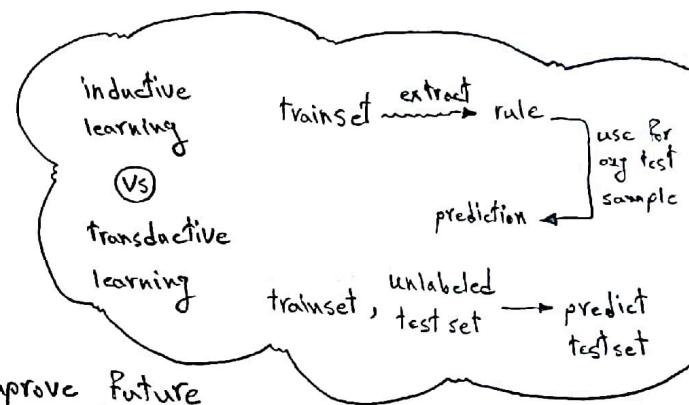


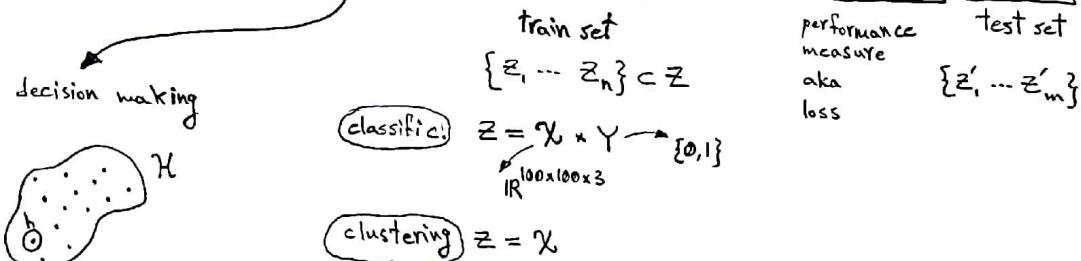
- Mohri → chapters 2–5, 10

- Shai Ben David (2nd resource)

- papers →
 - distributional robustness
 - distribution learning
 - neural tangent kernel (NTK)
 - norm based generalization bounds



Machine learning: learning from experience in order to improve future



Hypothesis set

$$|\mathcal{H}| = \text{IR}^5 \leftarrow \text{Gaus}^2 \text{ or } \text{GMM} \text{ like}$$

$$|\mathcal{H}| = \text{IR}^m \leftarrow \text{multi } m \text{ like NN like}$$



خوب و نیک trainset با learner

$$A : \bigcup_{i=1}^{\infty} z_i \rightarrow \mathcal{H}$$

$$S \rightarrow \overbrace{z_1, z_2, \dots, z_n}^{\text{train}} \overbrace{z_{n+1}, \dots}^{\text{test}}$$

stochastic process

as iid process S با S با \otimes

از یک توزیع fixed داشته باشد.

این توزیع درست نیست ولی ساده است

بمنای این توزیع تئوری خوب نمی‌باشد \otimes

نهض منظر بالا

نهض منظر بالا \otimes برای یک جامعه داده داده داده در انتشار میداد و به مرور مدت از آنها در trainset داشت \leftarrow online learning.

نهض منظر بالا \otimes ایست سلسله آزمون توزیع شده عوامل \leftarrow RL.

نهض منظر بالا \otimes توزیع شده از آنها داشته باشد \leftarrow adversarial.

نهض منظر بالا \otimes داریم که لیل های محدود داشت \leftarrow Oracle باشد و با درجه مون محدود

زیر مجموعه ممکن مالات $\mathcal{M}(z)$ observable space
 $(\mathcal{X} \times \mathcal{Y})$

$$z_1, \dots, z_n \mid z_{n+1}, \dots, z_{n+n_{test}}$$

iid $P \in \mathcal{M}(z)$

statistical risk

$$\overbrace{R(h)}^{\text{deterministic function}} = \min_{h \in \mathcal{H}} \mathbb{E}_P [l_h(z)]$$

$$\lim_{n_{test} \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{n_{test}} l_h(z_{i+n}) \stackrel{\text{a.s.}}{=} \mathbb{E}_P [l_h(z)] \xrightarrow[\text{goal}]{\text{ultimate}} h^* = \arg \min_{h \in \mathcal{H}} \mathbb{E}_P [l_h(z)]$$

ما نیز در حالت h ای دلایل تردیک loss کنیم $\hat{R}_n(h)$ باشیم

We would like to have:

$$R(h^*) \approx \hat{R}(h^*) \approx \hat{R}_n(h^*)$$

R.V. based on z_i

$$\left\{ \begin{array}{l} \hat{R}_n(h) \triangleq \frac{1}{n} \sum_{i=1}^n l_h(z_i) \\ \text{empirical risk} \end{array} \right.$$

$$\hat{h}^* \leftarrow A(z_1, \dots, z_n)$$

Empirical Risk Minimization (ERM)

$$\hat{h}^* = \arg \min_{h \in \mathcal{H}} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n l_h(z_i)}_{\hat{R}_n(h)} \right\}$$

$$\min_{h \in \mathcal{H}} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n l_h(z_i) \quad \text{intractable}$$

برای هر $h \in \mathcal{H}$ در مجموعه \mathcal{Z} برای n نمونه های z_1, \dots, z_n برای h داشتیم و $l_h(z_i)$ loss است و $\hat{R}_n(h)$ loss است

$$L_{\text{test}} = \frac{1}{2}, L_{\text{train}} = 0 \rightarrow \text{که}$$

$$\lim_{n \rightarrow \infty} \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l_h(z_i) \quad \text{با اینکه } n \rightarrow \infty \text{ می خواهیم دادیم}$$

$\hat{R}_n(\hat{h}_n^*)$

و باز بله داشت این \hat{h}_n^* یک h است که $\hat{R}_n(h)$ loss است

و عبارت دیگر هر دفعه که داریم $\hat{R}_n(\hat{h}_n^*)$ loss است

و \hat{h}_n^* را که $\hat{R}_n(\hat{h}_n^*)$ loss است را درد نمی خواهیم داشت

که در \mathcal{H} داشتیم n مولای عدد بزرگ است پس $\hat{R}_n(\hat{h}_n^*)$ loss است

statistics :: estimation

$$\text{estimator } \hat{\theta}(x_1, \dots, x_n) \triangleq \frac{1}{n} \sum x_i$$

$\hat{\theta} \rightsquigarrow \text{R.V.}$

$\theta \rightsquigarrow \text{Fixed but unknown}$

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

$$\text{Variance}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})^2]$$

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta})^2] = \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})^2]}_{\text{variance}(\hat{\theta})} + \underbrace{(\theta - \mathbb{E}\hat{\theta})^2}_{\text{bias}(\hat{\theta})^2} + \underbrace{\mathbb{E}[(\theta - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})]}_{0}$$

Cramer Rao Bound

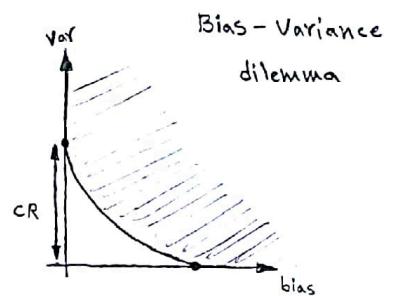
$\hat{\theta} : X^n \rightarrow \Theta$ (fixed n)

if $E\hat{\theta} = \theta$ (unbiased)

then $\text{Var}(\hat{\theta}) \geq \frac{1}{n I(\theta)}$ Fisher information

more general
if $E\hat{\theta} = \theta + \text{bias}(\theta)$

then $\text{Var}(\hat{\theta}) \geq \frac{(1 + \text{bias}'(\theta))^2}{n I(\theta)}$



دستیاری از جای احتمالی p داشت \oplus

Cramér Rao $\frac{1}{n I(\theta)}$

که $\text{MSE}(\hat{\theta}) \geq \text{Var}(\hat{\theta}) + \text{bias}^2(\theta)$ $\rightarrow E[(\hat{\theta} - \theta)^2] \geq \text{bias}^2(\theta) + \frac{(1 + \text{bias}'(\theta))^2}{n I(\theta)}$

بر حسب bias

: $\theta = h$ ل آندر

آندر نظر داشت \oplus
آندر نظر داشت \oplus
 $\hat{\theta} = h = c$ آندر \oplus

deterministic

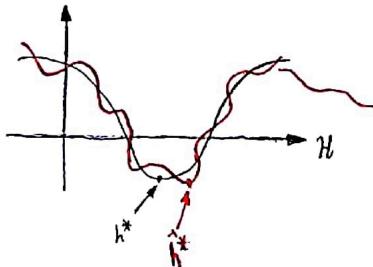
$$R(h) = \mathbb{E}_p[\ell_h(z)] \quad \forall h \in \mathcal{H}$$

ERM

$$\hat{h}^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_n(h)$$

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell_h(z_i) \quad z_1, \dots, z_n \sim p$$

R.V.



$$\hat{R}(\hat{h}^*) \rightarrow 0 \text{ or min}$$

bias reduction

$$R(\hat{h}^*) \leq \hat{R}(\hat{h}^*) + \xi$$

generalization

we want this to hold
with high probability

Concentration Bound

چون دیگر هر تابعی هست که می خواهد این را
گویند به عنوان نتیجه از generalization
که گویند به عنوان نتیجه از concentration bound
که گویند به عنوان نتیجه از generalization

$$\mathbb{E}_p[\ell_h(z)] \xrightarrow{\text{Markov}} \underbrace{\frac{1}{n} \sum_{i=1}^n \ell_h(z_i)}_{\mathbb{E}[S_n]} \xrightarrow{\text{R.V. } S_n} \frac{1}{n} \sum_{i=1}^n \ell_h(z_i)$$

Markov

- $X \geq 0 \rightarrow \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad \forall a > 0$
- $\mathbb{E}[X] = \mu$

$$\text{proof: } \mathbb{E}[X] = \int_0^\infty n f_X(n) d n = \underbrace{\int_0^a n f_X(n) d n}_{\geq 0} + \underbrace{\int_a^\infty n f_X(n) d n}_{\geq a \int_a^\infty f_X(n) d n} \geq a \mathbb{P}(X \geq a)$$

Chebyshev

- $\mathbb{E}[X] = \mu$
- $\text{Var}[X] = \sigma^2 \rightarrow \mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \quad \forall \varepsilon > 0$

$$\text{proof: } T \triangleq \left(\frac{X - \mu}{\sigma} \right)^2 \rightarrow T \geq 0 \xrightarrow{\mathbb{E}[T] = 1} \mathbb{P}\left(\left(\frac{X - \mu}{\sigma}\right)^2 \geq \varepsilon\right) \leq \frac{1}{\varepsilon}$$

$\lambda_h(z)$ $\xrightarrow{\text{E}} \checkmark$
 $\text{Var} \checkmark$ کرشن sample $\forall h \in \mathcal{H}$ اون ووت اگر $\lambda_h(z_i)$ ما هم جمل میشن

chebyshev for $\frac{1}{n} \sum_{i=1}^n \lambda_h(z_i)$ $\rightarrow \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \lambda_h(z_i) - \mathbb{E}_p[\lambda_h(z)] \right| \geq \varepsilon \right) \leq \underbrace{\frac{\sigma^2(\lambda_h)}{n \varepsilon^2}}_{\delta} \rightarrow \varepsilon = \frac{\sigma(\lambda_h)}{\sqrt{n \delta}}$

$\rightarrow \forall h \in \mathcal{H}$ with probability of at least $1-\delta$:

$$\left| \frac{1}{n} \sum_{i=1}^n \lambda_h(z_i) - \mathbb{E}_p[\lambda_h(z)] \right| \leq \frac{B}{2\sqrt{n\delta}}$$

$$0 \leq \lambda_h(z) \leq B \rightarrow \sigma(\lambda_h) \leq \frac{B}{2}$$

$$\forall h \in \mathcal{H} \quad \text{حالت } \frac{B}{2} \text{ میتوان بدیهی}$$

$$\mathbb{P}(z > t) \leq 2e^{-\frac{t^2}{2}} \quad \leftarrow \text{ایم } z \sim N(0,1) \text{ برای}$$

$\underset{n \rightarrow \infty}{\lim} \mathbb{P}(|\text{emp.} - \text{stat.}| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{2}}$ دست و دست یک تعداد بسیار کم، مجموعه میتوان در نظر گرفت tight میتواند chebyshev, markov گلوبال باشد

