Instructor : Avishek Ghosh

Course website : figure it out (content, assignments @ website, moodle)

Tue / Fri — 5.30 – 7 pm

Extra classes — Saturday 3.30 – 5 pm

Grading :

Homeworks	20%	(2–3)
Midsem	30%	
Endsem	30%	
Scribes	15%	
Class Participation	5%	

What is this course about ?

1. Analyzing ML algorithms from a statistical point of view
2. Involves a large of statistical tools / techniques that can be used independently

NOT about

1. Theory of Deep Learning
2. Not particularly algorithmic

Reference :

1. High dimensional Statistics — Martin Wainwright
2. Asymptotic Stats — A. W. Vanderbaart

Classification (Binary)

Given data points (X_1, Y_1) (X_2, Y_2) .... (X_n, Y_n) where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ (features)

$Y_i \in \{-1, +1\}$

Goal : Find a classifier $g : \mathcal{X} \to \{-1, +1\}$

How to obtain this ? — Notion of loss function

Binary loss (error) : $\mathbb{1}[g(x) \neq Y] = \begin{cases} 1, & \text{if } g(x) \neq Y \\ 0, & \text{if } g(x) = Y \end{cases}$

Construct "Empirical loss" $L_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{g(x_i) \neq y_i\}$

We select classifier for which $L_n(g)$ is minimized

$$\hat{g}_n = \arg\min_{g \in \mathcal{C}} L_n(g)$$, where $\mathcal{C}$ denotes family of classifier

<u>Problem</u> : 1. Performance on "unseen" data is not considered

2. $\mathcal{C}$ can be complicated, n can be much smaller

Upto this point, purely empirical (no statistics)

<u>Statistical Model</u>

We assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d samples from a joint distribution $\mathcal{D}$ having same distribution as (X, Y)

Then, for a classifier $g : \mathcal{X} \to \{\pm 1\}$, we can write,

$$\underbrace{L(g)}_{\text{Avg loss / expected loss}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\mathbb{1}\{g(X) \neq Y\}\right] = \mathbb{P}(g(X) \neq Y)$$

It is a good idea to study $L(\hat{g}_n)$

2 Questions :

1. Is $L(\hat{g}_n)$ comparable to $\overbrace{\inf_{g \in \mathcal{C}} L(g)}^{\text{Naive Bayes classifier}}$?

whether $\hat{g}_n$ is comparable with the best classifier in $\mathcal{C}$

2. Is $L(\hat{g}_n)$ comparable to $L_n(\hat{g}_n)$?

Comparison between "in-sample" error and average error for $\hat{g}_n$

Assume $g^* = \arg\min_{g \in \mathcal{C}} L(g)$ {Naive Bayes}

We write $L(\hat{g}_n) = L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + \overbrace{L_n(\hat{g}_n)}^{\text{lesser loss}} - L(g^*)$

$$\leq L(g^*) + \underbrace{L(\hat{g}_n) - L_n(\hat{g}_n)}_{\wedge} + \underbrace{L_n(g^*) - L(g^*)}$$

$$\Rightarrow L(\hat{g}_n) - L(g^*) \leq \sup_{g \in \mathcal{C}} |L(g) - L_n(g)| + \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$$

$$\Rightarrow \quad L(\hat{g}_n) - L(g^*) \le 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)| \qquad - (*)$$

Remark : ① is controlled by $(*)$

② $L(\hat{g}_n) - L_n(\hat{g}_n) \le \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$

Remark : Performance of $\hat{g}_n$ is governed by $\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$.

We use uniform law of large numbers to handle this

Empirical Process Theory

1. Uniform law of large numbers
2. Uniform central limit theorem

Uniform Law of large Nos.

Suppose $X_1, X_2, \ldots, X_n$ i.i.d random objects taking value in $\mathcal{X}$. Let $\mathcal{F}$ be class of real-valued function on $\mathcal{X}$, what can we say about

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| = Z$$

In particular,

① whether $Z \to 0$ when n is large ?

② Can we obtain non-asymptotic guarantees ? i.e. guarantees for every n

③ Can we provide conditions on f s.t. Z converges to 0 ?

Connection to ML and statistics

① Binary Classification —

$$X_i \mapsto (X_i, Y_i)$$

$$\mathcal{F} \mapsto \{ \mathbb{1}\{g(x) \ne y\} : g \in \mathcal{C} \}$$

② M - estimation

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i)$$

think as −ve of loss fn.

where $x_1 \cdots x_n$ are i.i.d observations, Θ is parameter space

m_θ are real valued fn parametrized by θ.

Examples,
1. $m_\theta(x) = \log p_\theta(x)$ ← Maximum likelihood estimator (MLE)
2. $m_\theta(x) = -(x-\theta)^2$ ← sample mean (Mean estimator)
3. $m_\theta(x) = -|x-\theta|$ ← Median estimator

In mean estimation, target quantity for $\hat{\theta}_n$ is

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}\, m_\theta(x)$$

Similar to binary classification example we want $\underbrace{d(\hat{\theta}_n, \theta^*)}_{\text{distance b/w } \hat{\theta}_n \text{ and } \theta^*}$ to be small

It turns out $d(\hat{\theta}_n, \theta^*)$ is governed by $2 \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i) - \mathbb{E}\, m_\theta(x) \right|$

which is an instance of uniform law of large numbers.

Strategy to Control Z

① Key observation : Z concentrates around $\mathbb{E} Z$ i.e. $\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E} f(x) \right|$

(concentration of measure)

② We control $\mathbb{E} Z$ through techniques like symmetrization (Rademacher Complexity) or chaining (VC dimension)

Remark : (Asymptotic result) $\mathcal{F}$ is called "Gliven Ko – Cantelli" if $Z \to 0$ almost-surely as $n \to \infty$

Assumption : $\sup_x |f(x)| \leq B \quad \forall f \in \mathcal{F}$

McDiarmid's inequality

Suppose $X_1, X_2, \ldots X_n$ and $g : \mathcal{X} \times \ldots \times \mathcal{X}_n \to \mathbb{R}$ satisfies "bounded difference".

$$|g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i$$

$$\forall x_1 \ldots x_n \quad \forall i \in [n]$$

Then we have

$$\mathbb{P}\{g(x_1 \ldots x_n) - \mathbb{E}\, g(x_1 \ldots x_n) \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\le -t \qquad \le \quad " \quad — \quad "$$

<u>Rmk</u> : The bounded difference says that a f^n that is not too sensitive on any of its argument concentrates.

