Stas Minsker

Intro

<u>machine learning</u> is the study of algorithms that can learn from data, gradually improving their performance. large datasets

<u>mathematical statistics</u> (ST) is the science of making decision in the face of uncertainty.

small datasets

[Ex] $Y = \alpha + \beta_1[\text{interesting}] + \beta_2[\text{genre}] + \beta_3[\text{Budget}] + \cdots + \varepsilon$

Y ↓ rating of a movie $\in [0, 100]$

α ↓ average rating (or ?)

Netflix: wants to predict rating of a movie - ML

Disney: wants to make a movie with high rating - ST

needs to understand whether the model is statiscally significant (hypothesis testing, etc. ...)

[Ex] Handwritten digits recognition - ml

ml:
- unsupervised learning — "raw data", no label — clustering, learning representation
- $\boxed{\text{supervised learning}}$ — labeled — classification, prediction
- reinforcement learning — "multi-armed bandit problem" — exploration v.s. exploitation

Realizable case:

Binary classification (X, Y) = (instance, label)
(observation, label)

[Ex] x - image, $Y \in \{+1, -1\}$ (eg: "cat", "dog")

$X \in S$ - set of all possible instances

statistical learning: we will assume that (X, Y) is random, in other words, it has a probility distribution P so we use language of probability theory

Superised Learning:

$(X, Y) \in S \times \{+1, -1\}$

P is the distribution of (X, Y)

i.e. $P(A) = \text{Probability}((X, Y) \in A)$

Π is the distribution of X

Imposing the probabilistic model on (X, Y) takes as info realm of Statistical Learning Theory

Goal: predict label Y based on the observation X

The prediction rule is a function $g: S \to \{-1, +1\}$

The quality of a prediction rule g is measured by the classification / generalization error

$L(g) = \text{Prob}(Y \neq g(X))$ (one prediction)

The training data is a sequence $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ of i.i.d. pairs with distribution P.

a prediction rule

An algorithm takes training data as an input and outputs $\hat{g}_n = \hat{g}_n((X_1, Y_1), \cdots, (X_n, Y_n))$

In general, we will consider 2 scenarios:

1) "Realizable" learning : there exists $g \in G$ s.t. $Y = g^*(X)$ with probability 1.

2) ? "Agnostic" learning : there is no $g \in G$ s.t. $Y = g^*(X)$ with probability 1.

Realizable scenario :

assume that the set G of all possible classification rules is finite.

By assumption, $\exists g^* \in G$: $Y = g^*(X)$ with prob 1

The Empirical Risk Minimization principle :

(training data) pick any $\hat{g}_n$ that agrees with the training data ($\hat{g}_n(X_i) = Y_i$, $i = 1, \cdots, n$)

Question : what is $L(\hat{g}_n)$?

what is $\text{prob}(L(\hat{g}_n) > \varepsilon) \leq$? , given $\varepsilon > 0$

Here, $L(g) = \text{Prob}(Y \neq g(X))$

let G_B = "bad" classification rules $= \{ g \in G : L(g) > \varepsilon \}$

$\text{prob}(L(\hat{g}_n) > \varepsilon) = \text{prob}(\hat{g}_n \in G_B)$

Takes $g \in G_B$, if $\hat{g}_n = g$, $g(X_i) = Y_i$, $i = 1, 2, \cdots, n$

$\text{prob}(g(X_1) \neq Y_1) > \varepsilon$

$\text{prob}(g(X_1) = Y_1) \leq 1 - \varepsilon$

$$\Rightarrow \text{prob}(g(X_i) = Y_i, i = 1, \cdots, n) = P(g(X_1) = Y_1) \cdot P(g(X_2) = Y_2) \cdots P(g(X_n) = Y_n)$$
$$= \prod_{i=1}^{n} \Pr(g(X_i) = Y_i)$$
$$\leq (1 - \varepsilon)^n$$

We show that $\forall g \in G_B$, $\Pr(\hat{g}_n = g) \leq (1 - \varepsilon)^n \leq e^{-\varepsilon n}$

(since $1 - \varepsilon \leq e^{-\varepsilon}$)