Hoeffding for a set of random variables

- $X_1, \dots, X_n \stackrel{iid}{\sim} p$
- $X_i \perp\!\!\!\perp X_j$ $\rightarrow \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X \right| \geq \varepsilon \right) \leq \underbrace{2e^{\frac{-2n\varepsilon^2}{B^2}}}_{\delta} \rightarrow \varepsilon = B \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$
- $0 \leq X_i \leq B$

$\rightarrow \forall h \in \mathcal{H}$ with probability of at least $1-\delta$:

$$\left| \frac{1}{n} \sum_{i=1}^n \lambda_h(z_i) - \mathbb{E}_p[\lambda_h(z_i)] \right| \leq B \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Mc Diarmid

- $f: \mathbb{R}^n \rightarrow \mathbb{R}$
- x_1, \dots, x_n i.i.d.

$$\Pr\left(\left|f(x_1, \dots, x_n) - \mathbb{E}[f(x_1, \dots, x_n)]\right| \geq \varepsilon\right) \leq 2e^{-\frac{2\varepsilon^2}{\sum c_i^2}}$$

- $\forall i \quad x_i \rightarrow x_{i-1}, x_{i+1}, \dots, x_n$ fix
 x_i change
- max diff = $c_i \geq 0$ (Bounded Difference property for f)

لما میکنیم

$$f(z_1, \dots, z_n) = \frac{1}{n} \sum_{i=1}^n \lambda_h(z_i)$$

$$0 \leq \lambda_h(z) \leq B \quad \rightarrow \text{max diff} = c_i = \frac{B}{n} \rightarrow \text{McDiarmid's Hoeffding bound}$$

برای سانحه این که بجای یک داده h برای همه h ها درست نشوند

h_1 fix $z_1, \dots, z_n \stackrel{\text{i.i.d}}{\sim} p$ $\Pr\left(\underbrace{|R_n(h_1) - R(h_1)|}_{A} \leq \dots\right) \geq 1-\delta$

h_2 fix $z'_1, \dots, z'_n \stackrel{\text{i.i.d}}{\sim} p$ $\Pr\left(\underbrace{|R_n(h_2) - R(h_2)|}_{B} \leq \dots\right) \geq 1-\delta$

$$\Pr(A, B) \geq (1-\delta)^2 \leftarrow A \perp\!\!\!\perp B \leftarrow \text{دو دیاست متمم اند} h_2, h_1$$

حالا h_1, h_2 جفت‌سون از یک دیاست باید ببینیم

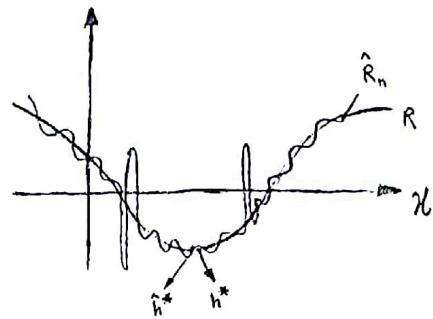
$$\Pr(A, B) \geq 1-2\delta + \underbrace{\delta^2}_{0} \leftarrow$$

وقتی می‌خواهیم ازای n ترکیب R , R_n که h_i ها ناید خرابش کن پس $\Pr(\bar{A} \cup \bar{B} \cup \dots)$ $\Pr(\bar{A})$ و $\Pr(\bar{B})$ را ترکیب کنیم، همچنان که $(\mathcal{H}_1, \mathcal{H}_2, \dots)$ uniform convergence for finite \mathcal{H} باشد

$$\Pr\left(\bigcup_{i=1}^m \bar{A}_i\right) \leq \sum_{i=1}^m \underbrace{\Pr(\bar{A}_i)}_{\leq \delta} \leq m\delta \rightarrow \Pr\left(\bigcap_{i=1}^m A_i\right) \geq 1-m\delta$$

point wise convergence

- Hypothesis set: $\mathcal{H} \subseteq Y^X$
- data distribution: $p \in M(Z)$
- Loss Function: $\ell: Y \times Y \rightarrow \mathbb{R}$



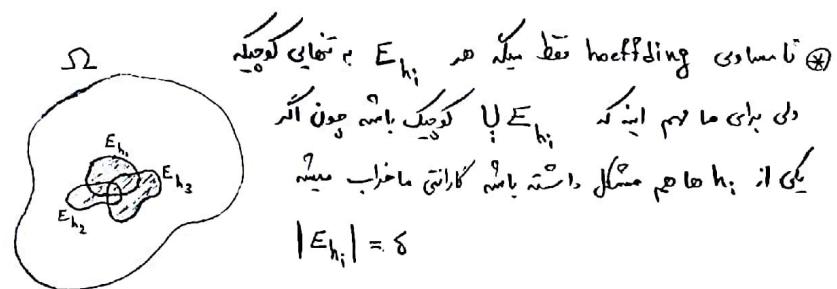
if $\ell(\cdot)$ is B -bounded for some $B > 0$

Hoeffding or McDiarmid

$$\forall h \in \mathcal{H} \text{ if } z_1, \dots, z_n \stackrel{iid}{\sim} p \text{ with probability of at least } 1-\delta \quad \Rightarrow \quad |R(h) - \hat{R}_n(h)| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \times B$$

for $\forall \delta > 0$

⊗ مه میز ب تله خوب میداد اما با این روش h^* و \hat{h}^* تقریب بزنم نه! و!
 چون در عمل ما اول z_i ها رو سپل کنیم و بعد ℓ انتگرال کنیم و اول h رو نکس کنیم بودم
 پس کلاً این نامساوی بقرار نیست ← چون روی یک تنه خاص که \hat{h}^* باشد نظر نمی خورد
 تقریم سرانه ایک با ازای هم ایک converges not quick باشیم.



Uniform Convergence

نیویورت

~~$\forall h \in \mathcal{H}$~~ if $z_1, \dots, z_n \stackrel{iid}{\sim} p$

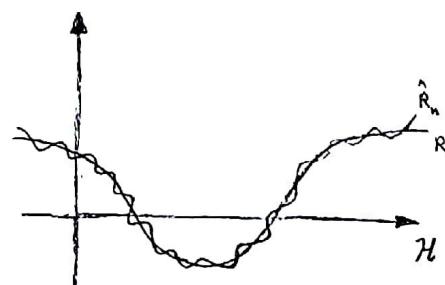
$$P\left(\max_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)| \leq B \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right) \geq 1-\delta$$

⊗ کران بزرگتر شد اما به جانش دریم که ازای هر h رو E_h محدود
 باشد $|h|$ با finite

$\hat{R}(\hat{h}^*)$ is small

$\rightarrow R(\hat{h}^*)$ is also small

\rightarrow is close to $R(h^*)$ and $\hat{R}_n(h^*)$



CDF of a R.V.

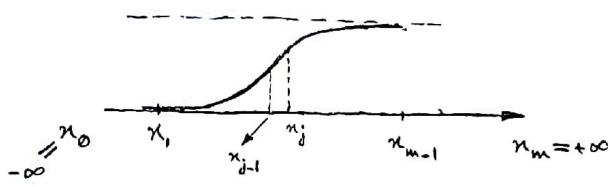
$$F(x) = \mathbb{P}(X \leq x)$$

$x \equiv h$, $F \equiv R$ در اینجا \oplus

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x)$$

pointwise converges \hat{F}_n برای
 رودازی $\left\{ \begin{array}{l} F(x) = \mathbb{E} \hat{F}_n(x) \end{array} \right.$

\oplus برای uniform convergence بطور کلی هر دو اینجا دوی $|F_n - F|$ محدود است نهایی نظر نیاد
محدودیت n به نظر نیاد این آنست بخوبی بینه!



\oplus x ها به مقدار انتخاب می‌شوند و دوی خود را بازه‌ها پاس uniform

$\forall x \in \mathbb{R}$
 \hookrightarrow def. $j \in \{1, \dots, m\}$ such that $x \in [x_{j-1}, x_j]$ $m \in \mathbb{N}$

$$\begin{cases} \hat{F}_n(x) - F(x) \leq \hat{F}_n(x_j) - F(x_{j-1}) = \hat{F}_n(x_j) - F(x_j) + \frac{1}{m} \\ \hat{F}_n(x) - F(x) \geq \hat{F}_n(x_{j-1}) - F(x_j) = \hat{F}_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m} \end{cases}$$

$$\rightarrow \left| \hat{F}_n(x) - F(x) \right| \leq \max_{j=1 \dots m} \left| \hat{F}_n(x_j) - F(x_j) \right| + \frac{1}{m}$$

\oplus از سر $|h|$ خلاص شده و به m انداد

\oplus درون مثلاً \sqrt{m} نیز ممکن و کافی بر اساس

تعادل بین m و \sqrt{m}

$$\rightarrow \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \leq \max_{j=1 \dots m} \left| \hat{F}_n(x_j) - F(x_j) \right| + \frac{1}{m}$$

deterministic

\oplus اینجا چون بازه‌های x_i را بدلی به x داشته باشیم در عین مرد
نیست که داشتن داریم $R(\hat{h}^*)$ دو $\sup_{x \in \mathbb{R}}$ دوی را داشتیم \oplus

$$\rightarrow \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

Glivenko – Cantelli Theorem

$\forall \varepsilon > 0$ choose $m \in \mathbb{N}$ such that $\frac{1}{m} < \varepsilon$

$$\forall j \quad P\left(\left|\hat{F}_n - F\right| > \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right) < \delta$$

$$\xrightarrow{\text{union bound}} P\left(\max_{j=1 \dots m} \left|\hat{F}_n(x_j) - F(x_j)\right| > \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right) < m\delta$$

حالات حدايیم برای سرت راست نامهای deterministic
عند تبلیغ بازدید کنیم

جون مساب کردن $\max_{j=1 \dots m} |\square|$ $\xrightarrow{\text{rename}} \delta \rightarrow \frac{\delta}{m}$

$$\sup_n \left| \hat{F}_n(x) - F(x) \right| \leq \frac{1}{m} + \sqrt{\frac{\log \frac{2m}{\delta}}{2n}}$$

$$\xrightarrow{\text{پس از } \sqrt{n} \text{ هم } n \text{ داشت، } m \text{ شد}} O\left(\frac{1}{\sqrt{n}}\right) + O\left(\sqrt{\frac{\log \frac{n}{\delta}}{n}}\right) = O\left(\sqrt{\frac{\log \frac{n}{\delta}}{n}}\right)$$

pointwise convergence

uniform convergence

for finite
H $\forall h \in \mathcal{H}$ $z_1, \dots, z_n \stackrel{iid}{\sim} p$ if $z_1, \dots, z_n \stackrel{iid}{\sim} p$ $\forall h \in \mathcal{H}$

$$\text{then } P(|\hat{R}_n(h) - R(h)| < \epsilon_{\delta, n}) \geq 1 - \delta$$

$$\forall \delta > 0 \quad \lim_{n \rightarrow \infty} \epsilon_{\delta, n} = 0$$

$$P(|\hat{R}_n(h) - R(h)| < \eta_{\delta, n}) \geq 1 - \delta$$

$$\forall \delta > 0 \quad \lim_{n \rightarrow \infty} \eta_{\delta, n} = 0$$

$$\equiv P\left(\sup_{h \in \mathcal{H}} |\hat{R}_n(h) - R(h)| < \eta_{\delta, n}\right) \geq 1 - \delta$$

• $\exists \delta > 0, \lim_{n \rightarrow \infty} \min_{h \in \mathcal{H}} \hat{R}_n(h) \geq R(h)$ \Rightarrow uniform conv. \mathcal{H} \oplus
 (الآن إن $\hat{R}_n(h) \geq R(h)$ $\forall h \in \mathcal{H}$ \Rightarrow $\min_{h \in \mathcal{H}} \hat{R}_n(h) \geq \min_{h \in \mathcal{H}} R(h)$)
 \Rightarrow صورة \mathcal{H} محوظة \Rightarrow $\min_{h \in \mathcal{H}} \hat{R}_n(h) \geq \min_{h \in \mathcal{H}} R(h)$

($\min_{h \in \mathcal{H}} \hat{R}_n(h) \geq \min_{h \in \mathcal{H}} R(h)$). \Rightarrow uniform conv. \mathcal{H} \oplus learnable \mathcal{H} \oplus

Mohri's chapter 2 $\xrightarrow{\text{finite } \mathcal{H}}$ pAC-Learnability of \mathcal{H}

إذا $\exists \delta > 0, \forall n \in \mathbb{N}, \forall \epsilon > 0$ $\exists N \in \mathbb{N}$ $\forall n \geq N$ $\forall h \in \mathcal{H}$ $R(h) \leq \hat{R}_n(h) \leq R(h) + \epsilon$ \Rightarrow uniform conv.