<u>Apply Mcdiarmid's to Z</u>

$$Z = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right|$$

we construct

$$g(x_1 \ldots x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E} f(x) \right|$$

$$g(x_1 \ldots x_i' \ldots x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j \ne i} f(x_j) + \frac{f(x_i')}{n} - \mathbb{E} f(x) \right|$$

$$= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^n f(x_j) - \mathbb{E} f(x) + \frac{f(x_i')}{n} - \frac{f(x_i)}{n} \right|$$

$$\le g(x_1 \ldots x_n) + \sup_{f \in \mathcal{F}} \left| \frac{f(x_i')}{n} \right| + \sup_{f \in \mathcal{F}} \left| \frac{f(x_i)}{n} \right|$$

$$\therefore \left| g(x_1 \ldots x_n) - g(x_1 \ldots x_1' \ldots x_n) \right| \le \boxed{\frac{2B}{n} = c_i}.$$

by switching x_i, x_i' to get both side

Hence, g satisfies bounded difference $\Rightarrow$ apply mcdiarmid's inequality.

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i \frac{4B^2}{n^2}}\right) = \exp\left(-\frac{nt^2}{2B^2}\right)$$

$||^{ly}$ $$\mathbb{P}(Z - \mathbb{E}Z \leq -t) \leq \underbrace{\exp\left(-\frac{nt^2}{2B^2}\right)}_{=\delta}$$

Then we say w.p. $\geq 1-\delta$,

$$Z \leq \mathbb{E}Z + \underbrace{B\sqrt{\frac{2}{n}\log\frac{1}{\delta}}}_{t}$$

t (very small bcz $\propto \frac{1}{\sqrt{n}}$)

Remark: We need to control $\mathbb{E}\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(x)\right|$

To control Z ① $Z \rightarrow \mathbb{E}Z$ ✓

② control $\mathbb{E}Z$?

Concentration Inequality (Hoeffding)

Suppose indep. $X_1 \dots X_n$ r.v. such that $a_i \leq X_i \leq b_i$ $\underbrace{\text{almost surely}}_{\text{w.p. } 1}$

where $a_1, \dots a_n, b_1, \dots, b_n$ are real numbers. Then for any $t \geq 0$

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof: Let $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$. Fix $\lambda \geq 0$. we have

$$\mathbb{P}(S \geq t) = \mathbb{P}(e^{\lambda S} \geq e^{\lambda t}) \underset{\text{Markov}}{\leq} \frac{\mathbb{E}(e^{\lambda S})}{e^{\lambda t}} = e^{-\lambda t}\,\mathbb{E}e^{\lambda t}$$

$$= \exp(-\lambda t + \psi_S(t)) \quad — ①$$

HW: Find Ω for which $e^{\lambda s} \geq e^{\lambda t}$ is true but $S \ngeq t$ $\exists x : t$

$$\psi_S(t) = \log \mathbb{E} e^{\lambda S}$$ (log-mgf / cumulant fn)

$$\psi_S(\lambda) = \log \mathbb{E}\, e^{\lambda \sum_i (X_i - \mathbb{E}X_i)} \xrightarrow{\text{independence}} = \log \prod_{i=1}^{n} \mathbb{E}\, e^{\lambda(X_i - \mathbb{E}X_i)}$$

$$= \sum_{i=1}^{n} \underbrace{\log \mathbb{E}\, e^{\lambda(X_i - \mathbb{E}X_i)}}$$

$$= \sum_{i=1}^{n} \psi_{X_i - \mathbb{E}X_i}(\lambda) \quad \text{— (11)}$$

Fix i, we analyze

$\psi_{\underbrace{X_i - \mathbb{E}X_i}_{U}}(\lambda)$. Need to bound ψ_U

$b_i - \mathbb{E}X_i \le U \le a_i - \mathbb{E}X_i$ almost surely.

Taylor series, $\psi_U(\lambda) = \cancel{\psi_U(0)}^{0} + \cancel{\psi_U'(0)}^{0}(\lambda) + \overbrace{\psi_U''(c)\left(\frac{\lambda^2}{2}\right)}^{\text{key term.}}$, $\quad 0 < c < \lambda$

* $\psi_U(0) = \log \mathbb{E}\, e^{0(X_i - \mathbb{E}X_i)} = 0$

* $\psi_U'(0) = \dfrac{d}{d\lambda} \log \mathbb{E}\, e^{\lambda U} = \dfrac{1}{\mathbb{E}e^{\lambda U}} \mathbb{E}(e^{\lambda U} \cdot U)$

Put $\lambda = 0$ $\quad = \dfrac{\mathbb{E}[U]}{(\ )} = 0 ..$

$$\psi_U(\lambda) = \psi_U''(c) \frac{\lambda^2}{2}$$

$$\psi_U''(\lambda) = \frac{d}{d\lambda} \frac{1}{\mathbb{E}e^{\lambda U}} \mathbb{E}(U e^{\lambda U}) = \frac{\mathbb{E}(e^{\lambda U})\, \mathbb{E}(U^2 e^{\lambda U}) - \mathbb{E}(U e^{\lambda U})^2}{\mathbb{E}(e^{\lambda U})^2}$$

We will show : $\psi_U''(\lambda) \ge 0$ for any λ

Consider a r.v. V whose density w.r.t. density of U is $e^{\lambda U} / \mathbb{E} e^{\lambda U}$

$$\underbrace{p_V}_{\text{density fn for random var. } V} = p_u \frac{e^{\lambda u}}{\mathbb{E} e^{\lambda u}} \qquad (u = 0 \Rightarrow p_u = 0 \Rightarrow p_V = 0)$$

$$\psi_U''(\lambda) = \mathbb{E}(V^2) - (\mathbb{E}V)^2 \leftarrow \text{variance of } V$$

$$\geq 0 .$$

Rmk : Support of U is $a_i - \mathbb{E}X_i$, $b_i - \mathbb{E}X_i$

support of V is " ——— " $(P_U = 0 \Rightarrow P_V = 0)$

Exercise : For any random variable V bound,

$$Var(V) \leq \left(\frac{b_i - a_i}{4}\right)^2 \leftarrow \tfrac{1}{2} \text{ pr } a_i, \ \tfrac{1}{2} \ b_i$$

$$\Rightarrow \psi_U(\lambda') = Var(V) \leq \frac{(b_i - a_i)^2}{4}$$

we have, $\boxed{\psi_U(\alpha) = \frac{\lambda^2}{2}\psi_U''(x) \leq \frac{\lambda^2}{8}(b_i - a_i)^2}$

Now, substituting (ii)

$$\psi_S(\lambda) = \sum_{i=1}^n \psi_{X_i - \mathbb{E}X_i}(\lambda) \leq \sum_{i=1}^n \frac{\lambda^2}{8}(b_i - a_i)^2$$

Substituting this in (i),

$$Pr(S \geq t) \leq \exp\left(-\lambda t + \sum_{i=1}^n \frac{\lambda^2}{8}(b_i - a_i)^2\right)$$

optimizing over λ, put $\lambda^* = \dfrac{4t}{\sum (b_i - a_i)^2}$

Putting $\lambda = \lambda^*$,

$$Pr(S \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

To get other side of concentration put $Y_i = -X_i$

Reading assignment:

1. Go through proof of lec 2
2. Revise martingales

Central Limit Theorem

$X_1, X_2, \ldots, X_n$ independent with mean μ, variance σ^2

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

(Reading: Convergence of r.v.
convergence in distribution)