Remainder : Union Bound : $P(A \cup B) \leq P(A) + P(B)$

$$\therefore \text{ If } G_B = \{g_1, \cdots, g_k\}, \text{ then } P(\hat{g}_n \in G_B) = P(\hat{g}_n = g_1 \text{ or } \hat{g}_n = g_2, \text{ or } \cdots, \text{ or } \hat{g}_n = g_k)$$
$$\leq P(\hat{g}_n = g_1) + P(\hat{g}_n = g_2) + \cdots + P(\hat{g}_n = g_k)$$
$$\leq k e^{-\varepsilon n}$$
$$\leq |G| e^{-\varepsilon n}$$

If one requires that $P(L(\hat{g}_n) \le \varepsilon) \geq 1-\delta$

then $|G| e^{-\varepsilon n} \le \delta$

$\Leftrightarrow n \geq \frac{\log \frac{|G|}{\delta}}{\varepsilon}$

E.g. if $G = 1024 = 2^{10}$, $\delta = 2^{-6}$, $\varepsilon = 0.01$, what is n? (最少有多少样本)

$n \geq \frac{16 \log 2}{0.01} = 160 \log 2$

Record some useful concepts:

ERM: pick a classifier that makes the smallest number of mistakes on observed data.

define: $L_n(g) = \frac{1}{n} \cdot \#\{1 \le j \le n : g(x_j) \neq Y_j\}$

Remark: By Law of Large Number, $L_n(g) \xrightarrow{P} L(g)$ since $\mathbb{E}\, L_n(g) = L(g)$

The ERM states that one pick $\hat{g}_n$ that minimize $L_n(g)$ over $g \in G$

PAC: ("Probably Approximately Correct") Learnability: Leslie Valiant '84

A class H of hypothesis / binary classifiers is PAC learnable:

if $\forall\, \varepsilon, \delta \in (0,1)$, any labeling function g^*, any distribution Π of X,

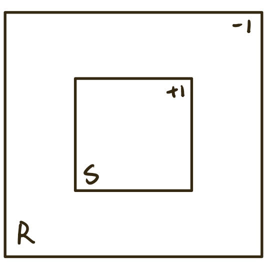
$\exists$ algorithm A. (e.g. ERM) and a function $\underbrace{n = n(\varepsilon, \delta, G)}_{\text{called sample complexity}}$ s.t. $Pr(L(\hat{g}_n) > \varepsilon) \le \delta$

推论: Any $\underbrace{\text{finite}}_{\text{有}|G|}$ set of binary classifiers is PAC-learnable

if we take $n \geq \frac{\log(\frac{|G|}{\delta})}{\varepsilon}$, we can satisfy $Pr(L(\hat{g}_n) > \varepsilon) \le \delta$

Infinite set of classifiers

[Example A]



Area(R) = 2 (大方块)

Area(S) = 1

$g^*(x) = \begin{cases} +1 & x \in S \\ -1 & x \in R \setminus S \end{cases}$

let $G = \{$all binary function $g: R \to \{+1, -1\}\}$

Training Data: $(X_1, Y_1), \cdots, (X_n, Y_n)$

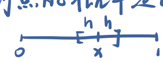
consider $\hat{g}_n(x) = \begin{cases} Y_i, & x = X_i \text{ for } i = 1, \cdots, n \\ -1, & \text{else} \end{cases}$

In particular, $\hat{g}_n$ is consistent with ERM (sample 里见过的都对)

But $\underset{\text{loss}}{L}(\hat{g}_n) = Pr(Y \neq \hat{g}_n(X)) = \frac{1}{2}$ → overfitting

(assume that X is chosen uniformly from R) 选到 samples 点的概率是0 ⇒ 几乎全是-1

$X \sim U(0,1)$



$Pr(X = x) \le Pr(X \in [x-h, x+h])$ $\forall h$

since h can be as small as it wants to

$\therefore Pr(X = x) = 0$

not all infinite $|G|$ is not PAC learnable ?

overfitting : $L(\hat{g}_n) \neq 0$?

if G is too "large", then any algorithm (in particular ERM) will produce a classifier with large misclassification error.

If G is "too large" → it is not PAC learnable

An algorithm A is a map that takes G and (X_i, Y_i), $i=1,\dots,n$ as input and outputs $\hat{g} \in G$.

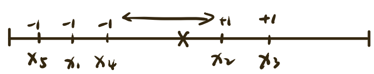
[Example B] $x \in [0,1]$

$$g^*(x) = \begin{cases} -1, & x \le x^* \\ +1, & x > x^* \end{cases}$$

$G = \{ g_y, \ y \in [0,1] \}$ ← 有定义的，不是 all classifier，但仍 infinite classifiers

$$g_y(x) = \begin{cases} -1, & x \le y \\ +1, & x > y \end{cases}$$ 只能分两段

$(X_1, Y_1), \dots, (X_n, Y_n)$ iid – training data



claim: G is PAC learnable ?