probably approximately correct

For any $p \in M(Z)$ if $z_1, \dots, z_n \stackrel{iid}{\sim} p$, $\forall \epsilon, \delta > 0$

$\xrightarrow{x \times y}$ approx probable
 \hat{R}_n $\xrightarrow{\min_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)| < \epsilon}$ \downarrow

there exists A such that $\hat{h}^* = A(z_1, \dots, z_n)$ if $n \geq \text{poly}(\frac{1}{\delta}, \frac{1}{\epsilon})$

$$\Rightarrow P\left(\left|R(\hat{h}^*) - \min_{h \in \mathcal{H}} R(h)\right| < \epsilon\right) \geq 1 - \delta$$

if it holds
we say \mathcal{H} is
Agnostic pAC Learnable

دلیل اینه بوسیله آگنوتیک است
 اگر \mathcal{H} دارای Bayes classifier باشد \Rightarrow آگنوتیک است
 با \mathcal{H} ت می‌توانی بگویی، بنابراین

= agnostic PAC learnable alg. $\mathcal{H} \subseteq A$ \leftarrow only B -bounded $\|w\|_B$, $|w| < \infty$ $\forall w$

uniform conv. for $|H| < \infty$

$$\sup_{h \in \mathcal{H}} |\hat{R}_n(h) - R(h)| \leq B \sqrt{\frac{\log \frac{2}{\delta} + \log |\mathcal{H}|}{2n}} \leq \frac{\varepsilon}{2} \rightarrow n \geq O\left(\frac{\log \frac{2}{\delta} + \log |\mathcal{H}|}{\varepsilon^2}\right) \times B^2$$

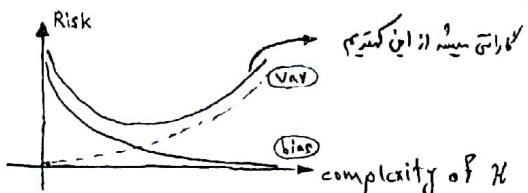
$\text{poly}(\frac{1}{\delta}, \frac{1}{\varepsilon})$

↓

$$\hat{R}_n(h^*) \leq \min_{h \in \mathcal{H}} R(h) + O\left(B \sqrt{\frac{\log \frac{2}{\delta} + \log |\mathcal{H}|}{2n}}\right)$$

~~$\hat{R}_n(h^*)$~~ (bias) (var)

structural Risk minimization (SRM alg.)



$$R(h) = \mathbb{E}_{p(x)} \left[\mathbb{1}(y \neq h(x)) \right]$$

$$\text{Bayes classifier} \rightarrow h^*(x) = \operatorname{argmax}_y p(y|x)$$

جایی داریم و هر جایی $p(y|x)$ label noise \oplus
باشد $\rightarrow p(y|x)$ توزیع determin. باشد

Agnostic PAC \rightarrow PAC : \oplus (realizable setting) only if Bayes classifier \rightarrow \oplus

مالا مطابق تبدیل های پیشتری بودیم ولی باز همین پیش
توزیع classification اگر به لحاظ PAC بود استفاده کرد \rightarrow classification \oplus است اگرچه agnostic PAC \oplus

For any $p \in M(X)$, $h^* \in \mathcal{H}$

$$\text{if } x_1, \dots, x_n \stackrel{iid}{\sim} p \rightarrow (x_1, j_1), \dots, (x_n, j_n)$$

$$\forall \varepsilon, \delta > 0 \text{ if } n \geq \text{poly}(\frac{1}{\delta}, \frac{1}{\varepsilon}) \text{, } \hat{h}^* = A(z_1, \dots, z_n)$$

$$\rightarrow \mathbb{P}(R(\hat{h}^*) \leq \varepsilon) \geq 1 - \delta$$

$$\min_{h \in \mathcal{H}} R(h)$$

Bayes classifier $\in \mathcal{H}$

If it holds,
we say \mathcal{H} is
PAC Learnable

$$\min_{h \in \mathcal{H}} R(h) = 0$$

classification loss = zero-one loss

Zero-one loss & $h^* \in \mathcal{H}$:

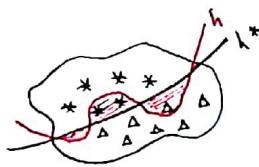
$$\hat{h}_{\text{ERM}}^* \triangleq \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}(y_i \neq h(x_i))}_{\equiv h^*(x_i) \neq h(x_i)}$$

$$\hat{R}_n(\hat{h}_{\text{ERM}}^*) = 0 \quad \mathbb{P}\left(\exists h \in \mathcal{H} \mid \hat{R}_n(h) = 0, R(h) \geq \varepsilon\right) \leq \delta$$

حالا اگر $|\mathcal{H}| \leq \infty$ و هم داشته باشیم:

$$\delta \leq \sum_{h \in \mathcal{H}} \mathbb{P}(\hat{R}_n(h) = 0 \cap R(h) \geq \varepsilon) \quad p(A \cap B) \leq p(A|B)$$

$$\leq \sum_{h \in \mathcal{H}} \underbrace{\mathbb{P}(\hat{R}_n(h) = 0 \mid R(h) \geq \varepsilon)}_{\leq (1-\varepsilon)^n} \quad \text{len(trainset)}$$



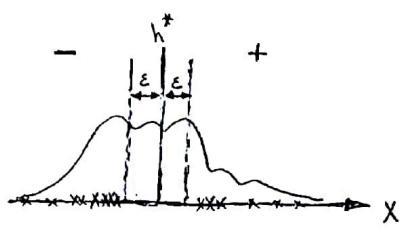
$$\leq |\mathcal{H}| (1-\varepsilon)^n$$

$$\leq |\mathcal{H}| e^{-n\varepsilon}$$

$$\longrightarrow n \geq 0 \left(\frac{\log \frac{1}{\delta} + \log |\mathcal{H}|}{\varepsilon} \right)$$

۴) ممکن است \mathcal{H} بزرگ باشد و p_{AC} آگوستیک باشد. p_{AC} learnable است و p_{AC} bandit ε -unbiased است. p_{AC} bandit ε -unbiased است و p_{AC} learnable است.

یک مثال برای حالات $|\mathcal{H}| = \infty$ داریم. فرض کنیم P و h^* و نیز دو نمونه یک دیاست. n تایی داریم و p_{AC} آگوستیک باشد. p_{AC} bandit ε -unbiased است. p_{AC} learnable است و p_{AC} ε -zero-one loss است.



الگوریتم پیشنهادی \rightarrow هر خط دخواه بین راست ترین سهیل سقی و چپ ترین سهیل مست-

أیت \rightarrow نسبت \hat{h}^* از دو نمونه می شود تغییر نمود، از هر طرف به اندازه ای که بجزء احتمال آن ع عدد از h^* فاصله باشد.

$$\bullet \quad \mathbb{P}\left(\hat{h}^* \text{ از } h^* \text{ نسبت در این بازه } \right) = \mathbb{P}\left(\text{چیزی سهیل در بازه } \text{ ای نباشد}\right) \leq \mathbb{P}\left(\text{چیزی سهیل در بازه } \text{ ای نباشد}\right) = (1-\varepsilon)^n$$

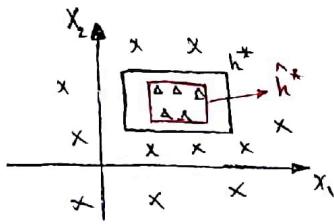
• \hat{h}^* صورت برای هست چه

$$\rightarrow \mathbb{P}\left(\hat{h}^* \text{ از } h^* \text{ نسبت در این بازه }\right) = \mathbb{P}\left(\text{چیزی سهیل در بازه } \text{ ای نباشد}\right) \stackrel{\text{union bound}}{\leq} 2(1-\varepsilon)^n \leq \delta$$

$$R(\hat{h}^*) - R(h^*) \geq \varepsilon$$

$$\rightarrow n \geq \frac{\log \frac{2}{\delta}}{\log \frac{1}{1-\varepsilon}} = O\left(\frac{\log \frac{2}{\delta}}{\varepsilon}\right)$$

یک مدل 2 بعدی



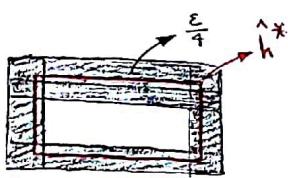
الگوریتم "fix and learn"

۴) دنباله کننده \hat{h}^* نویسنده مسئله باشد به ها و در برگیرد

و مینیمیزهای خطا train صفت خواهد بود

• جرم احتمال $\Pr[h > \hat{h}^*]$ already تست کردار دارد و بیشتر از اون ممکن نیست

• از هر فتح \hat{h} به اندیشه ای میان داخل که جرم $\frac{\varepsilon}{4}$ داشته باشد



$$\Pr(R(\hat{h}^*) > \varepsilon) \leq \Pr\left(\bigcup_{i=1}^n E_i\right)$$

$$\leq \sum_{i=1}^n \Pr(E_i) \leq 4e^{-\frac{n\varepsilon}{4}}$$

$$\rightarrow n \geq \frac{4}{\varepsilon} \log \frac{4}{\delta}$$

• در این مدل های $H = \mathcal{H}$ بازیگردی نسبت به \mathcal{H} دارند چون برای مسئله خاصی حل کردیم

• همینگوری تعداد امثال رو زیاد بگشم هم مینهاد همینگوری اثبات کرد PAC learnable میست. حتی دایره هم آرکیدی.

و PAC learnable دلکه to convex نیست

loss
 Function: $z \rightarrow [0,1]$
 uniform convergence For $\forall g \in \mathcal{G} \subseteq [0,1]^Z$ where $|\mathcal{G}| = \infty$
 $S = \{z_1, \dots, z_n\} \stackrel{iid}{\sim} P$
 $P(\forall g \in \mathcal{G} \rightarrow E_p[g(z)] \leq \frac{1}{n} \sum_{i=1}^n g(z_i) + o(1)) \geq 1 - \delta$
 union bound $\geq \frac{1}{n} \sum_{i=1}^n E_p[g(z_i)] + \text{finite}$
 این تابع ایمنی می‌گیرد PAC , Agnostic PAC

$\Phi(s) = \sup_{g \in \mathcal{G}} E_p(g) - \hat{E}_s(g)$ overfit میزان مدار
 این تابع $\Phi(s)$ را دارد و bounded diff. property دارد و با تئوری $\frac{1}{n}$ تئوری z_j داریم
 بر دلیل وجود محدودیت در انتخاب g که g کنترل شود
 تئوری iid تابع $\Phi(s)$ \rightarrow fano's inequality
 داریم bounded diff. prop. داریم
 McDiarmid

$\Phi(s') = \sup_{g \in \mathcal{G}} E_p(g) - \hat{E}_{s'}(g)$
 $= \frac{1}{n} \sum_{i=1}^n g(z_i) + \frac{g(z_j) - g(z'_j)}{n}$
 $\leq \sup_{g \in \mathcal{G}} \Phi(s) + \sup_{g \in \mathcal{G}} \frac{|g(z_j) - g(z'_j)|}{n}$
 ≈ 1 bounded $g \in \mathcal{G}$

$P(\Phi(s) \leq E_{S \sim P^n}[\Phi(s)] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}) \geq 1 - \delta$ ①

تابع $\Phi(s)$ بر میار خوب باید باشد $\Phi(s)$ learnability

$$\hat{g}^* = A(z_1, \dots, z_n, \varepsilon, \delta)$$

$$\begin{aligned}
 E[\hat{g}^*] - \hat{E}_s[\hat{g}^*] &\leq \sup_{g \in \mathcal{G}} \{E[g] - \hat{E}_s[g]\} \\
 \rightarrow E[\hat{g}^*] &\leq \underbrace{\hat{E}_s[\hat{g}^*]}_{R(h)} + \Phi(s) \xrightarrow{\text{①}} E[\hat{g}^*] \leq \hat{E}_s[\hat{g}^*] + \underbrace{E_{S \sim P^n}[\Phi(s)]}_{\substack{\text{complexity measure} \\ \text{only based on} \\ P, \mathcal{G}, n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}
 \end{aligned}$$

w.h.p

deterministic

$$S, S' \sim \rho^n, S \perp\!\!\!\perp S'$$

$$\mathbb{E}_{S \sim \rho^n} [\hat{\mathbb{E}}_{S'}[g]]$$

⊗ هنون باز جلسه قبل سی دوام قردن اضافه کنیم

$$\mathbb{E}_{S \sim \rho^n} [\Phi(S)] = \mathbb{E}_{S \sim \rho^n} \left[\sup_{g \in \mathcal{G}} \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right]$$

$$= \mathbb{E}_{S \sim \rho^n} \left[\sup_{g \in \mathcal{G}} \mathbb{E}_{S' \sim \rho^n} [\hat{\mathbb{E}}_{S' \sim \rho^n}[g] - \hat{\mathbb{E}}_S[g]] \right]$$

$$\leq \mathbb{E}_{S, S' \sim \rho^n} \left[\sup_{g \in \mathcal{G}} (\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g]) \right]$$