CLT: $\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow[\text{distribution}]{n\to\infty} N(0,1)$

Suppose n is very large (for $t \geq 0$)

$$\mathbb{P}\left(\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \geq t\right) \approx \mathbb{P}\left(N(0,1) \geq t\right)$$

Scaling r.v. by σ, variance scales by σ^2.

$$\mathbb{P}\left(\sqrt{n}\left(\overline{X}_n - \mu\right) \geq t\right) \approx \mathbb{P}\left(N(0,\sigma^2) \geq t\right)$$

We use MGF based method to upperbound

$$\mathbb{P}(N(0,\sigma^2) \geq t) \leq \mathbb{P}\left(e^{\lambda N(0,\sigma^2)} \geq e^{\lambda t}\right)$$

$$\leq e^{-\lambda t}\, \mathbb{E}\, e^{\lambda N(0,\sigma^2)}$$

$$= \exp\Big(-\lambda t + \underbrace{\psi_{N(0,\sigma^2)}}_{\text{log-mgf / cumulant}}(\lambda)\Big)$$

$\psi_{N(0,\sigma^2)}(\lambda) = \log \mathbb{E}\, e^{\lambda N(0,\sigma^2)}$

$$\mathbb{E}\, e^{\lambda N(0,\sigma^2)} = e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{(Exercise)}$$

Using this,

$$\mathbb{P}(N(0,\sigma^2) \geq t) \leq \exp\Big(\underbrace{-\lambda t + \frac{\sigma^2\lambda^2}{2}}_{\text{optimal } \lambda}\Big)$$

$$\boxed{\mathbb{P}(N(0,\sigma^2) \geq t) \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)}$$

* CLT: $\boxed{\mathbb{P}\left(\sqrt{n}(\overline{X}_n - \mu) \geq t\right) \lessapprox \exp\left(-\frac{t^2}{2\sigma^2}\right)}$ — I (CLT)

Using hoeffding's we get,

$X_1 \dots X_n$, $\mathbb{E}[X] = \mu$, $Var(X) = \sigma^2$, $a \le X \le b$ a.s.

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}(X) \ge t'\right) \le \exp\left(\frac{-2t'^2}{n(b-a)^2}\right)$$

Let $t' = \sqrt{n} \cdot t$

$$\Rightarrow \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \ge \frac{\sqrt{n}.t}{n}\right) \le \exp\left(-\frac{2t}{(b-a)^2}\right)$$

$$\Rightarrow \boxed{\mathbb{P}\left(\sqrt{n}\left(\bar{X}_n - \mu\right) \ge t\right) \le \exp\left(\frac{-2t}{(b-a)^2}\right)} \quad \text{II} \quad \text{(Hoeffding's)}$$

Comparison :

If $a \le X_i \le b$ a.s., then

$$\sigma^2 \le \frac{(b-a)^2}{4}$$

CLT : $\exp\left(\frac{-t^2}{2\sigma^2}\right)$

Hoeffding's : $\exp\left(\frac{-2t^2}{(b-a)^2}\right)$

With this, CLT bound is $\underbrace{\exp\left(\frac{-t^2}{2\sigma^2}\right)}_{\text{(CLT)}} \le \underbrace{\exp\left(\frac{-2t^2}{(b-a)^2}\right)}_{\text{(Hoeffding)}}$.

Remarks : * Error upper bound matches b/w CLT, hoeffding

* σ^2 can be much smaller than range, makes CLT bound sharper/stronger

* CLT holds asymptotically and is an approximation. On the other hand, hoeffding's is exact, non-asymptotic (holds for any n)

Berry Esses Thm : How close apx w.r.t CLT (approximation scales $\sim 1/\sqrt{n}$) (v. deep thm — not in course)

Note : MGF based techniques we use — (Cramer– chernoff Method.)

Confidence Interval

Suppose $X_1 \dots X_n$ i.i.d random variables with $\mathbb{E} X = \mu$, $\text{Var } X = \sigma^2$

$a \le X_i \le b$ almost surely.

Problem : We know (a, b, σ^2) and want to estimate μ.

We obtain a set (interval), also known as confidence interval (CI), where μ lies w.h.p.

CLT based bound (classical) (for $t \ge 0$)

By CLT, $\mathbb{P}\left(\left| \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \right| \le t \right) \xrightarrow{n \to \infty} \mathbb{P}(|N(0,1)| \le t)$

(union of two events)

Put $t = z_{\alpha/2}$ (α-quantile) s.t. $\mathbb{P}(|N(0,1)| \le z_{\alpha/2}) = 1 - \alpha$

w.p. $1-\alpha$,

$$-z_{\alpha/2} \le \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \le z_{\alpha/2}$$

$$\Rightarrow \boxed{\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}} \quad \text{C.I}$$

Length of interval $\frac{2\sigma}{\sqrt{n}} z_{\alpha/2}$ shrinks with n ($\to 0$ as $n \to \infty$)

Hoeffding's for C.I.

$$\mathbb{P}\left(\left| \sqrt{n} (\bar{X}_n - \mu) \right| \ge t \right) \le \underbrace{2 \exp\left(\frac{-2t^2}{(b-a)^2} \right)}_{\alpha}$$

i.e. $t = (b-a) \sqrt{\frac{1}{2} \log \frac{\alpha}{2}}$

So, w.p. $1-\alpha$,

$$\boxed{|\bar{X}_n - \mu| \le \frac{1}{\sqrt{n}} (b-a) \sqrt{\frac{1}{2} \log \frac{\alpha}{2}}}$$

μ lies in $[\bar{X}_n - \text{"} , \bar{X}_n + \text{"}]$ C.I.

Length of interval,

$\sigma \le \frac{b-a}{2}$

so, CLT is better (asymptotic)

Sub-Gaussian

Assume $X \sim \mathcal{N}(\mu, \sigma^2)$. We know that

$$\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\lambda^2\sigma^2/2} \quad \text{for any } \lambda.$$

This motivates us to define class of r.v. exhibiting similar properties.

Def: A r.v. X w/ mean μ is called sub-gaussian if there exists a positive number σ such that

$$\mathbb{E}\, e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2} \quad \text{for all } \lambda.$$

It is denoted as $X \sim \text{subG}(\sigma)$. σ is called the parameter of sub-Gaussian r.v.; σ^2 is a proxy for variance.

Examples ① Gaussian

② Rademacher r.v. :

$\epsilon \in \{-1, +1\}$ with equal probability

We want to show that ϵ is 1 sub gaussian

$\ell! \sim \ell^\ell$

$(2k)!$ much bigger than $2^k k!$

$$\mu = 0, \quad \sigma = 1 \qquad \mathbb{E}\, e^{\lambda\epsilon} = \frac{e^{\lambda}+e^{-\lambda}}{2} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{(\lambda^2)^k}{k!\,2^k}$$

$$\leq e^{\lambda^2/2}$$

③ Bounded r.v.

Suppose $a \leq X \leq b$ a.s.