→ To show this, we need to estimate its sample complexity

given $\varepsilon, \delta > 0$, if $n \ge n(\varepsilon, \delta)$, $\exists A$ such that, given a sample of size n, A outputs $\hat{g}$ such that $L(\hat{g}) < \varepsilon$ with prob $\ge 1-\delta$

let's use ERM: specifically, let $\hat{x} = \max\{X_j : Y_j = -1\}$

let $$\hat{g}_n(x) = \begin{cases} -1, & x \le \hat{x} \\ +1, & x > \hat{x} \end{cases}$$

$\Pr(L(\hat{g}) > \varepsilon) \le (1-\varepsilon)^n$ (show this!)

[Example 1] Task: identify counterfeit banknotes

We know that real banknotes (a) color change under the light $\in [0,1]$, with increment 0.1

(b) red/blue fibers $\in [0, 100]$, with increment 1

realizable ← Assume that for every "real" banknote, (a)(b) belong to a specific range, vice versa

Assume that you have 100 banknotes, known to be real or not

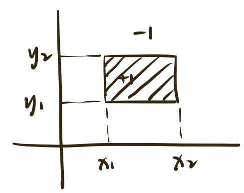
→ X : banknote, Y : label {real / fake}

$x \in S$ Domain set $= \{(x, y) : x \in [0:0.1:1],\ y \in [0:1:100]\}$

Training data: 100 banknotes

Hypothesis class G: $g : S \to \{+1, -1\}$

$$g(x) = \begin{cases} +1, & x \in [x_1, x_2],\ y \in [y_1, y_2] \\ -1, & \text{else} \end{cases}$$



$|G| \le \binom{11 \times 100}{2} \div 2$ — still manageable

Assume that we want a classifier that makes at most 5% mistake.

What is the probability that you will get such a classifier from a sample size 100.

We proved that

$$\Pr(L(\hat{g}_{ERM}) > 0.05) \le \underbrace{\frac{10^8}{2}}_{\text{代用}|G|}(1-0.05)^{100}$$

$$= \frac{10^8}{2} \cdot \frac{95^{100}}{10^{200}}$$

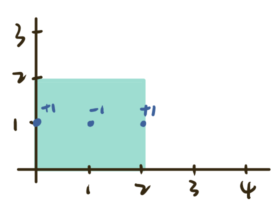
prob 小于1而是什么东西?

[Example 2] label (x_1, x_2). $x_1, x_2 \in$ II, $0 \le x_1 \le 4$, $0 \le x_2 \le 3$

$G = \{$rectangle with vertices $(x_1, x_2) \in [0,4] \times [0,3])\}$

Training set:

x_1	x_2	y
0	1	1
1	1	-1
2	1	1



let $g: [0,2] \times [0,2]$

$$\underset{(n=3)}{L_n(g)} = \frac{\#\{1 \le i \le n,\ Y \ne g(x_i)\}}{n} = \frac{1}{3}$$ ← empirical risk

ERM: line segment (0,1) to (2,1)

→ Agnostic learning (no perfect classifier)

"No Free Lunch" theorem

An algorithm A is a mapping from training data $(x_i, Y_i)_{i=1}^{n}$ to the class G of binary classifiers

Theorem: Assume that S is finite. let $(x_1, Y_1), \cdots, (x_n, Y_n)$ be the training data such that $n \le \frac{|S|}{2}$.
Then for any algorithm A, $\exists$ some distribution π over S and $g^*: Y = g^*(x)$ but
$Pr(\hat{g}(x) \neq g^*(x)) \geq \frac{1}{8}$ with prob $\frac{1}{8}$ where $\hat{g} = A((x_i, Y_i)_{i=1}^{n})$ P(习得算法犯错概率 > $\frac{1}{8}$) > $\frac{1}{8}$