برای
کلی
برای
 \mathbb{E}_S

$$\leq \mathbb{E}_S \mathbb{E}_{S, S' \sim \rho^n} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(z_i) \right]$$

$$+ \mathbb{E}_S \mathbb{E}_{S, S' \sim \rho^n} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(z'_i) \right]$$

$$= 2 \mathbb{E}_{\epsilon, S} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n g(z_i) \epsilon_i \right]$$

Rad_n(g)

پیشگیری از complexity measure کی

$$\text{if } |g| = 1 \rightarrow \text{Rad}_n(g) = \mathbb{E}_{\epsilon, S} \left[\frac{1}{n} \vec{g}_S \cdot \vec{\epsilon} \right]$$

$$= \mathbb{E}_S \left[\frac{1}{n} \vec{g}_S \right] \underbrace{\mathbb{E}_\epsilon [\vec{\epsilon}]}_0 = 0 \quad \checkmark$$

داد نیز داد

$$\lim_{n \rightarrow \infty} \text{Rad}_n(\text{lipschitz functions}) = 0 \rightarrow \text{lipschitz functions are learnable}$$

$$S = \{z_1, \dots, z_n\}$$

$$S' = \{z'_1, \dots, z'_n\} \quad z_i \perp\!\!\!\perp z'_i$$

$$\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g]$$

$$= \frac{1}{n} \sum_{i=1}^n [g(z_i) - g(z'_i)] \epsilon_i$$

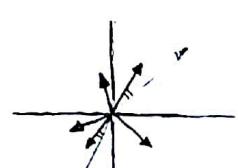
$$\epsilon_i \sim \begin{cases} \frac{1}{2} & +1 \\ \frac{1}{2} & -1 \end{cases} \quad \text{Rademacher R.V.}$$

swap $\frac{1}{2}$ احتمال داشته باشد ϵ_i کوون z_i, z'_i swap کوون z_i, z'_i برداشت ϵ_i دارد

$$\text{Rad}_n(g) = \mathbb{E}_{\epsilon, S} \left[\sup_{g \in \mathcal{G}} \frac{\langle \vec{g}_S, \vec{\epsilon} \rangle}{n} \right]$$

Rademacher complexity = complexity of g

$$\vec{g}_S \triangleq \begin{bmatrix} g(z_1) \\ \vdots \\ g(z_n) \end{bmatrix}$$



$$\vec{\epsilon} \triangleq \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

بردار \vec{g} یک مسیر نویز داشت می‌باشد و $\sup_{g \in \mathcal{G}}$ سعی می‌نماید \vec{g} را ایجاد کند تا \vec{g} باشد، هرچه فضای محدود تر باشد، دست $\sup_{g \in \mathcal{G}}$ باز آن

$$\hat{\text{Rad}}_n(g) \triangleq \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(z_i) \right]$$

McDiarmid \leftarrow obs bounded diff. prop. $\rho \circ U$

$\hat{\text{Rad}}_n(g) \leq \hat{\text{Rad}}_s(g)$ استناداً إلى ρ $\Rightarrow \hat{\text{Rad}}_s(g) \leq \text{Rad}(g)$
بناءً على $\hat{\text{Rad}}_s(g) \leq \text{Rad}(g)$

$$\xrightarrow{\text{جاكواري جاكواري}} \mathbb{E}[\hat{g}^*] \leq \mathbb{E}_s[\hat{g}^*] + 2\hat{\text{Rad}}_s(g) + (2+1)\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

عوالي g لـ Rad \Rightarrow

$$\mathbb{E}g = R(h) \xrightarrow{\text{جاكواري}} R(h) \leq \hat{R}_n(h) + 2\hat{\text{Rad}}_n(g, \rho) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

binary classification
 $h: X \rightarrow \{-1, 1\}$

$$\mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \underbrace{\mathbf{1}(y_i \neq h(x_i))}_{\frac{1 - y_i h(x_i)}{2}} \right]$$

$$= \mathbb{E} \left[\sup_{h \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{\epsilon_i}{2}}_{\text{دونه اداره}} + \underbrace{\sum_{i=1}^n \frac{1}{2n} (-y_i \epsilon_i) h(x_i)}_{\epsilon_i} \right]$$

$\sup_{h \in \mathcal{H}}$ مقدمة \mathbb{E}

$$= \frac{1}{2} \mathbb{E}_{x_1, \dots, x_n \sim P_x} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(x_i) \right] = \frac{1}{2} \text{Rad}(\mathcal{H}, P_x)$$

General

$$\mathbb{E}[g] \leq \hat{E}_n[g] + 2 \mathbb{E}_{z'_1, \dots, z'_n \sim P^n} \left[\frac{1}{n} \sup_{g \in \mathcal{G}} \sum_{i=1}^n \epsilon_i g(z'_i) \right] + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

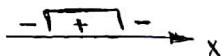
$\text{Rad}(g, P)$

for binary classification

zero-one loss

$$\mathcal{H} \subseteq \{-1, +1\}^X$$

$$R(h) \leq \hat{R}_n(h) + \text{Rad}(\mathcal{H}, P) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

if \mathcal{H} :  $d_{VC}(\mathcal{H}) = 2$ $\begin{cases} \bullet -\infty \rightarrow 1 \\ \bullet 1+n \rightarrow n \\ \bullet -1+n \rightarrow (n-1)+\dots+1 = \binom{n}{2} \end{cases}$

then for n samples we have $1+n+\binom{n}{2} = O(n^2)$ dichotomies

این تعداد حالت شکل \mathcal{H} را labeling $\mathcal{C}_{\mathcal{H}}$ می‌گیرد

$\mathcal{C}_{\mathcal{H}}$ برای یک classifier \mathcal{H} تعداد d بخواهد

Growth Function

(Labeling power of \mathcal{H}) binary classifier \mathcal{H} توانایی \mathcal{H} را برای یک classifier \mathcal{H} تعداد d بخواهد

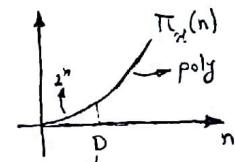
$$\forall h: X \rightarrow \{-1, 1\}$$

$$\forall n \in \mathbb{N}: \Gamma_{\mathcal{H}}(n) \triangleq \max_{x_1, \dots, x_n \in X} \left| \{(h(x_1), \dots, h(x_n)) \mid h \in \mathcal{H}\} \right| \leq 2^n$$

برای هر n تعداد $\Gamma_{\mathcal{H}}(n)$ حالت شکل \mathcal{H} است

وایدیت \mathcal{H} distribution agnostic

ابتدا $d+1$ تا n توانایی \mathcal{H} را داشته باشند و آنها 2^{d+1} تا n توانایی \mathcal{H} را دارند (یعنی \mathcal{H} شکل n توانایی را شکل $d+2$ توانایی را نمی‌شکله باشد)



VC Dimension (\mathcal{H})

breakpoint D را نقطه ای می‌گویند که $\Gamma_{\mathcal{H}}(D) = 2^D$ و $\Gamma_{\mathcal{H}}(D+1) > 2^D$ باشد. از نظریه یک دفعه پیش از D توانایی \mathcal{H} داشتند و در D توانایی \mathcal{H} داشتند.

VC Dimension

$$d_{VC}(\mathcal{H}) = \max \left\{ D \in \mathbb{N} \mid \Gamma_{\mathcal{H}}(D) = 2^D \right\} = \text{breakpoint} - 1$$

برای \mathcal{H} توانایی \mathcal{H} را $\text{Rad}_{\mathcal{H}}$ می‌گیریم. $\text{Rad}_{\mathcal{H}}$ را Rad می‌گیریم و Rad را $\sup_{P \in \mathcal{P}} \frac{\text{Rad}}{n}$ می‌گیریم. P مجموعه داره.

Sauer's Lemma

if $\mathcal{H} \subseteq \{-1, 1\}^X$ such that $d_{VC}(\mathcal{H}) = D < \infty$

then $\forall n \in \mathbb{N}$,

$$\Gamma_{\mathcal{H}}(n) \leq \sum_{i=0}^D \binom{n}{i} \leq \left(\frac{en}{D}\right)^D$$

proof: at Mohri's book $\binom{n-1}{i-1} + \binom{n-1}{i} = \binom{n}{i}$

ولی خوب به اینرا Rad جنجال نیست و فقط برای VC توانایی \mathcal{H} می‌گیریم.

if $\mathcal{H} = \{\text{sign}(\sin(wx)) \mid w \in \mathbb{R}\} \rightarrow d_{VC}(\mathcal{H}) = \infty$

برای هر n توانایی \mathcal{H} توانایی \mathcal{H} را شکل n توانایی \mathcal{H} دارد.

برای \mathcal{H} بسیار بزرگ شود و $\text{Rad}_{\mathcal{H}}$ بزرگ شود. $\text{Rad}_{\mathcal{H}} = \text{Rad}$ است.

$\text{Rad}_{\mathcal{H}} = \text{Rad}$ است.

برای $\text{Rad}_{\mathcal{H}} = \text{Rad}$ داره.

برای $\text{Rad}_{\mathcal{H}} = \text{Rad}$ داره.

برای $\text{Rad}_{\mathcal{H}} = \text{Rad}$ داره.

$$W = \text{Rad} \left[1 + \sum_{j=1}^n \frac{(1-\delta_j)}{2} 4^j \right]$$

شکل \mathcal{H} را شکل n توانایی \mathcal{H} دارد.

if $\mathcal{H} = \{\text{convex shapes}\} \rightarrow d_{VC}(\mathcal{H}) = \infty$

Massart's Lemma

$$\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_n) \end{bmatrix} = \text{dichotomy}$$

if $A = \{\alpha^{(1)} \dots \alpha^{(|A|)}\}$ $\alpha^{(j)} \in \mathbb{R}^n$, $\|\alpha^{(j)}\|_2^2 \leq r^2$

$$\text{then } \mathbb{E}_{\epsilon_1 \dots \epsilon_n \sim \text{Rad}} \left[\frac{1}{n} \sup_{\alpha \in A} \sum_{i=1}^n \alpha_i \epsilon_i \right] \leq \frac{\sqrt{2 \log |A|}}{n} r$$

میکار و دخنای بخواه اگر داخل
دیگر کره n بعدی سیر بردار
داری باز نمای تا بردار، انتهه باشی
 $\mathbb{E}_{\epsilon_1 \dots \epsilon_n \sim \text{Rad}} \left[\frac{1}{n} \sup_{h \in H} \sum_{i=1}^n h(x_i) \epsilon_i \right] \leq \sqrt{\frac{\log |H|}{2n}}$
tight \hat{R}_n

for binary classification and zero-one loss :

$$R(h) \stackrel{1-\delta}{\leq} \hat{R}_n(h) + \mathbb{E}_{x_1 \dots x_n} \mathbb{E}_{\epsilon_1 \dots \epsilon_n} \left[\frac{1}{n} \sup_{h \in H} \sum_{i=1}^n h(x_i) \epsilon_i \right] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\leq \frac{1}{n} \sqrt{2 \log \text{TC}_H(n)} \sqrt{n}$$

$$A = \left\{ \alpha^{(1)} = \begin{bmatrix} h_1(x_1) \\ \vdots \\ h_1(x_n) \end{bmatrix}, \alpha^{(2)} = \begin{bmatrix} h_2(x_1) \\ \vdots \\ h_2(x_n) \end{bmatrix}, \dots \right\} \xrightarrow{\text{by definition}} |A| = \text{TC}_H(n)$$

$$\rightarrow R(h) \stackrel{\text{Prob. } 1-\delta}{\leq} \hat{R}_n(h) + \mathbb{E}_{x_1 \dots x_n} \left[\sqrt{\frac{2}{n} \log \text{TC}_H(n)} \right] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \leq \hat{R}_n(h) + \sqrt{\frac{2D}{n} \log \left(\frac{ne}{D} \right)} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

we ignore this by assuming

$\text{TC}_H(n)$ is not dependant on p

but interaction between g and p is lost

الآن \oplus VC, Rademacher در ترتیب کردیم

یک بازه دلخواه را تقریباً در اینجا بازه جامد می‌دانیم:

For binary classification and zero-one loss:

$$z_1, \dots, z_n \stackrel{\text{iid}}{\sim} p$$

$$P\left(\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)| > \epsilon\right) \leq 4 \cdot \text{TC}_{\mathcal{H}}(2n) \cdot e^{-\frac{n\epsilon^2}{8}}$$

$$= R(h) \leq \hat{R}_n(h) + \sqrt{\frac{8D \log \frac{2n\epsilon}{\delta}}{n} + \frac{8 \log \frac{4}{\delta}}{n}}$$

prob. $1-\delta$

⊗ این قطعه درایه حرف زدن حالا \hat{R}_n lower bound و R upper bound است. حالا بسیار ساده شدن upper bound از tightness

Lower Bounds

① let \mathcal{H} have $d_{VC}(\mathcal{H}) = D > 1$.

$$\hat{R}_n(h_s) = 0 \rightarrow \inf_{h \in \mathcal{H}} R(h)$$

then for any learning alg. A ,

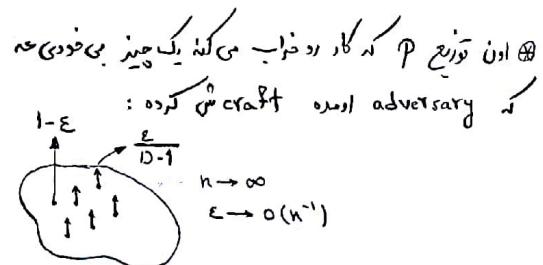
اگر n عدد زیاد و نتایج خاکل یک ترتیب

there exists a distribution p over X

وجود دارد که در آن به خواهی بود

and a target function $f \in \mathcal{H}$ such that:

$$P_{S \sim p^n} \left(R(h_s, f) > \frac{D-1}{32n} \right) \geq \frac{1}{100}$$



⊗ جون به ازای یک توزیع لعنتی می‌توان generalization حدیت پذیر است

② $\hat{R}_n \neq 0 \rightarrow \inf_{h \in \mathcal{H}} R(h) \text{ unrealizable}$

$$P_{S \sim p^n} \left(R(h_{A(S)}) - \inf_{h \in \mathcal{H}} R(h) \geq \sqrt{\frac{D}{320n}} \right) \geq \frac{1}{64}$$

SVM

Binary Linear Classification

linearly separable case ($\hat{R}_n = 0$)
and zero-one loss
the trainset given
testset is given

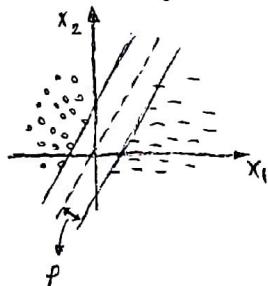
$$\mathcal{H} = \left\{ h(n) = \text{sign}(w^T n + b) \mid \begin{array}{l} w \in \mathbb{R}^d \\ b \in \mathbb{R} \end{array} \right\}$$

$$S = \{(x_i, y_i)\}_{i=1}^n$$

$$w^*, b^* \rightarrow R(w^*, b^*) \leq \sqrt{\frac{2(d+1)}{n} \log \frac{n\epsilon}{d+1}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

من میزان بیک بازیست از این پر

Maximum Margin classification



$$\rho = \min_{i \in [n]} \frac{|w^T n_i + b|}{\|w\|_2}$$

$$\frac{1}{\alpha} (w^T n_i + b - \alpha) > 0$$

$$\begin{aligned} w &\rightarrow \frac{w}{\alpha} \\ b &\rightarrow \frac{b}{\alpha} \end{aligned}$$

canonical form

$$\begin{cases} \text{if } y_i = 1 \rightarrow w^T n_i + b \geq 1 \\ \text{if } y_i = -1 \rightarrow w^T n_i + b \leq -1 \end{cases}$$

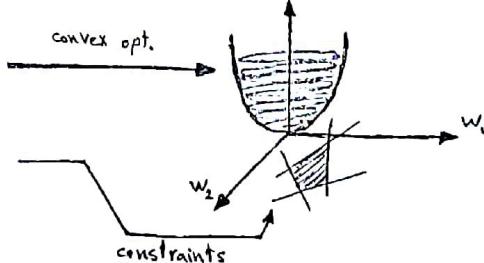
$$\rightarrow y_i (w^T n_i + b) \geq 1 \quad \forall i \in [n]$$

convex

objective (primal form)

$$\max_{w,b} \rho = \max_{w,b} \frac{1}{\|w\|_2} = \min_{w,b} \frac{1}{2} \|w\|_2^2$$

$$\text{st. } y_i (w^T n_i + b) \geq 1 \quad \forall i \in [n]$$



مثلاً QP مثل non-convex مثل

Lagrangian

$$L(w, b, \underbrace{\alpha_1, \dots, \alpha_n}_{\alpha_i \geq 0}) = \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^n \alpha_i (y_i (w^T n_i + b) - 1) \quad \alpha_i \geq 0$$

$$\bullet \nabla_w L = 0 \rightarrow w^* - \sum_{i=1}^n \alpha_i y_i x_i = 0$$

$$\bullet \nabla_b L = 0 \rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

$$\bullet \text{Complementary slackness: } \alpha_i [y_i (w^T n_i + b) - 1] = 0$$

ابن فرق و نیاز دارم
برابر dual بشه. یعنی تراویح هم دید
که برابر بسن

concave

convex

dual form : $\max_{\alpha_1, \dots, \alpha_n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j (y_i y_j x_i^T x_j)$

s.t. $\alpha_i \geq 0 \quad \sum_{i=1}^n \alpha_i = 0$

$w^* = \sum_{i=1}^n \alpha_i y_i n_i$

$b^* = y_i - w^{*T} n_i \quad \forall n_i \in \{\text{SV}\}$

$\vec{w}, \vec{n} \leftarrow$ primal $\begin{cases} \text{if } \vec{w} \neq 0 \\ \text{if } \vec{n} \neq 0 \end{cases} \oplus$

(n) $\vec{w}, \vec{n} \leftarrow$ dual $\begin{cases} \text{if } \vec{w} \neq 0 \\ \text{if } \vec{n} \neq 0 \end{cases} \oplus$

$$\nabla^2 = \frac{1}{2} \left[y_i y_j x_i^T x_j \right]_{i,j=1}^n$$

$A^T A = \text{Gram matrix}$

$A = \begin{bmatrix} y_1 n_1 & \dots & y_1 n_n \\ \vdots & \ddots & \vdots \\ y_n n_1 & \dots & y_n n_n \end{bmatrix}_{d \times n}$

if pd \rightarrow unique SVs

if psd \rightarrow more pairs of SVs

in both cases margin is the same

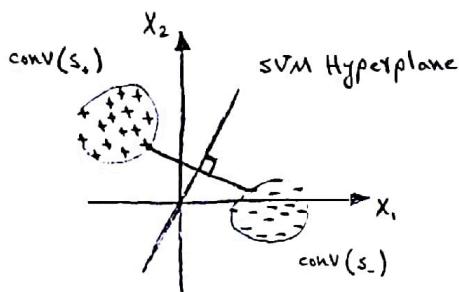
اگر مجموعهای جوابی برای $\vec{\alpha}$ psd نبودند که یک solution ممکن است که مجموعهای جوابی بوده و \vec{w} نیز مقادیر جوابی نداشت. لسان معانی داشت.

Matrix Form of Dual Form

$$\min_{\alpha} \frac{1}{2} \alpha^T \underbrace{\begin{bmatrix} y_1 y_1^T n_1 & \dots & y_1 y_n^T n_n \\ \vdots & & \vdots \\ y_n y_1^T n_1 & \dots & y_n y_n^T n_n \end{bmatrix}}_{\text{quadratic coefficients}} \alpha + (-1^T) \alpha$$

s.t. $y^T \alpha = 0, \quad 0 \leq \alpha \leq \infty$

SVM خواص منزلي



فرزن لين با داده های دسته داشته باشند pair of data

که d_{vc} بخوبی generalization \rightarrow svr اگر \oplus

فرزن لين با داده های iid داده های w^*, b^* را می سین
بعد از از داده های iid و داده های آزمون می سین. احتمال انتباوه \oplus

$$\text{احتمال خطأ} \leq \frac{N_{sv}}{n+1} \longrightarrow E_s \left[R(h_{SVM}^*(s)) \right] \leq \frac{E[N_{sv}]}{n+1}$$

trained on $n+1$ iid samples

↓

trained on n iid samples

\oplus باز جزوی نیست جون E_s در نمی کردن

$$\mathcal{H} = \left\{ x \rightarrow \text{sign}(w^T x) \mid \min_{x \in \mathcal{X}} |w^T x| = 1, \|w\|_2 \leq \frac{1}{\Delta} \right\}$$

فرزن لين با مارجین Δ

\oplus مرجین طوری می شود در تابع w در تابع R در نظر گرفت
دایلکتریکی می نماییم

\oplus مرجین نزدیک لین هم داده ها را کرو به مساح R مدار داریم

$$d_{vc}(\mathcal{H}) \leq \Delta + 1$$

جون d_{vc}

$$D = \sum_{i=1}^D 1 \leq \sum_{i=1}^D y_i (w^T x_i) = w^T \sum_{i=1}^D y_i x_i \quad \forall i (w^T x_i + b) y_i \geq 1$$

\oplus مرجین نزدیک لین هم داده ها را کرو به مساح R مدار داریم

$$\leq \|w\|_2 \left\| \sum_{i=1}^D y_i x_i \right\|_2 \leq \frac{1}{\Delta} E_y \left[\left\| \sum_{i=1}^D y_i x_i \right\|_2^2 \right]$$

$$\leq \frac{1}{\Delta} \int E_y \left[\left\| \sum_{i=1}^D y_i x_i \right\|_2^2 \right] = \frac{1}{\Delta} \int \sum_{i,j=1}^D E[y_i y_j] x_i^T x_j = \frac{1}{\Delta} \int \sum_{i,j=1}^D \underbrace{E[y_i]}_0 \underbrace{E[y_j]}_0 x_i^T x_j$$

$$\cdots \int \sum_{i=1}^D E[y_i^2] x_i^T x_i = \frac{1}{\Delta} \int \sum_{i=1}^D \underbrace{\|x_i\|_2^2}_{\leq R^2} \leq \frac{1}{\Delta} \int D R^2 \implies D \leq \left(\frac{R}{\Delta} \right)^2$$

$$\rightarrow d_{VC}(H) \leq \min \left\{ d+1, \left(\frac{R}{\Delta}\right)^2 \right\}$$

از سر d تا n خلاص
 $\frac{R}{\Delta} \ll \sqrt{d}$

آن دادنیم می توانیم باشد بهترین
 generalization را
 نهایی مفہومی و بایاسی و خراب (محدود) کنیم

$\forall h \in \mathcal{H}$

$$R(h) \leq \hat{R}_n(h) + O\left(\sqrt{\frac{\left(\frac{R}{\Delta}\right)^2 + \log \frac{1}{\delta}}{n}}\right)$$

$$\forall i : x_i \in \{sv\} \quad b = y_i - w^T x_i$$

$$\rightarrow \forall i : \underbrace{\sum_{i=1}^n \alpha_i y_i b}_{\| \alpha \|_1} = \underbrace{\sum_{i=1}^n \alpha_i \frac{y_i^2}{1}}_{\| \alpha \|_1} - \underbrace{\sum_{i=1}^n \alpha_i y_i w^T x_i}_{\| w \|_2^2} \rightarrow \| w \|_2 = \sqrt{\| \alpha \|_1}$$

$x_i \notin \{sv\}$
 ازای i هایی که
 α_i صفر و میانگین نداشت

Margin Theory

$$\mathcal{H} \subseteq \mathbb{R}^X \neq \text{classifier}$$

$$(y_i, x_i) \stackrel{\text{iid}}{\sim} P^n$$

$$\forall h \in \mathcal{H}$$

$$R(h) \leq R_p(h) \leq \hat{R}_p(h) + 2 \text{Rad}(\Phi_p \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$\mathbb{E}_{(x_i, y_i)} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \Phi_p(y_i h(x_i)) \right]$