Show that $X \sim \text{SubG}\left(\frac{b-a}{2}\right)$

Last class: $\log \mathbb{E}\, e^{\lambda X} \leq \frac{(b-a)^2}{8}\lambda^2$

$$\Rightarrow \mathbb{E}\, e^{\lambda X} \leq \exp\left(\lambda^2 \frac{(b-a)^2}{8}\right) \leftarrow \text{exactly what we want.}$$

Hoeffding's inequality for sub Gaussian [not just for bounded r.v.]

Suppose $X_1 \cdots X_n$ are sub $G(\sigma_i)$ and $\mathbb{E} X_i = \mu_i$ then,

for all $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} (X_i - \mathbb{E} X_i) \geq t \right) \leq \exp\left(\frac{-t^2}{2 \sum_{i=1}^{n} \sigma_i^2} \right)$$

$$\mathbb{P}\left(\sum_{i=1}^{n} (\quad '' \quad) \leq -t \right) \leq \quad ''$$

Remark : $\sigma_i = \frac{b_i - a_i}{2}$, we get back hoeffding's for bounded r.v. (proved in lec2)

Proof : Use Cramer-chernoff technique, bound on $\mathbb{E} e^{\lambda X}$ comes from defn of sub-gaussian directly!

Reading exercise :

Chap 2 from HDS (Martin) — Sub-gaussian (bigger class: sub-exponential)

— Martingales.

17/1

Last time : Concentration of measure

- Hoeffding's ineq for bounded r.v.
  CLT
- Confidence Interval (C.I.) construction
- Sub-Gaussian X is subG(σ)

  if $\mathbb{E} e^{\lambda X} \leq e^{\lambda^2 \sigma^2 / 2} \; \forall \lambda$.

  - Hoeffding's ineq for sub-gaussian

Properties of Sub-Gaussian r.v.

① $X_1, X_2 \sim$ subG(σ_1), SubG(σ_2) and they are independent w/ mean μ_1, μ_2

$$X_1 + X_2 \sim \mathbb{E}\, e^{\lambda(X_1 + X_2 - (\mu_1 + \mu_2))} = \mathbb{E}\, e^{\lambda(X_1 - \mu_1)}\, \mathbb{E}\, e^{\lambda(X_2 - \mu_2)} \leq e^{\frac{\lambda^2}{2}(\sigma_1^2 + \sigma_2^2)}$$

$$\sim \text{sub}G\left(\sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Reading exercise : Sub-exponential r.v. (chapter 2 from HDS)

* Property : If $X \sim SubG$ then $X^2 \sim$ sub exponential

Norm : $Y = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$, where $X_1 \dots X_d \sim SubG$

$\|Y\|_2^2 = X_1^2 + \dots + X_d^2 \sim$ Sub exp.

Martingale Based Concentration ← Doob 1956 (rejected Shannon's paper :p)

* Bounded, Sub-G , Sub-exp . we require $X_1, \dots, X_n$ to be independent

* Independence doesn't hold in many settings – online Learning / Time-Series

* Need to deal w/ dependent r.v's

Martingale

$\underbrace{F_1}_{Y_1} \subseteq \underbrace{F_2}_{Y_2} \subseteq \dots$ (sets)

Pair $(Y_k, F_k)_{k=1}^{\infty}$ is called a martingale if:

random variables

1. $\mathbb{E}[|Y_k|] < \infty$

2. $\mathbb{E}[Y_{k+1} \mid F_k] = Y_k$ a.s. for $k \geq 1$

Ex : ① (Partial Sum)

Let $x_1, x_2, x_3, \dots$ be i.i.d random variables w/ mean μ

Define, $Y_k = \sum_{i=1}^{n} x_i - k\mu$ w/ finite variance

This is a martingale because

① $\mathbb{E}[|Y_{k+1}|] < \infty$

② $\mathbb{E}[Y_{k+1} \mid \underbrace{X_1 \dots X_k}_{\text{contains all information abt } Y_k}] = \mathbb{E}[\sum_{i=1}^{k} X_i - k\mu + X_{k+1} - \mu \mid X_1 \dots X_k]$

$= \mathbb{E}(Y_k \mid X_1 \dots X_k) + \underbrace{\mathbb{E}(X_{k+1} - \mu \mid X_1 \dots X_k)}_{0}$

$= Y_k$

Example : (Doob Martingale)

Given a sequence of independent random variables $\{X_k\}_{k=1}^{n}$. We define

$Y_k = \mathbb{E}[f(X) \mid X_1 \dots X_k]$ for $k = 1 \dots n$ and

$Y_0 = \mathbb{E}[f(X)]$; where $X = (X_1, \dots, X_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$

with $\mathbb{E}|f(X)| < \infty$

We now show this is a martingale

$Y_n = f(X)$. So, $f(X) - \mathbb{E}[f(X)] = Y_n - Y_0$

$$= \sum_{k=1}^{n} (\underbrace{Y_k - Y_{k-1}}_{\text{difference}})$$

Claim : $\{Y_k\}_{k=1}^{n}$ is a martingale w.r.t. $\{X_k\}_{k=1}^{n}$

$\mathcal{F}_k = \sigma(X_1 \dots X_k)$ ← set w/ all possible functions of $X_1 \dots X_k$ (σ-field)

over $X_1 \dots X_k$ ← absorbs $|X_1 \dots X_k$

① $\mathbb{E}[|Y_k|] = \mathbb{E}\,|\mathbb{E}(f(X) \mid X_1 \dots X_k]|$

$\leq \mathbb{E}\,|f(X)| < \infty$

triangle inequality/jensen

Prop : $\mathbb{E}\,\mathbb{E}(Y \mid X] = \mathbb{E}(Y)$ (law of iterated expectation)

② $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[\mathbb{E}(f(X) \mid X_1 \dots X_{k+1}] \mid X_1 \dots X_k]$

Tower rule : smaller set wins

$= \mathbb{E}[f(X) \mid X_1 \dots X_k]$

$= Y_k$

Tower rule : $\mathbb{E}[\mathbb{E}[Z \mid S_1] \mid S_2] = \mathbb{E}[Z \mid S_1]$ if $S_1 \subseteq S_2$

Martingale difference

A sequence $(D_k, \mathcal{F}_k)_{k=1}^{\infty}$ is called martingale difference if D_k is adapted to $\mathcal{F}_k$ and

① $\mathbb{E}|D_k| < \infty$ ⑪ $\mathbb{E}(D_{k+1} \mid \mathcal{F}_k) = 0$

Natural way : $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ martingale , $D_k = Y_k - Y_{k-1}$

① $\mathbb{E}[D_k] = \mathbb{E}[|Y_k - Y_{k-1}|] \le \mathbb{E}|Y_k| + \mathbb{E}|Y_{k-1}| < \infty$

② $\mathbb{E}[D_{k+1} | \mathcal{F}_k] = \mathbb{E}[Y_k - Y_{k+1} | \mathcal{F}_k] = Y_k - Y_k = 0$ a.s.