Proof: let $\mathcal{X} = \{(x_1, Y_1), \cdots, (x_n, Y_n)\}$

consider $\max_{g^*, \pi} E_{\mathcal{X}} \underbrace{E_{x \sim \pi} I\{\hat{g}(x) \neq g^*(x)\}}_{Pr(\hat{g}(x) \neq g^*(x))} \underset{\text{want}}{\geq} c > 0$ $\frac{1}{8}$

pick $g_*^{(1)}, g_*^{(2)}, \cdots, g_*^{(k)}$

$\max(\quad) \geq \frac{1}{k}\sum_{j=1}^{k} E_{\mathcal{X}} E_{x \sim \pi} I\{\hat{g}(x) \neq g_*^{(j)}(x)\}$ 最大值 > expected

consider a random g^* s.t. $\forall x \in S$, $g_*(x) = \begin{cases} +1 & \text{with prob } \frac{1}{2} \\ -1 & \text{with prob } \frac{1}{2} \end{cases}$ independently.

call this distribution Q

$$\max_{g^*, \pi} E_{\mathcal{X}} E_x I\{\hat{g}(x) \neq g_*(x)\} \geq E_{g^* \sim Q} E_{\mathcal{X}} E_x I\{\hat{g}(x) \neq g^*(x)\}$$
$$= E_{\mathcal{X}} E_x \underline{E_{g^* \sim Q} I\{\hat{g}(x) \neq g^*(x)\}}$$
$$\geq E_{\mathcal{X}} E_x \tfrac{1}{2} I\{x \notin \{x_1, \cdots, x_n\}\}$$

$E_{g^* \sim Q} I\{\hat{g}(x) \neq g_*(x)\} = \begin{cases} 0 & \text{if } x \in \{x_1, \cdots, x_n\} \\ \frac{1}{2} & \text{if } x \notin \{x_1, \cdots, x_n\} \end{cases}$ 如果见过，那么会是对的 / 如果没见过，那么有一半概率会对

② $\begin{cases} \geq E_{\mathcal{X}} \frac{1}{2} \cdot \frac{1}{2} \\ = \frac{1}{4} \end{cases}$

[lemma] let Z be a r.v. such that $0 \le Z \le 1$, then $Pr(Z \geq \delta) \geq E(Z) - \delta$

proof: $EZ = EZ \cdot I\{Z \le \delta\} + EZ \cdot I\{Z > \delta\}$
$\le \delta + Pr(Z > \delta)$

$\therefore$ Applying the lemma, we get that

$$\max_{g^*} Pr_{\mathcal{X}}\left(Pr(\hat{g}(x) \neq g_*(x)) \geq \tfrac{1}{8}\right) \geq \tfrac{1}{4} - \tfrac{1}{8} = \tfrac{1}{8}$$

$(\because E_{\mathcal{X}}(Pr(\hat{g}(x) \neq g^*(x))) = \frac{1}{4})$

Review of conditional probabilities and conditional expectations

[lemma] let $\mathbb{Z}$ be a r.v. s.t. $Var(\mathbb{Z}) < \infty$

Then, $E\mathbb{Z} = \underset{z \in \mathbb{R}}{\arg\min}\ E(\mathbb{Z} - z)^2$

Proof: let $f(z) = E(\mathbb{Z} - z)^2$

Then, $f'(z) = 2E(\mathbb{Z} - z) = 0 \iff z = E\mathbb{Z}$

Finally, $f''(z) = 2 \Rightarrow E\mathbb{Z}$ is the minimizer

(开口向上)

Now, let $\mathbb{Z}, W$ be such that $Var(\mathbb{Z})$ and $Var(W)$ are finite.

Then $E[\mathbb{Z} | W = y] = \underset{z \in \mathbb{R}}{\arg\min}\ E_{\mathbb{Z}|W=y} (\mathbb{Z} - z)^2$

Clearly, $z = z(y)$ above. Therefore, $E[\mathbb{Z} | W = y]$ is a function of W that minimizes $E[\mathbb{Z} - f(w))^2]$ over all funtions f.

