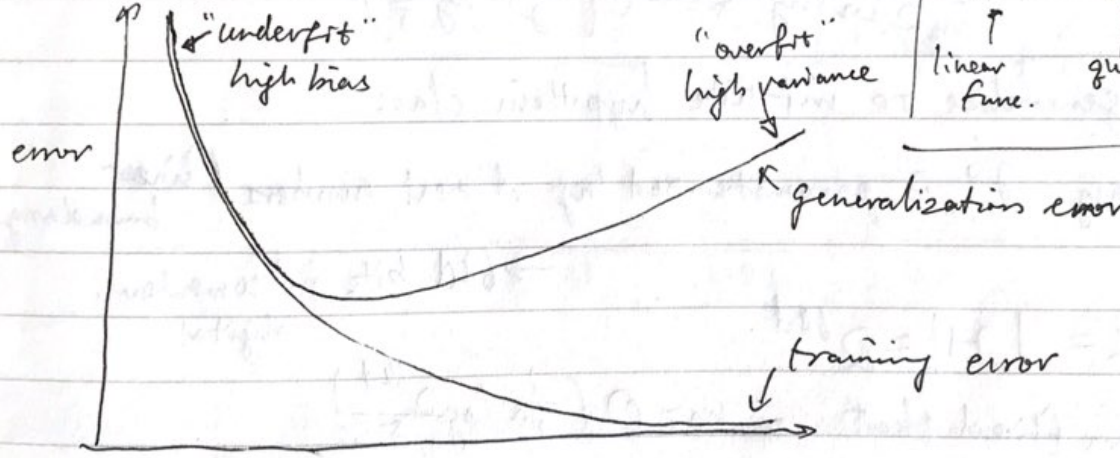
"bias" "variance"

Set $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$, we know ① holds w.p. $1-\delta$. (i.e. $|\varepsilon(h) - \hat{\varepsilon}(h)| \le \gamma \ \forall h \in \mathcal{H}$)

which implies $(*)$

$\mathcal{H} \subseteq \mathcal{H}'$

linear func. quadratic func.



model Complexity (Degree of polynomial. size of $\mathcal{H}$ etc.)
(τ in locally-weighted ~~linear~~ linear regression)

Corollary: Let $|\mathcal{H}| = k$, let any δ, γ be fixed.

Then form
$$\varepsilon(\hat{h}) \le \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$$

w.p. $1-\delta$, it suffices that

$$m \ge \frac{1}{2\gamma^2}\log\frac{2k}{\delta} = O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right)$$ ← Important when generalize to infinite hypothesis class $\mathcal{H}$.

Lecture 10

VC dimension
Model Selection
- Cross validation
- feature selection
Bayesian statistics & Regularization

Let $|\mathcal{H}|=k$, and let γ, δ be fixed.
Then for $\varepsilon(\hat{h}) \le \min_{h\in\mathcal{H}} \varepsilon(h) + 2\gamma$ w.p. $1-\delta$,
it suffices that

$$m \ge \frac{1}{2\gamma^2} \log \frac{2k}{\delta} = O\left(\frac{1}{\gamma^2} \log\frac{k}{\delta}\right)$$

* generalize to infinite hypothesis class:

Say $\mathcal{H}$ is parameterized by d real numbers (linear boundary
→ 64d bits in digital computer.

$k = |\mathcal{H}| = 2^{64d}$.

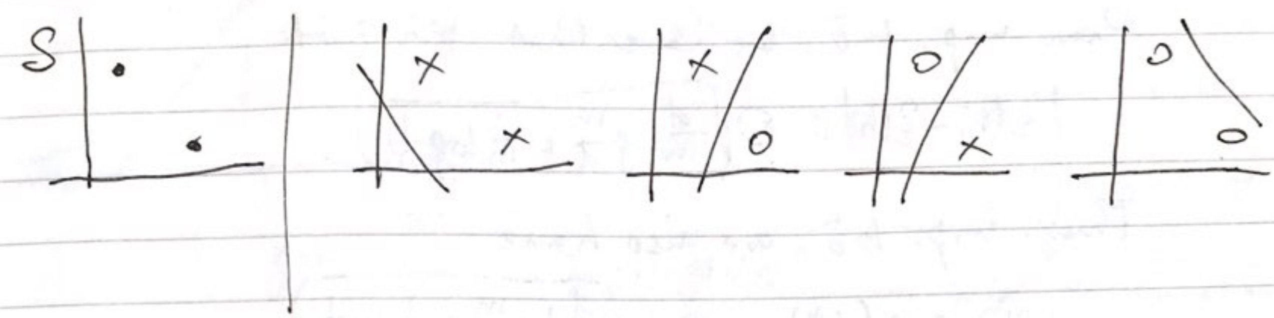
Sufficed that: $m \ge O\left(\frac{1}{\gamma^2} \log\frac{2^{64d}}{\delta}\right)$