Thm (Azuma - Hoeffding's inequality) [Hoeffding's for martingale differences]

Let $(D_k, \mathcal{F}_k)_{k=1}^{\infty}$ be a martingale difference sequence where $a_k \le D_k \le b_k$ a.s. for $k = 1 \dots n$. Then for all $t \ge 0$,

$$\mathbb{P}\left[\sum_{k=1}^{n} D_k \ge t\right] \le \exp\left(\frac{-2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right) \quad \text{and}$$

$$\mathbb{P}\left[\sum_{k=1}^{n} D_k \le -t\right] \le \exp\left(\frac{-2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right)$$

* Bounded martingale difference concentrates.

* Azuma - Hoeffdings for Martingale sequences — skipping, similar.

Proof : Similar to Hoeffding's bound for bounded r.v., in particular we use Cramer – Chernoff method

Let $S = \sum_{k=1}^{n} D_k$; $\mathbb{P}(S \ge t) \le \mathbb{P}(e^{\lambda S} \ge e^{\lambda t}) \overset{(\text{markov})}{\le} e^{-\lambda t}\, \mathbb{E}\, e^{\lambda S}$

$$\le \exp(-\lambda t + \psi_S(\lambda)) \quad \text{where } \psi_S(\lambda) = \log \mathbb{E}\, e^{\lambda S}$$

$$= \log \mathbb{E}\, e^{\lambda \sum_{k=1}^{n} D_k}$$

Let us look at

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n} D_k} \,\middle|\, \mathcal{F}_{n-1}\right] = \mathbb{E}\left[e^{\lambda D_n} e^{\lambda \sum_{k=1}^{n-1} D_k} \,\middle|\, \mathcal{F}_{n-1}\right]$$

$$= e^{\lambda \sum_{k=1}^{n-1} D_k} \underbrace{\mathbb{E}\left[e^{\lambda D_n} | \mathcal{F}_{n-1}\right]}_{\text{MGF of a mean zero, bounded random variable}}$$

Using same technique as hoeffding ←

$$\le \left(e^{\lambda \sum_{k=1}^{n-1} D_k}\right) \cdot \exp\left(\frac{\lambda^2}{8}(b_n - a_n)^2\right)$$

Last time we were looking at

$$\mathbb{E} \sup_{f \in \mathcal{F}} P_n f - Pf := \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E} f(x) \right|$$

Symmetrization   chaining

* Rademacher complexity :

for a set A, draw n elements $a_1 \cdots a_n$

(Empirical) Rademacher avg : $\hat{R}_n(A) = \mathbb{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i \right|$

ε_i: i.i.d. ± 1 w/equal prob (w.e.p)

* Examples : ℓ_2 - Ball ($1/\sqrt{n}$), ℓ_1 - Ball ($1/n$)

Today :- Connect R. C. to U.L.L.N

- Symmetrization argument
- Bounds of $\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf|$

Empirical Process Setup

- $X_1, X_2, \cdots, X_n \sim$ i.i.d $\mathcal{D}$
- $\mathcal{F}$: class of real valued functions

Let,

$$\mathcal{F}(X_1, X_2, \ldots, X_n) = \{ f(X_1), \ldots, f(X_n) : f \in \mathcal{F} \}$$

A random subset of $\mathbb{R}^n$ ← compute rademacher complexity of this

$$\hat{R}_n(\mathcal{F}(X_1 \ldots X_n)) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \varepsilon_i \right| \quad \text{(Empirical Rademacher Complexity)}$$

$$R_n(\mathcal{F}) = \mathbb{E}_{X_1 \ldots X_n} \hat{R}(\mathcal{F}(X_1 \ldots X_n)) \leftarrow \text{Rademacher Complexity of } \mathcal{F}$$

$$= \mathbb{E}_{X_1 \ldots X_n} \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \varepsilon_i \right| \quad \text{— (1)}$$

Note : $\mathcal{F}$ is a class of fn, $\mathcal{F}(X_1 \cdots X_n)$ is a random subset of $\mathbb{R}^n$

Theorem (Symmetrization)

We have $\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq 2 R_n(\mathcal{F})$ where $R_n(\mathcal{F})$ is defined in (1)

Proof : $X_1, \dots, X_n \sim$ i.i.d (X)

Withdraw $(X_1', \dots, X_n')$ from same distribution (X') s.t. $(X_1' \dots X_n') \perp\!\!\!\perp (X_1, \dots, X_n)$

component-wise

(independent)

$$\mathbb{E}_X f(X) = \mathbb{E}_{X'}\left(\frac{1}{n} \sum_{i=1}^n f(X_i')\right) \quad \text{— (I)}$$

$$* \quad \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_X f(X) \right|$$

Using (I)

independent of X'

$$= \mathbb{E}_X \sup_{f \in \mathcal{F}} \Big(\frac{1}{n} \sum_{i=1}^n f(X_i) - \underbrace{\mathbb{E}_{X'}\Big(\frac{1}{n} \sum_{i=1}^n f(X_i')\Big)}_{\text{constant}} \Big)$$

$$\mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X'}\left(\frac{1}{n} \sum_{i=1}^n f(X_i) - f(X_i') \right) \right|$$

(Symmetrization argument)

* $\sup(\cdot)$ is convex, $|\cdot|$ is convex, so use jensens inequality

$$\leq \mathbb{E}_{X, X'} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X_i')) \right|$$

Claim : $f(X_i) - f(X_i')$ is distributed identically to

$$(f(X_i) - f(X_i'))\, \varepsilon_i$$

Intuition : ε_i flips sign, but X_i, X_i' i.i.d, so they switch also w.p. $1/2$.

$$= \mathbb{E}_{X, X', \varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(X_i')) \right|$$

$\sup(A+B) \leq \sup(A) + \sup(B)$

$$\leq \mathbb{E}_{X, X', \varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i)) \right| + \quad ''$$

$$= 2 R_n(\mathcal{F})$$

∎

Remark : In most cases we condition on $X_1 \cdots X_n$ $(x_1 \cdots x_n)$. So, in this setting
$$R_n(\mathcal{F}) = \hat{R}_n(\mathcal{F})$$

Simple Bounds on $R_n(\mathcal{F})$

Lemma (Massat's Lemma) : Suppose A is a finite subset of $\mathbb{R}^n$ with cardinality $|A|$. Then,
$$R_n(A) = \mathbb{E} \max_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right| \le \sqrt{\frac{6 \log(2|A|)}{n}} \max_{a \in A} \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2}$$

Proof : For non-negative X, we write
$$\mathbb{E} X = \int_0^\infty P(X > x)\, dx \qquad (\text{Exercise})$$