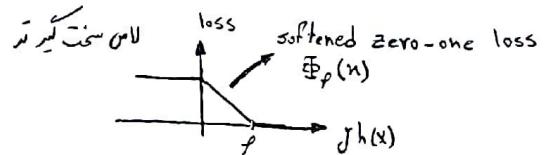
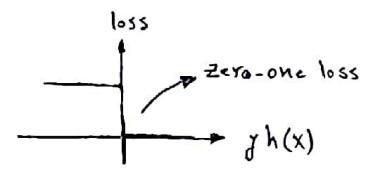
($\rho > 1$) $\frac{1}{\rho} \text{Rad}(\mathcal{H})$ \rightarrow Rad complexity ω
 باين ω \rightarrow Rad complexity ω
 دلی خواست باش که ادلهای ρ فیکس بشه بعدن سهیل بگیر
 کرد optimize ω , ρ \rightarrow ρ

only relevant ω Rad complexity ω که درون \mathcal{H} با استثنای $\mathcal{H} \cap \{(0,1)\}$ باشد \oplus
 $\text{Rad}(\alpha \mathcal{H}) = \alpha \text{Rad}(\mathcal{H})$ جون

$$(x_1, y_1), \dots, (x_n, y_n)$$

$$\forall h \in \mathcal{H} \text{ and } \forall \rho \in (0,1)$$

$$R(h) \leq \hat{R}_{n,\rho}(h) + \frac{4}{\rho} \text{Rad}_n(\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \sqrt{\frac{\log \log \frac{2}{\delta}}{n}}$$



$$R(h) = \mathbb{E} [\text{zero one loss}]$$

$$R_p(h) = \mathbb{E} [\Phi_p(j_h(x))]$$

$$R(h) \leq R_p(h)$$

$$\hat{R}(h) \leq \hat{R}_p(h)$$

Φ is γ -Lipschitz

$$\text{if } \forall n, j \quad |\Phi(n) - \Phi(j)| \leq \gamma \|n-j\|$$

$$\approx \frac{1}{\rho} \text{Lipschitz} \Rightarrow \Phi_p \text{ is Lipschitz}$$

Talagrand's Lemma

$g \in \mathbb{R}^X$ $\Phi: \mathbb{R} \rightarrow \mathbb{R}$ is γ -Lipschitz

$$\text{Rad}_n(\Phi \circ g) \leq \gamma \text{Rad}_n(g)$$

نهی سهل نام پنداشتنی کرده

Soft SVM

$$\min_{w, b, \xi} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i^p \quad p \geq 1 \quad \xi_i \geq 0$$

$p=1$ (hinge loss)

$C \rightarrow \infty$ hard SVM

$$\text{s.t. } y_i (w^T x_i + b) \geq 1 - \xi_i$$

$$\forall i \in [n]$$

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$\rightarrow w^* = \sum_{i=1}^n \alpha_i y_i x_i, \quad \sum_{i=1}^n \alpha_i y_i = 0, \quad C - \alpha_i - \beta_i = 0$$

$$\alpha_i \geq 0, \quad \beta_i \geq 0, \quad \text{complementary slackness}$$



dual form

$$\max_{\alpha_1, \dots, \alpha_n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

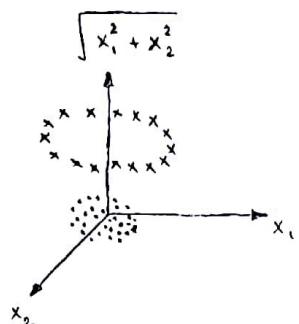
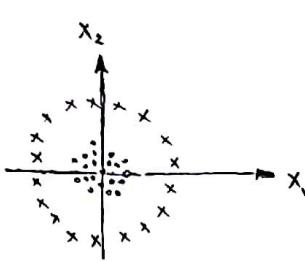
$$\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C$$

$$\forall h \in \mathcal{H} = \left\{ n \rightarrow w^T n \mid \|w\| < \frac{1}{\Delta} \right\}$$

$$R(h) \leq \frac{1}{n} \sum_{i=1}^n \xi_i + 2 \sqrt{\frac{R^2}{n \Delta^2}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\xi_i \triangleq \max(1 - y_i (w^T x_i), 0)$$

kernel methods



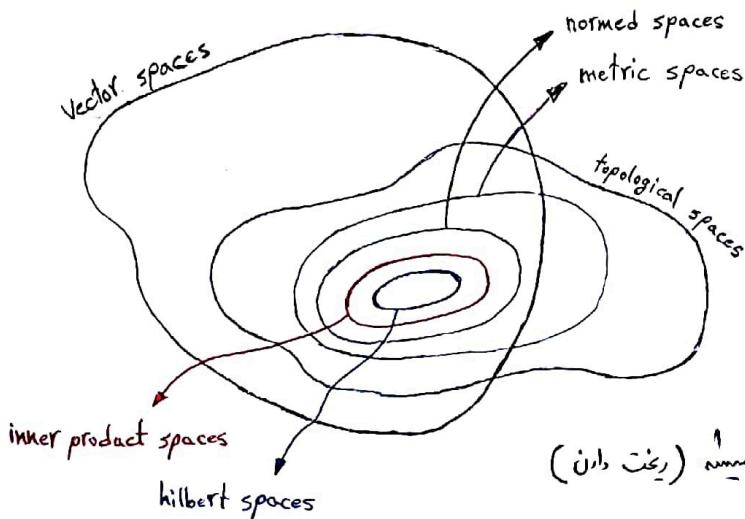
$$\Phi : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ x_2 \\ \sqrt{x_1^2 + x_2^2} \end{bmatrix}$$

ⓐ این تعبیه فضای از SUM و هرچیزی که ضرب داشل دردی ها رو داشته باشیم میتوان از آن استفاده کرد

kernel Function

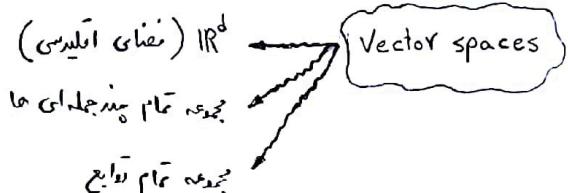
$$k: X \times X \rightarrow \mathbb{R}$$

feature space چون دو عکس نموداری داشته باشیم که اسکالر جمعون بده $\Phi(x_i)^T \Phi(x_j) \leftrightarrow k(x_i, x_j)$ پس اگر کنیم که :



inner product spaces زیرمجموعه ای از Hilbert spaces
نمایند و complete کر

اگر با انتخاب دنباله درست کردی ب اینها خود را هم کنیم



مقتلهایی که میتوانند مطلع میشون (یعنی دارند) topological spaces

مقتلهای اولیه ای که برای مطالعه اعفانی سرتاسر تعریف شده Metric spaces

$$d: X \times X \rightarrow \mathbb{R}_{\geq 0} \quad (i) d(x, y) = d(y, x) \quad (ii) d(x, y) + d(y, z) \geq d(x, z) \quad (iii) d(x, y) = 0 \iff x = y$$

Normed spaces

$$p: X \rightarrow \mathbb{R}_{\geq 0} \quad (i) p(x+y) \leq p(x) + p(y) \quad (ii) p(sx) = |s|p(x) \quad (iii) p(x)=0 \iff x=0$$

Inner product spaces

$$\langle \cdot, \cdot \rangle: X \times X \rightarrow \mathbb{R} \quad (i) \langle x, y \rangle = \langle y, x \rangle \quad (ii) \langle \alpha x + \beta z, y \rangle = \alpha \langle x, y \rangle + \beta \langle z, y \rangle$$

$$(iii) \langle x, x \rangle \geq 0 \text{ and } \langle x, x \rangle = 0 \iff x = 0$$

For Hilbert space $\mathbb{H} = \left\{ f: \mathbb{R} \rightarrow \mathbb{R} \mid f \in L^2(\mathbb{R}) \right\}$

: $\int_{-\infty}^{\infty}$

inner product $\langle f, g \rangle \triangleq \int_{-\infty}^{\infty} f(u) g(u) du$

norm $\|f\|_{\mathbb{H}} \triangleq \sqrt{\int_{-\infty}^{\infty} f(u)^2 du}$

definition of kernel

$$k: X \times X \rightarrow \mathbb{R}$$

↓
input space $\subseteq \mathbb{R}^d$

usually compact \rightsquigarrow bounded and closed

positive definite symmetric kernel

definition of PDS

- having Mercer's condition

$$\forall c: X \rightarrow \mathbb{R}, c \in L^2(X)$$

must be compact

$$\forall n \in \mathbb{N} \quad \forall x, x' \in X \quad \int \int_{X \times X} c(u)c(u') k(u, u') du du' \geq 0, \quad k(u, u') = k(u', u)$$

$\Rightarrow k$ is symmetric

- having a stronger condition

$$\forall n \in \mathbb{N} \quad \forall x_1, \dots, x_n \in X$$

$$\left[k(x_i, x_j) \right]_{i,j=1}^n \geq 0$$

PSD matrix

PDS \Leftrightarrow دلیل تئن زدن دایبات \otimes
بودن محدوده

RkHS Theorem

reproducible kernel Hilbert space

If a kernel k is PDS on X then there exists $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and $\Phi: X \rightarrow \mathcal{H}$
such that $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$

Reproducing property

$$\forall h \in \mathcal{H} \quad \forall x \in X \quad h(x) = \langle h, k(x, \cdot) \rangle$$

proof For $x \in X, \forall x' \in X$

$$\underbrace{\Phi(x)(x')}_{\in \mathcal{H}} \triangleq k(x, x')$$

vector space \hookrightarrow با بر این ترتیب

و مکمل . \hookrightarrow inner product space

Hilbert space \leftarrow باید ignore

h از \mathcal{H} است

$$\text{pre Hilbert } \mathcal{H}_0 = \left\{ \sum_{i \in I} a_i \Phi(x_i) \mid a_i \in \mathbb{R}, x_i \in X, \text{card}(I) < \infty \right\}$$

$$f, g \in \mathcal{H}_0 \rightarrow \langle f, g \rangle_{\mathcal{H}_0} \triangleq \sum_{\substack{i \in I \\ j \in J}} a_i b_j k(x_i, x'_j)$$

$$I, \{a_i\}, \{x_i\} \rightarrow f = \sum_{i \in I} a_i \Phi(x_i)$$

$$J, \{b_j\}, \{x'_j\} \rightarrow g = \sum_{j \in J} b_j \Phi(x'_j)$$

$$h(n) = \langle h, k(n, \cdot) \rangle$$

نیم ابتداء هست

پرکاردن سوابع هاست

Hilbert space \longleftrightarrow pre Hilbert space

\equiv

$\mathbb{R} \longleftrightarrow \mathbb{N}$

$$\left\langle \sum_{i=1}^n a_i \Phi(n_i), \sum_{j=1}^l 1 \cdot \Phi(n_j) \right\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^n a_i \underbrace{\langle \Phi(n_i), \Phi(n) \rangle}_{k(n, n_i)}$$

لین تعریف که برای فضای داخلی کردیم

: Reproducing property مسود

از طرف دیگر طبق تعریف $(h(n))$ داریم

Hilbert فضاهای پایه های محدود
حصن (البتہ دوستی پایه نیست. بلکه

$$h(n) = \sum_{i=1}^n a_i \Phi(n_i)(n) = \sum_{i=1}^n a_i k(n, n_i)$$

$$\Phi(x) \triangleq k(x, \cdot)$$

$$\rightarrow h(n) = \langle h, k(n, \cdot) \rangle \quad \checkmark$$

polynomial kernel is PSD

$$k(n, n') \triangleq (n^T n' + c)^r \quad c \geq 0, r \in \mathbb{N}$$

$n, n' \in \mathbb{R}^d$

For $d=2, r=2$ $\Phi : \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \rightarrow \begin{bmatrix} n_1^2 \\ n_2^2 \\ \sqrt{2}n_1 \cdot n_2 \\ \sqrt{2c} n_1 \\ \sqrt{2c} n_2 \\ c \end{bmatrix}$

$$\Phi : \mathbb{R}^d \rightarrow \frac{\binom{d+r}{r}}{\mathcal{H}}$$

Gaussian aka RBF kernel is PSD

$$k(n, n') \triangleq e^{-\frac{\|n - n'\|_2^2}{2\sigma^2}} \quad \sigma > 0$$

$$\Phi : \mathbb{R}^d \rightarrow \frac{\mathbb{R}^\infty}{\mathcal{H}}$$

$$\|\Phi(n)\|_{\mathcal{H}} = \sqrt{k(n, n)} = 1$$

If $k(x, x')$ is PSD, its normalized version $\frac{k(x, x')}{\sqrt{k(x, x)k(x', x')}}$ is also PSD

حالا برای استفاده از کرنل در SVM هرچاکی داشتیم بی جای $\Phi(x_i)$ داریم

$$\rightarrow w^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i) \quad \text{برای explicit form of RBF برای}$$

$$\rightarrow \langle w^*, \Phi(x) \rangle + b^*$$

$$= \sum_{i=1}^n \alpha_i^* y_i \underbrace{\langle \Phi(x_i), \Phi(x) \rangle}_{k(x_i, x)} + b^*$$

$$\xrightarrow{\text{RBF}} \text{Sign} \left(\sum_{i=1}^n \alpha_i y_i e^{-\frac{\|x - x_i\|_2^2}{2\sigma^2}} + b \right)$$

$$\xrightarrow{\substack{\text{all } \alpha_i = 1 \\ \text{all } y_i \\ b = 0}} \text{sign} \left(\sum_{i=1}^n y_i e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}} \right) \quad (x = x_j) \text{ است بلطفاً اگر داده ای آمریخ است}$$

$$= \text{sign} \left(y_j \times 1 + \underbrace{\sum_{i \neq j} y_i e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}}_{\rightarrow 0} \right)$$

اگر x از تردیکردن ناصله باشد هم خواهد بود
بلطفاً $e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$ صفر می شود

$$= \text{sign}(y_j) = y_j$$

از Kernelها استفاده می‌کنیم که چون میدانی در دل بودجه خلی سه بخوبی کرد.