[Exercise] let $y(w) = E[\mathbb{Z} | W]$. Prove that for any function g, $E(\mathbb{Z} - y(w)) \cdot g(w) = 0$

let $h(\mathbb{Z})$ be an arbitrary function of $\mathbb{Z}$ s.t. $Eh^2(\mathbb{Z}) < \infty$

then $E[(w - \hat{f}(\mathbb{Z})) \cdot h(\mathbb{Z})] = 0$ where $\hat{f}(\mathbb{Z}) = E[w | \mathbb{Z}]$

$\langle x, Y \rangle = \mathbb{E}(x, Y)$

Bayes Classifier

$y(x) = E[Y \mid X = x] \qquad Y \in \{+1, -1\}$

Theorem let S be a finite set. X has (discrete) distribution Π over S. Then the best possible binary classifier is given by $g^*(x) = \text{sign}(E(Y \mid X = x))$

$g^*(x)$ is known as <u>Bayes classifier</u>

Proof: let $g: S \to \{\pm 1\}$ be arbitrary

probability of $X=x$

Then $\Pr(Y \neq g(x)) = \sum_{x \in S} \Pr(Y \neq g(x) \mid X = x) \cdot \Pi(x)$

$\Pr(Y=1 \mid X=x) + \Pr(Y=-1 \mid X=x) = 1$

$$\left.\begin{cases} \Pr(Y=1 \mid X=x) = \dfrac{1+y(x)}{2} \\ \Pr(Y=-1 \mid X=x) = \dfrac{1-y(x)}{2} \end{cases}\right\} \Leftrightarrow \begin{array}{l} \Pr(Y=t \mid X=x) = \dfrac{1+ty(x)}{2}, \\ \Pr(Y \neq t \mid X=x) = \dfrac{1-ty(x)}{2}, \end{array} \quad t \in \{\pm 1\}$$

为正更好

Then $\underbrace{\Pr(Y \neq g(x))}_{\text{想要 minimize}} = \sum_{x \in S} \dfrac{1 - g(x)y(x)}{2} \cdot \Pi(x)$

$$\geq \sum_{x \in S} \frac{1 - |y(x)|}{2} \Pi(x)$$

Equality is achieved when $g(x)y(x) = |y(x)|$ for all $x \in S$

$\Leftrightarrow$ <u>$g(x) = \text{sign}(y(x))$</u> $\to$ Bayes Classifier

Bayes Risk

$$L^* = L(g^*) = \sum_{x \in S} \left(\frac{1 - |y(x)|}{2} \right) \Pi(x)$$

[Example] 5 cards are drawn at random. 2 are reviewed

$$Y = \begin{cases} 1, & 5 \text{ cards contain an Ace} \\ -1, & \text{otherwise} \end{cases}$$

Find the Bayes classifier and its risk.

Remark: The risk of the Bayes classifier is called the Bayes risk: $L^* = \Pr(Y \neq g^*(X))$

Solution: $S \in \{1, 0\} \to \begin{cases} x=1 & \text{if the pair of cards have at least 1 Ace} \\ x=0 & \text{otherwise} \end{cases}$

$y(x) = E[Y \mid X = x] = 1 \cdot \Pr(Y=1 \mid X=x) + (-1) \cdot \Pr(Y=-1 \mid X=x)$

$\Pr(Y=-1 \mid X=1) = 0$

$\Pr(Y=-1 \mid X=0) = \dfrac{\binom{46}{3}}{\binom{50}{3}} \approx 0.77$

$\Pr(Y=1 \mid X=1) = 1$

$\Pr(Y=1 \mid X=0) = 1 - \dfrac{\binom{46}{3}}{\binom{50}{3}}$

$\overset{x=0}{y(0)} = 1 \cdot (1 - 0.77) + (-1) \cdot 0.77 = -0.54$

$\overset{x=1}{y(1)} = 1 \cdot (1-0) + (-1) \cdot 0 = 1$

$g^*(x) = \text{sign}(y(x)) = \begin{cases} 1, & x=1 \\ -1, & x=0 \end{cases}$

$P(X=1) = 1 - P(\text{both cards are not Aces}) = 1 - \frac{48}{52} \cdot \frac{47}{51} = 0.15$

$P(X=0) = 1 - 0.15 = 0.85$

$$L^* = \frac{1-|\eta(0)|}{2} \cdot \pi(0) + \frac{1-|\eta(1)|}{2} \cdot \pi(1)$$
$$= \frac{1-0.54}{2} \cdot 0.85 + \frac{1-1}{2} \cdot 0.15$$
$$\approx 0.2$$

Agnostic PAC-learnability

A class G of binary classifiers is agnostic PAC-learnable (APACL) if $\exists\, m(G;\varepsilon,\delta)$ and an algorithm A such that $\forall \varepsilon, \delta > 0$, any distribution P of (X, Y), $L(\underbrace{\hat{h}_n}_{\text{output of } A}) \le \underbrace{\min_{g \in G} L(g)}_{L(g^*)} + \varepsilon$ with prob $\ge 1-\delta$ as long as $n \ge m(G;\varepsilon;\delta)$

Next goal: understand which classes are PAC learnable. We will start with finite classes, and then will study infinite classes.