Notation : $\sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2} = \|a\|_2$, $\sum_{i=1}^n \frac{\varepsilon_i a_i}{n} = \frac{a^T \epsilon}{n} = a^T \tilde{\varepsilon}$ $\left(\tilde{\varepsilon} = \frac{\varepsilon}{n}\right)$

$$\mathbb{E} \exp\left[\frac{(a^T \tilde{\varepsilon})^2}{6\|a\|_2^2}\right] = \int_0^\infty P\left\{\exp\left(\frac{(a^T\tilde{\varepsilon})^2}{6\|a\|_2^2}\right) > x\right\} dx$$

$$= \int_0^1 \underbrace{P(\cdot)}_{\le 1} dx + \int_1^\infty P\left\{\underbrace{|a^T \tilde{\varepsilon}|}_{\frac{1}{n}\sum_{i=1}^n a_i \varepsilon_i} > 6\|a\|_2 \sqrt{\log(x)}\right\} dx$$

(=1 actually)

$\frac{1}{n}\sum_{i=1}^n a_i \varepsilon_i \leftarrow$ mean 0 so direct hoeffding

$$= 1 + \underbrace{2}_{\text{due to } |a^T\tilde{\varepsilon}|} \int_1^\infty \exp\left(-\frac{2 \times 6\|a\|_2^2 \log x}{4\|a\|_2^2}\right) dx \qquad -1 \le \varepsilon_i \le 1$$

$$= 1 + 2\int_1^\infty x^{-3}\, dx = 2 \qquad \text{— (I)}$$

We have,
$$\mathbb{E} \exp\left[\max_{a \in A} \frac{|a^T \tilde{\varepsilon}|^2}{6\|a\|_2^2}\right] = \mathbb{E} \max_{a\in A} \exp \frac{|a^T\tilde{\varepsilon}|^2}{6\|a\|_2^2} \qquad (e^x \text{ is increasing})$$

$$\le \sum_{i=1}^n \mathbb{E} \exp\left[\frac{(a^T\tilde{\varepsilon})^2}{6\|a\|_2^2}\right] \le 2|A| \qquad \text{— (II)}$$

(III) can be re-written as, $(\max\{a,b\})^2 = \max\{a^2, b^2\}$ for $a,b \geq 0$

$$\mathbb{E} \exp\left(\left[\max_{a\in A} \frac{|a^T \tilde{\varepsilon}|}{\sqrt{6}\,\|a\|_2}\right]^2\right) \leq 2|A| \qquad \text{— (IV)}$$

* Fact : Show that $x \mapsto e^{x^2}$ is convex $(x > 0)$ (composition)

Using jensen's inequality, $\max\left(\frac{a}{b}\right) \geq \frac{\max(a)}{\max(b)}$

$$\exp\left(\left[\mathbb{E} \max_{a\in A} \frac{|a^T \tilde{\varepsilon}|}{\sqrt{6}\,\|a\|_2}\right]^2\right) \leq \mathbb{E} \exp\left(\left[\max_{a\in A} \frac{|a^T \tilde{\varepsilon}|}{\sqrt{6}\,\|a\|_2}\right]^2\right) \leq 2|A|$$

$$\mathbb{E} \max_{a\in A} |a^T \varepsilon| \leq \sqrt{6 \log 2|A|}\ \max \|a\|_2$$

(some bug in pf ← in eqn (II))

Application of ULLN

Apply Mossart's Lemma to $R_n(\mathcal{F})$

Assumption : $\mathcal{F}$ is Boolean , $f(x) \in \{0,1\}\ \forall x \in \mathcal{X}, f \in \mathcal{F}$

(*) Classification, testing

** Let us fix $\mathcal{F}$ and $(X_1 \cdots X_n) = (x_1 \cdots x_n)$

$$\mathbb{E} \sup_{f\in\mathcal{F}} |P_n f - Pf| \leq 2R_n(\mathcal{F}) = 2R_n(x_1 \cdots x_n)$$

$$\leq \sqrt{\frac{6 \log\left(2\,|\mathcal{F}(x_1 \cdots x_n)|\right)}{n}}\ \underbrace{\max_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2}}_{\leq 1}$$

$$\mathcal{F}(x_1 \cdots x_n) = \{f(x_1) \cdots f(x_n) : f \in \mathcal{F}\}$$

* If $|\mathcal{F}(x_1 \cdots x_n)| \sim 2^n$

$$\mathbb{E} \sup_{f\in\mathcal{F}} |P_n f - Pf| \leq \sqrt{\frac{6 \log 2}{n} + 6 \log 2}$$

Rmk : Growth of $\mathcal{F}(x_1 \cdots x_n)$ is important (different from the size of function class)

* Polynomial discrimination $\quad P(n) = n^{\alpha}$

$$|\mathcal{F}(x_1 \cdots x_n)| \sim \mathrm{poly}(n) \qquad \mathbb{E} \sup_{f\in\mathcal{F}} |P_n f - Pf| \leq \sqrt{\frac{\alpha \log n}{n}}$$

Checked by VC dimension.

Last class : Symmetrization to show that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \leq 2 R_n(\mathcal{F}(x_1 \cdots x_n))$$

where $\mathcal{F}(x_1 \cdots x_n) = \{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n$

We condition on $X_1 = x_1 ; \cdots ; X_n = x_n$

$\rightarrow \mathbb{E}_{X_1 \cdots X_n} R_n(\mathcal{F}(X_1 \cdots X_n))$ can use instead

(Recall)

Lemma (Massart's Lemma) : Suppose A is a finite subset of $\mathbb{R}^n$ with cardinality $|A|$. Then,

$$R_n(A) = \mathbb{E} \max_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right| \leq \sqrt{\frac{6 \log(2|A|)}{n}} \max_{a \in A} \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2}$$

Using this, $R_n(\mathcal{F}(x_1 \cdots x_n)) \lesssim \sqrt{\frac{\log(|\mathcal{F}(x_1 \cdots x_n)|)}{n}} \max_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(x_i)}$

ignoring universal constants

* Want upper bound on $|\mathcal{F}(x_1 \cdots x_n)|$

Assumption : $\mathcal{F}$ is boolean. $f(x_i) \in \{0, 1\}$ $\forall x_i, f \in \mathcal{F}$

* $\mathcal{F}$ Boolean, $|\mathcal{F}(x_1 \cdots x_n)| \leq 2^n$ Is this useful ? (no)

$\Downarrow$ $R_n(\ ) \lesssim 1$

set or $\mathcal{F}$ is not learnable (no matter how many samples)

* Polynomial Discrimination :

$\mathcal{F}$ has polynomial discrimination if there exists a polynomial $p(\cdot)$ s.t.

$$|\mathcal{F}(x_1 \cdots x_n)| \leq p(n)$$

Ex : If $p(n) = n^{\alpha}$, then $R_n(\mathcal{F}) \lesssim \sqrt{\frac{\alpha \log n}{n}}$

Q : How to check if $\mathcal{F}$ has polynomial discrimination ?