Representer Theorem

assume $L: \mathbb{R}^n \rightarrow \mathbb{R}$, $G: \mathbb{R} \rightarrow \mathbb{R}$ and G is strictly increasing

$x_1, \dots, x_n \in \mathcal{X}$, k is pDS on \mathcal{X} $\longrightarrow \mathcal{H} \subset \mathbb{R}^n$ $\| \cdot \|_{\mathcal{H}}$

یک مقسوم بر ترکیب kernel PCA, kernel SVM
مدیران های ازین هستند

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} G(\|h\|_{\mathcal{H}}) + L(h(x_1), \dots, h(x_n))$$

بنابراین فضای بعدی های
برای L می‌توانند محسن شوند.

$$\longrightarrow \exists \alpha_1, \dots, \alpha_n \in \mathbb{R} \quad h^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$$

(دیگر دوام نیست)

جنس $h \in \mathcal{H}$ هنوز ابر مسنه جدا نشده در پایه بالاتر (یعنی w است)

تبیین ترکیب بودن که همراه با آنیم h رو به صورت $\sum_{i=1}^n \alpha_i k(x_i, \cdot)$ بود. یک عدد بزرگ دل بخواه بود!

proof is said to be easy!

kernel Bounds :: Rademacher

$$\Omega = \left\{ \langle w, \Phi(\cdot) \rangle \mid w \in \mathcal{H}, \|w\|_{\mathcal{H}} < \Delta \right\} = \{f: X \rightarrow \mathbb{R}\}$$

$$S = \left\{ (\underbrace{x_i}_{j_i}) \right\}_{i=1}^n \stackrel{iid}{\sim} P^n$$

$$\mathbb{E}_{x \sim p} \left[\langle w, \Phi(x) \rangle \right] \leq \hat{\mathbb{E}}_n \left[\langle w, \Phi(x) \rangle \right] + 2 \mathbb{E}_{\substack{x_1, \dots, x_n \\ \epsilon_1, \dots, \epsilon_n}} \left[\frac{1}{n} \sup_{w \in \Omega} \sum_{i=1}^n \epsilon_i \langle w, \Phi(x_i) \rangle \right] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\longrightarrow \mathbb{E}_{x, y \sim p} \left[\psi_p \left(\frac{1}{p} \langle w, \Phi(x) \rangle \right) \right] \leq \hat{\mathbb{E}}_n \left[\dots \right] + \frac{2}{p} \mathbb{E}_{\substack{x_1, \dots, x_n \\ \epsilon_1, \dots, \epsilon_n}} \left[\dots \right] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$\psi_p(n)$ is $\frac{1}{p}$ -Lipschitz

$$\frac{2}{\rho} \mathbb{E}_{\substack{x_1, \dots, x_n \\ \epsilon_1, \dots, \epsilon_n}} \left[\frac{1}{n} \sup_{w \in \Omega} \sum_{i=1}^n \epsilon_i \langle w, \Phi(x_i) \rangle \right]$$

$$= \frac{2}{\rho} \mathbb{E} \left[\underbrace{\sup_{w \in \Omega} \left\langle w, \frac{1}{n} \sum_{i=1}^n \epsilon_i \Phi(x_i) \right\rangle}_{\leq \Delta} \right]$$

$$\leq \|w\|_{\mathcal{H}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \Phi(x_i) \right\|_{\mathcal{H}}$$

$$\leq \frac{2\Delta}{n\rho} \mathbb{E}_{\substack{x_1, \dots, x_n \\ \epsilon_1, \dots, \epsilon_n}} \left[\left\| \sum_{i=1}^n \epsilon_i \Phi(x_i) \right\|_{\mathcal{H}} \right] = \frac{2\Delta}{n\rho} \sqrt{\mathbb{E}^2[\|\cdot\|]} \stackrel{\text{Jensen}}{\leq} \frac{2\Delta}{n\rho} \sqrt{\mathbb{E}[\|\cdot\|_{\mathcal{H}}^2]}$$

$$\leq \frac{2\Delta}{n\rho} \sqrt{\mathbb{E} \left[\sum_{i,j=1}^n \epsilon_i \epsilon_j \langle \Phi(x_i), \Phi(x_j) \rangle \right]} = \frac{2\Delta}{n\rho} \sqrt{\sum_{i=1}^n \mathbb{E}[k(x_i, x_i)]} = \frac{2\Delta}{\sqrt{n}\rho} \sqrt{\mathbb{E} k(x, x)}$$

For RBF $\frac{2\Delta}{\sqrt{n}\rho}$

$$\langle w, \Phi(x) \rangle = \sum_{i=1}^n \gamma_i e^{-\frac{\|x_i - x\|^2}{2\sigma^2}}$$

$$w = \sum_{i=1}^n \gamma_i k(x_i, \cdot) \longrightarrow \|w\|_{\mathcal{H}}^2 = [\gamma_1 \dots \gamma_n] \begin{bmatrix} 1 & \cancel{\gamma_1} & \dots & \cancel{\gamma_n} \\ \cancel{\gamma_1} & \ddots & \ddots & 1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix}$$

$$\longrightarrow \|w\|_{\mathcal{H}}^2 = \sum_{i=1}^n \gamma_i^2 \simeq n \longrightarrow \|w\|_{\mathcal{H}} \simeq \sqrt{n}$$

Boosting ⊂ Ensemble Learning

Ada Boost \rightarrow For binary classification

initial \rightarrow to weak learner \rightarrow iterative \rightarrow Boosting \rightarrow \oplus

step 0

$$(x_i, y_i)_{i=1}^n \quad D_0(i) = \frac{1}{n} \quad h \in \mathcal{H} \quad \text{a set of weak classifiers}$$

$$i = 1 \dots n$$

$$\hat{R}_0(h) = \hat{\mathbb{E}}_{D_0} [\mathbb{1}(y \neq h(x))] \rightarrow (h_0, \varepsilon_0) = (\underset{h \in \mathcal{H}}{\operatorname{argmin}}, \underset{h \in \mathcal{H}}{\min}) \hat{R}_0(h)$$

step 1
↓
⋮

$$D_t(i) = \frac{A_0^{-y_i h_0(x_i)}}{Z_0} D_0(i) \quad A_0 \triangleq \sqrt{\frac{1-\varepsilon_0}{\varepsilon_0}} = e^{\alpha_0}$$

$$(h_t, \varepsilon_t) = (\underset{h \in \mathcal{H}}{\operatorname{argmin}}, \underset{h \in \mathcal{H}}{\min}) \hat{R}_t(h)$$

$$R_t = \mathbb{E}_{D_t} [\mathbb{1}(y \neq h(x))] = \frac{1}{n} \sum_{j=1}^n D_t(j) \mathbb{1}(y_j \neq h(x_j))$$

$$T \in \mathbb{N}$$

$$g^*(.) = \sum_{t=0}^{T-1} \overbrace{\log \left(\frac{1-\varepsilon_t}{z_t} \right)}^{\alpha_t} h_t(.)$$

$$D_{t+1}(i) = D_t(i) \frac{e^{-\alpha_t y_i h_t(x_i)}}{z_t}$$

$$z_t = \sum_{i=1}^n D_t(i) e^{-\alpha_t y_i h_t(x_i)}$$

$$= e^{\alpha_t} \underbrace{\sum_{i \in \text{incorrect}} D_t(i)}_{\varepsilon_t} + e^{-\alpha_t} \underbrace{\sum_{i \in \text{correct}} D_t(i)}_{1-\varepsilon_t} = 2 \sqrt{\varepsilon_t (1-\varepsilon_t)}$$

$$= \sqrt{1 - 4(\frac{1}{2} - \varepsilon_t)^2} \leq e^{-2(\frac{1}{2} - \varepsilon_t)^2} \quad (*)$$

$\gamma \equiv$ margin

$$(A, \mathcal{H}) \text{ is called } \gamma\text{-weak if} \\ R \neq \frac{1}{2} - \gamma$$

$$\mathbb{E}_{S \sim P_n} [R(h(s))] \leq \frac{1}{2} - \gamma$$

$$\hat{R}_n(g^*) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}_{(y_i \neq \text{sign}(g^*(x_i)))}}_{\mathbb{1}_{y_i g^*(x_i) < 0}} \leq \frac{1}{n} \sum_{i=1}^n e^{-y_i g^*(x_i)} \\ = \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=0}^{T-1} e^{-\alpha_t y_i h_t(x_i)} \frac{z_0 \dots z_{T-1}}{z_0 \dots z_{T-1}} \right) = z_0 \dots z_{T-1}$$

$$\rightarrow \hat{R}_n(g^*) \stackrel{*}{\leq} e^{-2 \sum_{t=0}^{T-1} \underbrace{\left(\frac{1}{2} - \varepsilon_t\right)^2}_{\geq \gamma}} \leq e^{-2\gamma^2 T} \quad \text{weak learner is } \gamma \text{ weak}$$

AdaBoost pseudo code → next page

AdaBoost Generalization

$$g^* \in \mathcal{G} = \left\{ \text{sign} \left(\sum_{t=0}^{T-1} \alpha_t h_t(\cdot) \right) \mid \begin{array}{l} h_t \in \mathcal{H}, t=1 \dots T \\ \alpha_t \in \mathbb{R} \end{array} \right\} \subseteq \{-1, 1\}^X \quad \text{Final classifier}$$

$$d_{VC}(g) \leq 2(d+1)(T+1) \log((T+1)e) \longrightarrow O(dT \log T)$$

\downarrow
 $d_{VC}(\mathcal{H})$

$$R(g) \leq \underbrace{\frac{e^{-2\gamma^2 T}}{\geq \hat{R}_n(g)}}_{\text{test}} + \tilde{O}\left(\sqrt{\frac{dT}{n}}\right)$$

AdaBoost تكتل اعمى و مروي \oplus
وهو ERM بحسب نسبت عصري generalization

AdaBoost pseudo code

$$D_1(i) = \frac{1}{n} \quad \forall i \in [n] \quad , \quad k=1$$

(start) $h_k, \varepsilon_k = (\underset{h \in \mathcal{H}}{\operatorname{arg\,min}}, \underset{h \in \mathcal{H}}{\min}) \sum_{i=1}^n D_k(i) \mathbb{1}(j_i \neq h(x_i))$

$$\alpha_k = \frac{1}{2} \lg \frac{1 - \varepsilon_k}{\varepsilon_k}$$

$$D_{k+1}(i) = D_k(i) \frac{e^{-\alpha_k j_i h_k(x_i)}}{z_k} \quad z_k = 2 \sqrt{\varepsilon_k (1 - \varepsilon_k)}$$

$$k = k+1$$

$$\rightsquigarrow \begin{cases} h_1 \dots h_k \\ \varepsilon_1 \dots \varepsilon_k \\ \alpha_1 \dots \alpha_k \end{cases} \rightarrow g^*(\cdot) = \operatorname{sign} \left(\sum_{k=1}^K \alpha_k h_k(\cdot) \right) : X \rightarrow \{\pm 1\}$$

VC Dimension of AdaBoost

assume x_1, x_2, \dots, x_n

$$\begin{bmatrix} h_1(x_1) & h_1(x_2) & \dots & h_1(x_n) \\ \vdots \\ h_k(x_1) & h_k(x_2) & \dots & h_k(x_n) \end{bmatrix} \rightsquigarrow \text{TC}_{\mathcal{H}}(n)$$

$$\left. \begin{array}{l} \text{عدد حالات برای } \\ \text{کل مجموعه } \\ \text{کل } \mathbb{R}^d \end{array} \right\} \leq \text{TC}_{\mathcal{H}}(n)^k \leq \left(\frac{ne}{d} \right)^{kd}$$

$$\left. \begin{array}{l} \text{عدد حالات درست شده} \\ \text{کل } \mathbb{R}^d \end{array} \right\} \leq \text{TC}_{\mathcal{H}_{\text{HP}}}(n) \leq \left(\frac{ne}{k+1} \right)^{k+1}$$

$$\left. \begin{array}{l} \left(\frac{ne}{d} \right)^{kd} \left(\frac{ne}{k+1} \right)^{k+1} \geq \text{TC}_g(n) \end{array} \right\}$$

$$\rightarrow d_{VC}(g) \leq \max \left\{ n \mid 2^n \leq \underbrace{\left(\frac{ne}{d} \right)^{kd} \left(\frac{ne}{k+1} \right)^{k+1}}_{\leq n^{kd+k+1}} \right\}$$