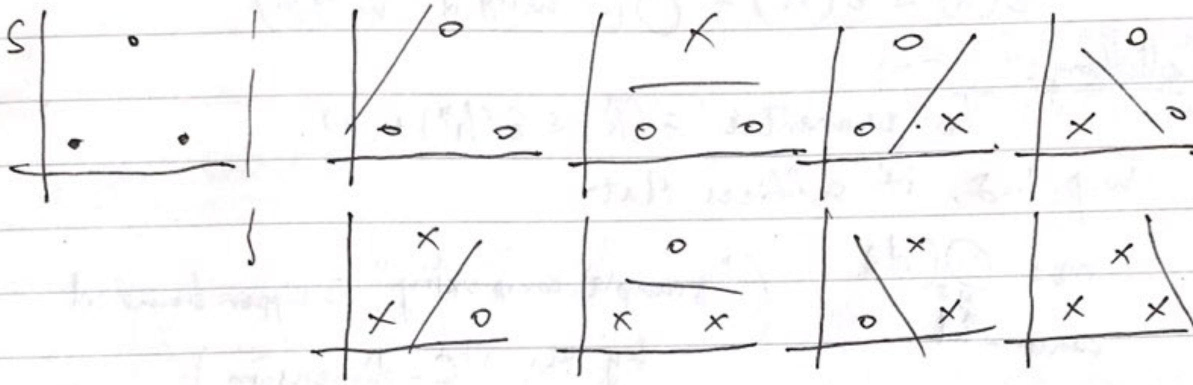
$$= O\left(\frac{d}{\gamma^2} \log\frac{1}{\delta}\right).$$

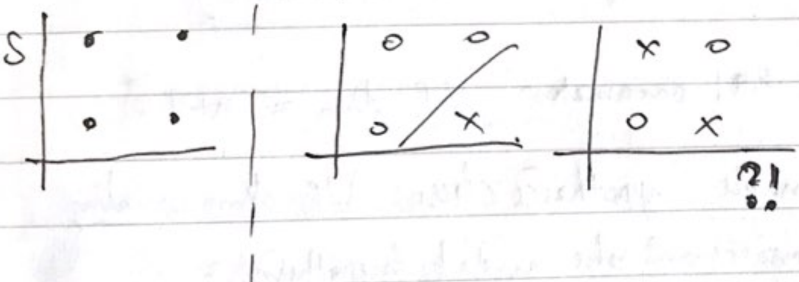
intuitively

Definition: Given a set $S = \{x^{(1)} \ldots x^{(d)}\}$. more formal
we say $\mathcal{H}$ shatters S if $\mathcal{H}$ can realize any labeling
on it.

$\mathcal{H} = \{\text{linear classifiers in 2D}\}$





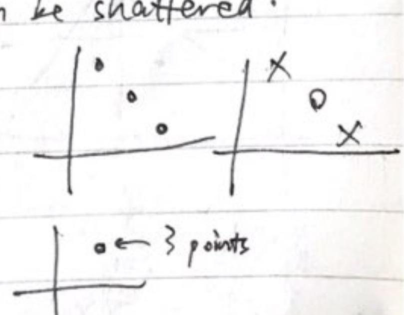


Definition: The Vapnik-Chervonenkis dimension of $\mathcal{H}$ (i.e. $VC(\mathcal{H})$) is the size of the largest set shattered by $\mathcal{H}$

E.g. $\mathcal{H} = \{\text{linear classifiers in 2D}\}$.

$VC(\mathcal{H}) = 3$.

* Only need exist one set of 3 points can be shattered.



More generally, in n dimension,

$VC(\{\text{linear classifiers in } n \text{ dim}\}) = n+1$.

Theorem. Let $\mathcal{H}$ be given and let $VC(\mathcal{H}) = d$.
then w.p. $1-\delta$, we have that $\forall h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \le O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

Thus, w.p. $1-\delta$, we also have

$$\varepsilon(\hat{h}) \le \varepsilon(h^*) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

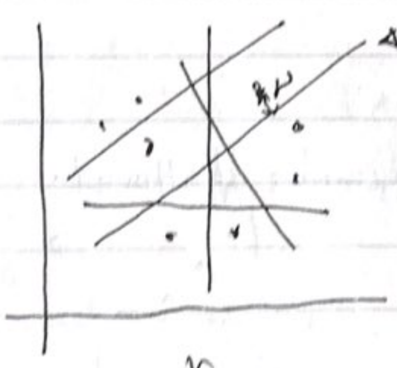
Collollary: To guarantee $\varepsilon(\hat{h}) \le \varepsilon(h^*) + 2\gamma$,
w.p. $1-\delta$, it suffices that

$m = O_{\gamma,\delta}(d)$ ("sample complexity" is upper bounded by the VC-dimension)

const

E.g. logistic, $n+1$ parameter VC dim $= n+1$.

* For most reasonable hypothesis classes, VC-dim usually linear in # parameters of the model/hypothesis.
* Class of linear seperaters with large margin actually has low VC-dim.



If $\|x^{(i)}\|_2 \le R$.

$VC(\mathcal{H}) \le \left\lceil \frac{R^2}{4\gamma^2} \right\rceil + 1$.

SVM automatically find low VC-dim classifier.

$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$

$\|x\|_2^2 = \sum_{i=1}^{\infty} x_i^2$ (converge condition).