Key Idea: concept of uniform closeness of the true and empirical risks.

Definition The training set $X = (X_1, Y_1), \cdots, (X_n, Y_n)$ is called ε-representative if $\forall g \in G$, $|L_n(g) - L(g)| \le \varepsilon$

$L(g) = \Pr(Y \ne g(X))$

$L_n(g) = \frac{1}{n} \cdot \#\{1 \le j \le n;\ g(X_j) \ne Y_j\} = \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \ne g(X_j)\}$

$|\text{empirical risk} - \text{true risk}| \le \varepsilon$

[Lemma] Assume that X is $\frac{\varepsilon}{2}$ representative, and let $\hat{g}_n$ be the minimizer of the empirical risk: $\hat{g}_n = \underset{g \in G}{\operatorname{argmin}}\, L_n(g)$. Then

$$L(\hat{g}_n) = \Pr(Y \ne \hat{g}_n(X) \mid X) \le \min_{g \in G} L(g) + \varepsilon$$

[proof] let $\bar{g} = \underset{g \in G}{\operatorname{argmin}}\, L(g)$. Then

$$L(\hat{g}_n) = L_n(\hat{g}_n) + L(\hat{g}_n) - L_n(\hat{g}_n)$$
$$\le L_n(\bar{g}) + \max_{g \in G} |L(g) - L_n(g)|$$

$L_n(\hat{g}_n) \le L_n(\bar{g})$

已是empirical最优解

$$\le L(\bar{g}) + \frac{\varepsilon}{2} + |L_n(\bar{g}) - L(\bar{g})|$$
$$\le L(\bar{g}) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$
$$\le L(\bar{g}) + \varepsilon$$

Bias- Complexity Tradeoff

This concept refers to the following error decomposition:

let $\hat{g}$ be the output of a learning algorithm A given the training data $\mathcal{X} = (X_1, Y_1), \cdots, (X_n, Y_n)$

Then $L(\hat{g}) = Pr(Y \neq \hat{g}(X) | \mathcal{X})$

$$= L(\hat{g}) - \min_{g \in G} L(g) + \underbrace{\min_{g \in G} L(g)}_{\geq \text{ Bayes Risk } L^*}$$

If G is large, $\min_{g \in G} L(g)$ is small. But $L(\hat{g}) - \min_{g \in G} L(g)$ is large

Finite Classes are agnostic PAC learnable

Question: What is the smallest sample size sufficient to guarantee that it is ε-representative with probability at least $1-\delta$?

Assume $|G| < \infty$, then $Pr(\forall g \in G, |L_n(g) - L(g)| \leq \varepsilon)$

$$= 1 - Pr(\exists g \in G, |L_n(g) - L(g)| > \varepsilon)$$

why can apply union bound ?

$Pr(\exists g \in G, |L_n(g) - L(g)| > \varepsilon)$

$= Pr(\bigcup_{g \in G} \{|L_n(g) - L(g)| > \varepsilon\})$

$\leq \sum_{g \in G} Pr(|L_n(g) - L(g)| > \varepsilon)$

we need to show the measure of $L_n(g)$ is concentrated around its expected value

$\leq |G| \max_{g \in G} Pr(|L_n(g) - L(g)| > \varepsilon)$

apply chebyshev's inequality: $P(|X - E(X)| \geq k) \leq \frac{\sigma^2}{k^2}$

Fix $g \in G$, $Pr(|\frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - L(g)| > \varepsilon)$

$L(g) = E L_n(g)$

$Z_j = I\{Y_j \neq g(X_j)\}$

$Z_1, \cdots, Z_n$ are i.i.d. $\to \mathbb{Z}$

$$Pr(|\frac{1}{n}\sum Z_j - E\mathbb{Z}| > \varepsilon) \leq \frac{Var(\frac{1}{n}\sum Z_j)}{\varepsilon^2}$$

$$Var(\frac{1}{n}\sum Z_j) = \sum Var(\frac{1}{n} Z_j) = n \cdot \frac{1}{n^2} Var(\mathbb{Z}) = \frac{Var(\mathbb{Z})}{n}$$