$(d, k+1 \geq e \text{ وضیع!})$

$$n \leq \underbrace{(kd+k+1)}_a \lg_2 n \xrightarrow{\text{lemma}} n \leq 2a \lg_2 a$$

$$\rightarrow d_{VC}(g) \leq 2(kd + k+1) \lg_2(kd+k+1) = O(kd \lg(kd))$$

$$\rightarrow R(g^*) \leq e^{-2\gamma^2 k} + \tilde{O}\left(\sqrt{\frac{kd}{n}}\right)$$

↙
test error of zero-one loss

Regression

$$x_1, \dots, x_n \stackrel{iid}{\sim} P_x \rightsquigarrow \text{unknown}$$

$$f: X \rightarrow \mathbb{R} \rightsquigarrow \text{true function unknown}$$

$$\text{Dataset} = \{(x_i, f(x_i))\}_{i=1}^n \rightsquigarrow \text{known}$$

$$\mathcal{H} \subseteq \mathbb{R}^X \rightsquigarrow \text{known}$$

ERM:

$$\hat{h}^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n L\left(\hat{y}_i, h(x_i)\right) \quad L(y, y') = |y - y'|^p \quad p \in [1, \infty)$$

$$R = \mathbb{E}_{x \sim P} [L(\hat{h}^*(x), f(x))]$$

↓
جودون محدودون بـ bounded with zero-one loss
عن قيد عدد المدمنون
 $\forall h \in \mathcal{H} \quad \forall x \in X \rightarrow L(h(x), f(x)) \leq M$

$$\forall h \in \mathcal{H}, \quad x_1, \dots, x_n \stackrel{iid}{\sim} P, \quad p=1$$

$$|R(h) - \hat{R}(h)| = \left| \frac{1}{n} \sum (h(x_i) - f(x_i)) - \mathbb{E}[|h(x_i) - f(x_i)|] \right|$$

$|\mathcal{H}| < \infty \rightarrow \text{union bound}$
McDiarmid

$$\mathbb{P}(\exists h \in \mathcal{H}, R(h) \geq \hat{R}(h) + \varepsilon) \leq |\mathcal{H}| e^{-\frac{2n\varepsilon^2}{m^2}}$$

uniform bound
for finite \mathcal{H}

≡

$$R(h) \stackrel{1-\delta}{\leq} \hat{R}_n(h) + M \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

Learning Guarantee for $|\mathcal{H}| = \infty$

$$R(h) \leq \hat{R}_n(h) + 2\text{Rad}_n(g) + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \leq \hat{R}_n(h) + 2\hat{\text{Rad}}_n(g) + 3M \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\text{Rad}_n(g) = \mathbb{E}_{\substack{x \\ \epsilon}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i L(h(x_i), f(x_i)) \right]$$

assume $L(y, y') = |y - y'|^p = \Phi \circ (h(n) - f(n))$

$\Phi(\cdot) = |\cdot|^p$ is pM^{p-1} lipschitz

حالا $\text{Rad}_n(\mathcal{H}) \leq \text{Rad}_n(g)$ باشی

$$\rightarrow R(h) \leq \hat{R}_n(h) + 2 \mathbb{E}_{\substack{x \\ \epsilon}} \left[\frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^n \epsilon_i \underbrace{|h(x_i) - f(x_i)|^p}_{\Phi \circ (h(n) - f(n))} \right] + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\begin{aligned} \rightarrow R(h) &\leq \hat{R}_n(h) + M^p \sqrt{\frac{\log \frac{1}{\delta}}{2n}} + 2pM^{p-1} \mathbb{E}_{\substack{x \\ \epsilon}} \left[\frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^n \epsilon_i (h(x_i) - f(x_i))^p \right] \\ &= \underbrace{\mathbb{E}_{\substack{x \\ \epsilon}} \left[\frac{1}{n} \sup \sum \epsilon_i h(x_i) \right]}_{\Phi \circ (h(n) - f(n))} - \underbrace{\mathbb{E}_{\substack{x \\ \epsilon}} \left[\frac{1}{n} \sup \sum \epsilon_i f(x_i) \right]}_{0} = \text{Rad}_n(\mathcal{H}) \end{aligned}$$

$g \in \mathbb{R}^X \rightsquigarrow \text{pseudo Dimension of } g \text{ or } \text{pdim}(g)$

Shattering for Regression

$x_1, \dots, x_n \in X$ is said to be shattered by g if $\exists t_1, \dots, t_n \in \mathbb{R}$ such that

$$\text{card} \left\{ \begin{bmatrix} \text{sign}(g(x_1) - t_1) \\ \vdots \\ \text{sign}(g(x_n) - t_n) \end{bmatrix} \mid g \in g \right\} = 2^n$$

برای هر couple داده t_i با x_i و \otimes بسیاری حالت ممکن $\text{card}\{\dots\} = 2^n$

$\text{pdim}(g)$ is the largest $n \in \mathbb{N}$ such that $\exists x_1, \dots, x_n \in X$

can be shattered by g

- $\text{pdim}(g) = d_{VC} \left(\left\{ n \rightarrow \mathbb{1}_{g(x_i) - t_i \geq 0} \mid g \in g, t \in \mathbb{R} \right\} \right)$

Theorem 10.4 Mohri

$$g = \left\{ x \mapsto w^T x + b \mid \begin{array}{l} w \in \mathbb{R}^d \\ b \in \mathbb{R} \end{array} \right\} \implies \text{pdim}(g) = d+1$$

Theorem 10.5 Mohri

\mathcal{G} = vector space of real valued functions $\implies \text{pdim}(g) = \dim(g)$

$$\text{e.g. } \left\{ g(x) = \sum_{i,j} A_{ij} x_j x_i + \sum_i B_i x_i + c \mid A_{ij}, B_i, c \in \mathbb{R} \right\}$$

$$\rightarrow \text{pdim}(g) = \underbrace{\binom{d}{2}}_{2 \text{ d.o.f.}} + d + \underbrace{1}_{1 \text{ d.o.f.}}$$

Theorem

if previous and $\text{pdim}(g) = d$

$\implies \forall h \in \mathcal{H}$ (uniform):

$$\mathbb{E}_x [L(h(x), f(x))] \leq \frac{1}{n} \sum_{i=1}^n L(h(x_i), f(x_i)) + M \sqrt{\frac{2d \log \frac{n}{\delta}}{n}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

(uniform) \hat{h} $\in \mathcal{H}$ \approx $\hat{h} \in \mathcal{H}$ min

proof

$$\text{Lemma: } \mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t) dt$$

$$\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)| = \sup_{h \in \mathcal{H}} \left| \int_0^M P_x(L(h(n), f(n)) \geq t) dt \right| \quad \text{statistical}$$

$$- \int_0^M \hat{P}_s(L(h(x), f(x)) \geq t) dt \quad \text{empirical}$$

$$\leq \sup_{h \in \mathcal{H}} M \sup_{t \in [0, M]} \left| P_x(L \geq t) - \hat{P}_s(L \geq t) \right|$$

$$\left| \int_0^M g(t) dt \right| \leq M \sup_{t \in [0, M]} |g(t)|$$

$$\leq M \sup_{h \in \mathcal{H}} \left| \frac{P_x(L \geq t)}{\mathbb{E} c_{h,t}(n)} - \frac{\hat{P}_s(L \geq t)}{\hat{\mathbb{E}}_s c_{h,t}(n)} \right|$$

$$\leq \sqrt{\frac{2d \log \frac{n}{\delta}}{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\mathcal{C} = \left\{ c : n \rightarrow \mathbb{1}_{L(h(n), f(n)) \geq t} \mid t \in \mathbb{R} \right\}$$

$$\rightarrow d_{VC}(\mathcal{C}) = d$$

Rad
+
VC
+
Massart

Using kernels in Regression

$$\Phi: \mathcal{X} \rightarrow \mathcal{H}$$

أكمل تعلم في المنهجية
Learning in the kernel

$$\{n \rightarrow \langle w, \Phi(x) \rangle_{\mathcal{H}} \mid w \in \mathcal{H}, \|w\|_{\mathcal{H}} \leq \Delta\}$$

$$\rightarrow \text{ERM: } \min_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n L(y_i, \langle w, \Phi(x_i) \rangle)$$

$$\text{s.t. } \|w\|_{\mathcal{H}} \leq \Delta$$

$$\equiv \min_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n L(y_i, \langle w, \Phi(x_i) \rangle) + \lambda \|w\|_{\mathcal{H}}$$

representer theorem $w^* = \sum_{j=1}^n \alpha_j \underbrace{k(x_j, \cdot)}_{\Phi(x_j)}$

$$\text{dual} \quad \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n L(y_i, \sum_{j=1}^n \alpha_j k(x_i, x_j))$$

$$\text{s.t. } \underbrace{\alpha \mathbf{k} \alpha}_{\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)} \leq \Delta^2$$

$$\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$$

Theorem 10.7 Mohri

Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a pds kernel, $\Phi: \mathcal{X} \rightarrow \mathcal{H}$ a feature mapping associated to k

and $\mathcal{H} = \{n \rightarrow w \cdot \Phi(x) \mid \|w\|_{\mathcal{H}} \leq \Delta\}$ assume that there exists $r > 0$ such that

$k(x, x) \leq r^2$ and $|f(x)| \leq \Delta r$ for all $x \in \mathcal{X}$. then for any $\delta > 0$ with probability of at least $1 - \delta$, we have for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_n(h) + \frac{8r^2 \Delta^2}{\sqrt{n}} \left(1 + \frac{1}{2} \sqrt{\frac{\log \frac{1}{\delta}}{2}} \right)$$

loss Function:
 $|y - \hat{y}|$ or $|y - \hat{y}|^2$

$$R(h) \leq \hat{R}_n(h) + \frac{2\sigma^2 \Delta^2}{\sqrt{n}} \left(\sqrt{\frac{\text{trace}(k)}{nr^2}} + \frac{3}{4} \sqrt{\frac{\log \frac{n}{\delta}}{2}} \right)$$

⊗ نظر میاد با محدود کردن $\|w\|_2$ برای ب دست مارک

Ridge Regression

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^n (y_i - w^T x_i - b)^2 + \lambda \|w\|_2^2$$

جوابت فرم بسته دار

→ ERM on $\{n \rightarrow w^T x_i + b \mid \|w\|_2 \leq \Delta\}$

Lasso Regression

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^n (y_i - w^T x_i - b)$$

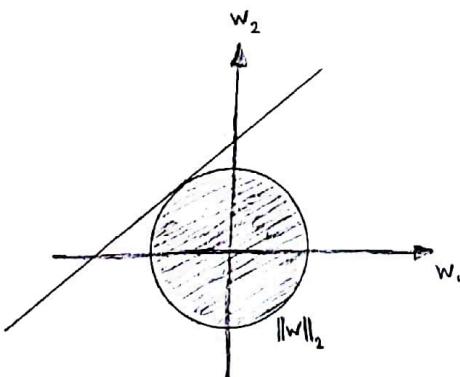
$$\text{s.t. } \|w\|_0 \leq s$$

\equiv

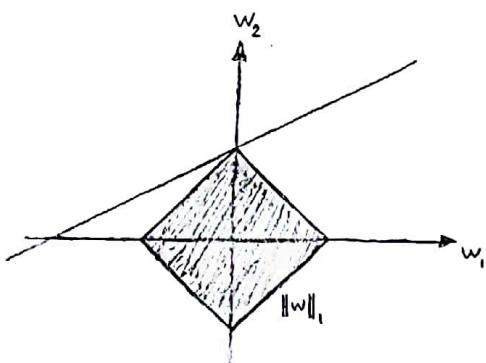
$$\text{s.t. } \|w\|_1 \leq \Delta$$

$$\begin{aligned} y_1 &= w^T x_1 + \xi_1 \\ &\vdots \\ y_n &= w^T x_n + \xi_n \end{aligned}$$

($\Delta \neq s$) ابتدئاً مساحت ممکن را با s و Δ بازدید کنید و حل کرد و دویناً معادل هستن



Ridge regression \Rightarrow محدود



Lasso regression \Rightarrow محدود

امان و Sparse و محدود