$\hookrightarrow$ VC dimension (Vapnik Chervonenkis) [A combinatorial object]

Defn (VC dimension)

Shattering : A finite subset $\{x_1 \ldots x_m\}$ is said to be shattered by a Boolean class $\mathcal{F}$ if

$$|\mathcal{F}(x_1 \cdots x_m)| = 2^m \quad ; \quad \mathcal{F}(x_1 \cdots x_m) = \{0,1\}^m$$

The VC-dimension D of $\mathcal{F}$ is the maximum integer D for which any set of $\{x_1 \cdots x_D\}$ is shattered by $\mathcal{F}$.

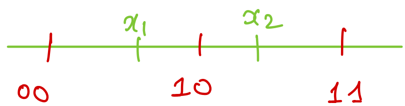
* If $\{x_1 \cdots x_D\}$ shattered for every D, VC dim is ∞.

Examples : ① Intervals on $\mathbb{R}$

$$S_{left} = \{ \mathbb{1}_{(-\infty, a]} : a \in \mathbb{R} \} \rightarrow \quad D = 1 \quad \{x_1\} \text{ is shattered}$$

where $\mathbb{1}_{(-\infty,a]}(x) = \begin{cases} 1, & \text{if } x \le a \\ 0, & \text{o/w} \end{cases}$

$$\mathbb{1}_{(-\infty, x_1 - \epsilon)}, \ \mathbb{1}_{(-\infty, x_1 + \epsilon)}$$

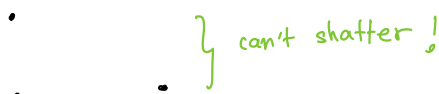


(cannot enumerate 01)

② Extend to $\mathbb{R}^2$

$$S_{rect} = \{ \mathbb{1}_{[a_1, b_1] \times [a_2, b_2]} : a_i, b_i \in \mathbb{R} \}$$

$$VC(S_{rect}) = 2$$



③ **Lemma** : Let $\mathcal{V}$ as a D-dimensional vector space on real functions on $\mathcal{X}$

Let, $\mathcal{F} = \{ \mathbb{1}_{(f \ge 0)} : f \in \mathcal{V} \}$. Then, $VC(\mathcal{F})$ is atmost D.

Pf : Let $\{x_1 \ldots x_{D+1}\}$ and consider $T = \{ f(x_1) \cdots f(x_{D+1}) : f \in \mathcal{V} \}$

$\therefore \exists$ some coefficients $\alpha \in \mathbb{R}^{D+1} \neq 0$.

$$\sum_{i=1}^{D+1} \alpha_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{V} \quad \longleftarrow V^* \text{ (evaluation = dual space)}$$

Since $\alpha \neq 0$, wlog, $\exists$ index k s.t. $\alpha_k > 0$

Let's assume $\mathcal{F}$ shatters $\{x_1 \cdots x_{D+1}\}$, then there exist $f \in \mathcal{V}$ s.t.

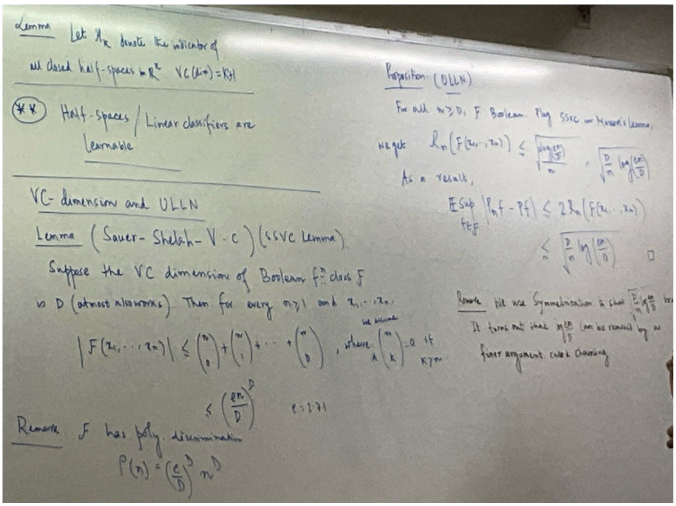
$$f(x_i) < 0 \quad \text{for all } i \text{ s.t. } \alpha_i > 0$$
$$f(x_i) \geq 0 \quad \text{"} \qquad \text{"} \quad \alpha_i \leq 0$$

With this,

$$\sum_{i=1}^{D+1} \alpha_i f(x_i) = \underbrace{\sum_{\because \alpha_i > 0} \alpha_i f(x_i)}_{<0} + \underbrace{\cdots}_{\leq 0} < 0$$

a contradiction. ■

VC dimension and ULLN



Remark: SSVC lemma can be proved using induction and "down-shifting"
(read pf in HDS - Chapter 4 Prop. 4.18)

Chaining

Covering and Packing :

Let (T, ρ) denotes a metric space T with associated metric ρ.

$(\rho : T \times T \to \mathbb{R})$

ⓐ $\rho(\theta, \tilde{\theta}) \geq 0$, $\rho(\theta, \tilde{\theta}) = 0$ iff $\theta = \tilde{\theta}$ (non negativity)

ⓑ $\rho(\theta, \tilde{\theta}) = \rho(\tilde{\theta}, \theta)$ (symmetric)

ⓒ $\rho(\theta, \tilde{\theta}) + \rho(\tilde{\theta}, \bar{\theta}) \geq \rho(\theta, \bar{\theta})$ (Triangle inequality)

Ex : $\mathbb{R}^d$ (or subset)

$$\rho(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|_2 = \left(\sum_{j=1}^d (\theta_i - \tilde{\theta}_i)^2\right)^{1/2}$$

Ex : Boolean cube $= \{0,1\}^d$

$$\rho(\theta, \tilde{\theta}) = d_H(\theta, \tilde{\theta}) = \#\text{ coordinates at which } \theta, \tilde{\theta} \text{ differ} = \sum_{j=1}^d \mathbb{1}(\theta_j \neq \tilde{\theta}_j)$$

$$\text{Normalised hamming dist} = \frac{\sum_{j=1}^d \mathbb{1}(\theta_j \neq \tilde{\theta}_j)}{d}$$

Ex : $C[0,1]$: set of all continuous functions in $[0,1]$

Metric : (Sup-norm) $\sup_{x \in [0,1]} |f(x) - g(x)| = \rho(f,g) = \|f-g\|_\infty$

Ex : L^2 : space of square integrable fns on $[0,1]$

$$\|f-g\|_2 = \left[\int_0^1 (f(x) - g(x))^2 \, dx\right]^{1/2}$$

Ex : $L^2(\mu ; [0,1])$

$$\|f-g\|_2 = \left[\int_0^1 (f(x) - g(x))^2 \, d\mu(x)\right]^{1/2}$$

Covering Number : A δ-cover of a set T w.r.t ρ is a set $\{\theta^1, \dots, \theta^N\} \subset T$ s.t. $\forall\, \theta \in T\ \exists\, \theta^i$ s.t. $\rho(\theta, \theta^i) \leq \delta$

Covering Number, $N(\delta, T, \rho)$ is the cardinality of the smallest cover.