$\leq |G| \frac{Var(\mathbb{Z})}{n\varepsilon^2} \to P(\mathcal{X}$ is not ε-representative$)$

$$\frac{|G| Var(\mathbb{Z})}{n\varepsilon^2} \leq \delta \Rightarrow n \geq \frac{|G| Var(\mathbb{Z})}{\delta \varepsilon^2}$$

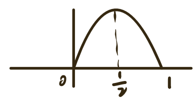
! Bad Bound

$Pr(Z=1) = Pr(Y \neq g(x)) = P_g$

$Pr(Z=0) = 1 - P_g$

$Var(Z) = P_g(1-P_g) \leq \frac{1}{4}$ → suppose Bernoulli distribution

$f(x) = x(1-x)$



$\therefore Pr(\mathcal{X}$ is not ε-representative $) \leq |G| \cdot \frac{1}{4n\varepsilon^2}$

[Exercise] $|G| = 1000$

we want $\mathcal{X}$ ε-representative with prob at least 0.9 $(\delta = 0.1)$ and $\varepsilon = 0.1$

$|G| \cdot \frac{1}{4n\varepsilon^2} \leq 0.1 = \delta$

$n \geq \frac{|G|}{4\delta\varepsilon^2} = \frac{10^3}{4\times 10^{-3}} = 250{,}000$ 很大, bad estimation

$\therefore \mathcal{X}$ is ε-representative with prob $1-\delta$

$\Leftrightarrow \hat{g}_n$ obtained by ERM satisfy $L(\hat{g}_n) \leq \min_{g\in G} L(g) + 2\varepsilon$ with prob $1-\delta$

Hoeffding's Inequality

Lemma 4.5 (Hoeffding's Inequality). *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-2m\epsilon^2/(b-a)^2\right).$$

apply Hoeffding's inequality to question:

let θ_i be the random variable $L_n(g)$, $\theta_1, \cdots, \theta_n$ are i.i.d.

(∵ $L_n(g)$ 选取 sample data 也是 i.i.d. 的)

$L_n(g) = \frac{1}{n}\sum_{i=1}^{n}\theta_i$, $L(g) = EL_n(g) = \mu$

$\therefore P(|\frac{1}{n}\sum_{i=1}^{n}\theta_i - \mu| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$

$\therefore Pr(\exists g \in G, |L_n(g) - L(g)| > \varepsilon) \leq \sum_{g\in G} 2e^{-2n\varepsilon^2} = 2|G|e^{-2n\varepsilon^2}$

$\therefore n \geq \frac{\log(\frac{2|G|}{\delta})}{2\varepsilon^2}$ then $Pr(\exists g \in G, |L_n(g) - L(g)| > \varepsilon)$

对于上面 E.x., $n \geq \frac{1}{2} \cdot 100 \cdot \log(\frac{2\cdot 10^3}{0.1}) \leq 2000$

more reasonable estimation

[Exercise] $S = \{0,1,2,3,4\}$ X is binomial $B(4, \frac{1}{2})$, $Y \in \{+1, -1\}$

$P(Y=1 \mid X=x) = \frac{1}{2}$ (label is random guess)

$x \sim B(n,p)$

$\Pr(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$

consider $g(x) = \begin{cases} 1 & , x \text{ is even} \\ -1 & , x \text{ is odd} \end{cases}$

interpretation: $L(g) = \Pr(Y \neq g(x)) = \frac{1}{2}$

$\eta(x) = E[Y \mid X=x] = 0$

$g^*(x) = \mathrm{sign}(\eta(x)) =$ either $+1$ or -1

Now, $P(Y=1 \mid X=x) = \begin{cases} \frac{3}{4} & x=0,1,2 \\ \frac{1}{4} & x=3,4 \end{cases}$

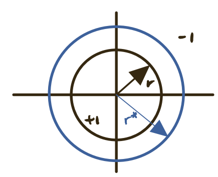
$\Rightarrow P(Y=-1 \mid X=x) = \begin{cases} \frac{1}{4} & x=0,1,2 \\ \frac{3}{4} & x=3,4 \end{cases}$