$*$ $\delta_1 \leq \delta_2$, $N(\delta_1, T, \rho) \geq N(\delta_2, T, \rho)$

Ex. Unit hypercube

$d=1$: $[-1,1]$, $\rho(\theta, \theta') = |\theta - \theta'|$

$$N(\delta, [-1,1], \rho) \leq \underbrace{1/\delta + 1}_{\text{a valid cover construction.}} \quad (\text{break into intervals of } 2\delta)$$

$$\left\lceil \frac{2 - 2\delta}{2\delta} \right\rceil + 1$$

Extend to d

$$N(\delta, [-1,+1]^d, \|\cdot\|_\infty) \leq \left(1 + \frac{1}{\delta}\right)^d \leftarrow \text{due to grid}$$

Ex : Binary Hypercube : $\{-1,+1\}^d = H_d$

$$\rho(\theta, \theta') = \frac{1}{d} \sum_{j=1}^{n} \mathbb{1}_{\{\theta_j \neq \theta_j'\}}$$

fix δd — take all combo. $(1-\delta)d$

$$N_H(\delta, H_d; \rho) \leq \underbrace{2^{\lceil (1-\delta)d \rceil}}_{\text{turns out to be a lower bound}}$$

$T(S) = \{\theta \in H_d \mid \theta_j = 0\ \forall j \notin S\}$

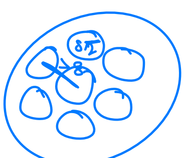
$S = \{1, 2, \dots, \lceil (1-\delta)d \rceil\}$

Match $\tilde{\theta}$ exactly on $S \Rightarrow d_H \leq \delta$.

Packing Number

A δ-packing of a set T w.r.t. a metric ρ is $\{\theta^1, \dots, \theta^N\} \subset T$ s.t. $\rho(\theta^i, \theta^j) > \delta\ \forall\ i,j \in [N], i \neq j$

Packing Number $M(\delta, T, \rho)$ is the cardinality of maximum such set.



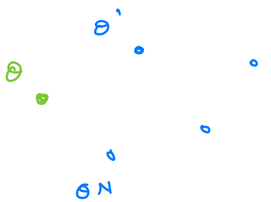
Lemma: For $\delta > 0$ we have

$$M(2\delta, T, \rho) \underset{(b)}{\leq} N(\delta, T, \rho) \underset{(a)}{\leq} M(\delta, T, \rho)$$

Remark: These numbers are orderwise equivalent.

Pf: (a) Let $\{\theta^1 \dots \theta^N\}$ be maximal set of δ-separated pts for (T, δ)

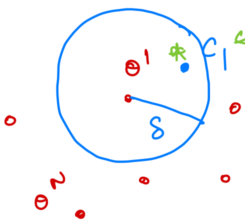
$M(\delta, T, \rho) = N$. Also a valid cover



all dist $\geq \delta$

$\Rightarrow$ Add to $M(\delta, T, \rho)$ to get a better packing number !

(b) Let $\{\theta^1, \dots, \theta^N\}$ be min set of pts δ-cover



~~exactly~~ atmost one center $\Big\}$ every center in ≥ 1 ball (cover)

$$\Rightarrow M(2\delta, T, \rho) \leq N(\delta, T, \rho)$$

■

Examples.

Proposition: Let $\|\cdot\|$ denote some norm $\mathbb{R}^d$. Also, $B_R = \{x \in \mathbb{R}^d, \|x\| < R\}$

Then $\delta > 0$, we have,

$$M(\delta R, B_R, \|\cdot\|) \leq \left(1 + \frac{2}{\delta}\right)^d$$

(easy to see for ∞ norm)

Proof: (Volumetric argument)

Let $x_1 \dots x_N$ denote any set of points in B_R that are δR separated

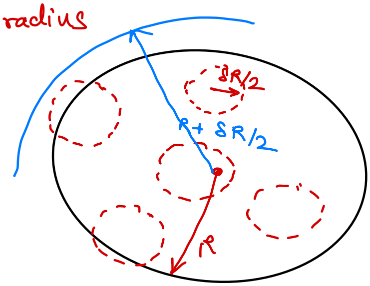
i.e., $\|x_i - x_j\| > \delta R \quad \forall\, i \neq j$

Then, the closed balls

$$\bar{B}\left(x_i, \frac{\delta R}{2}\right) = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x_i - x\| \leq \frac{\delta R}{2} \right\}$$

are disjoint

Moreover, all these balls are contained within a ball of radius



So, $M(\cdot)\cdot c\left(\frac{\delta R}{2}\right)^d \le c\left(R+\frac{\delta R}{2}\right)^d$

$$\Rightarrow M(\cdot) \le \left(1+\frac{2}{\delta}\right)^d$$

■

Normalize : $M(\delta, B_R, \|\cdot\|) \le \left(1+\frac{2R}{\delta}\right)^d$

Notation : Covering number / δ-net

Example : Cover / Pack a function class.

Proposition : Let $\Theta \subseteq \mathbb{R}^d$ be a non-empty subset with diameter D $\quad \rho(\theta_1,\theta_2) \le D \;\; \forall \theta_1,\theta_2 \in \Theta$

Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ satisfying $|f_{\theta_1}(x) - f_{\theta_2}(x)| \le \Gamma(x)\|\theta_1-\theta_2\|_2$

Fix a measure Q. $\rho^2(f,g) = \int_x (f(x)-g(x))^2 \, dQ(x)$

Then $\delta > 0$,

$$M(\delta, \mathcal{F}, \rho) \le \left(1 + \frac{2D\|\Gamma\|_Q}{\delta}\right)^d \quad \text{where } \|\Gamma\|_Q^2 = \int_x \Gamma(x)^2 \, dQ(x)$$

Pf : Need to find R.

$$\int_x \left(f_{\theta_1}(x) - f_{\theta_2}(x)\right)^2 dQ(x) \le \int_x \Gamma(x)^2 \|\theta_1-\theta_2\|^2 \, dQ(x)$$

$$\le \|\Gamma_Q\|^2 \|\theta_1 - \theta_2\|^2$$

$$\Rightarrow \rho(f_{\theta_1}, f_{\theta_2}) \le \|\Gamma\|_Q \|\theta_1-\theta_2\|_2 \quad \text{— ①}$$

From ①, to find a δ-packing for $\mathcal{F}$, it is sufficient to find a $\frac{\delta}{\|\Gamma_Q\|}$ packing in Θ.

$$M(\delta, \mathcal{F}, \rho) \le M\left(\frac{\delta}{\|\Gamma\|_q}, \Theta, \|\cdot\|_2\right) \rightarrow \left(\text{packing for lhs} \Rightarrow \text{packing for rhs.}\right)$$

$$B(a, D) = \{x \in \mathbb{R}^d \mid \|x - a\| \le D\}$$

$$\le M\left(\frac{\delta}{\|\Gamma\|_q}, B(a, D), \|\cdot\|_2\right) \overset{\text{previous lemma}}{\le} \left(1 + \frac{2D\|\Gamma\|_q}{\delta}\right)^d$$

$\blacksquare$