$L(g) = \Pr(Y \neq g(x)) = \sum \Pr(Y \neq g(x) \mid X=x) \cdot \Pr(X=x)$

$= \underbrace{(\frac{1}{2})^4}_{x=0} \times \frac{1}{4} + \underbrace{\binom{4}{1}\times(\frac{1}{2})^4}_{x=1} \times \frac{3}{4} + \underbrace{\binom{4}{2}\times(\frac{1}{2})^4}_{x=2} \times \frac{1}{4} + \underbrace{\binom{4}{3}\times(\frac{1}{2})^4}_{x=3} \times \frac{1}{4} + \underbrace{\binom{4}{4}\times(\frac{1}{2})^4}_{x=4} \times \frac{3}{4}$

[Exercise] $S = \mathbb{R}^2$, $G = \{g_r, r>0\}$

$g_r(x) = \begin{cases} +1 & , \|x\| \le r \\ -1 & , \|x\| > r \end{cases}$. Assume realizability



$Y = g_{r^*}(x)$

Show that G is PAC-learnable

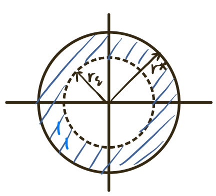
$\to$ show 2 things: (a) An algorithm $A \to$ ERM

(b) $m(\varepsilon, \delta)$ s.t. $\hat{g} = \hat{A}((X_1,Y_1), \cdots, (X_n,Y_n))$ satisfies $\Pr(L(\hat{g}) \ge \varepsilon) \le \delta$

ERM outputs $\hat{g}$ that minimizes $\#\{1 \le j \le n ; Y_j \neq g(x_j)\}$

$\therefore L_n(\hat{g}) = 0$

$\hat{r} = \min\{r > 0 : \|x_j\| < r \Leftrightarrow Y_j = +1\}$



let r_ε be s.t. $\Pr(r_\varepsilon \le \|x\| \le r_*) = \varepsilon$

(i) $\Pr(\|x\| \le r_*) < \varepsilon \Rightarrow$ Circle $(r=r^*)$ 面积小于 ε $\Rightarrow$ 所有 classifier loss 都小于 ε

(ii) $\Pr(\|x\| \le r_*) > \varepsilon$

$P(L(\hat{g}_r) \ge \varepsilon) = \Pr(\hat{r} < r_\varepsilon) = \Pr($there are no instances with label $+1$ in the ring between circle of radius r_ε and $r^*)$

$= \Pr(\|x_j\| > r_* \text{ or } \|x_j\| < r_\varepsilon \text{ for all } j)$

$= \prod_{j=1}^{n} \Pr(\|x_j\| > r_* \text{ or } \|x_j\| < r_\varepsilon)$

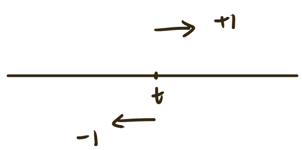
$= (1-\epsilon)^n \le e^{-\epsilon n}$

$(\because Pr(r_\epsilon \le \|x\| \le r^*) = \epsilon)$

To have $e^{-\epsilon n} \ge \delta$, $n \ge \dfrac{\log(\frac{1}{\delta})}{\epsilon}$ $\quad$ ($\Leftrightarrow$ finite class : $n \ge \dfrac{\log(\frac{|G|}{\delta})}{\epsilon}$)

[Exercise] let $S = \mathbb{R}$, $G = \{g_t, t \in \mathbb{R}\}$, $g_t(x) = \begin{cases} +1, & x \ge t \\ -1, & x < t \end{cases}$

Prove that G is PAC-learnable assume realizability



Vapnik - Chervonenkis

Question: Which classes G are agnostic PAC learnable ?

observation : let $(x_1, y_1), \cdots, (x_n, y_n)$ be the training data , and G is the concept class

consider $G_c = \{(g(x_1), \cdots, g(x_n)), g \in G\}$

note that $|G_c| \leq 2^n$

from exercise , we have g_r (圆形 classifier)

g_r $\|x_1\|, \cdots, \|x_n\|$

$i_1, \cdots, i_n$ is a permutation such that $\|x_{i_1}\| \leq \|x_{i_2}\| \leq \cdots \leq \|x_{i_n}\|$

$\Rightarrow \{(g_r(x_{i_1}), \cdots, g_r(x_{i_n})), g_r \in G\}$

can tell $y_{i_j} = \begin{cases} +1 \Rightarrow y_{ik} = +1 \quad \forall k \leq j \\ -1 \Rightarrow y_{ik} = -1 \quad \forall k \geq j \end{cases}$

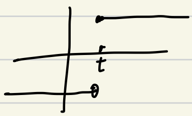
$\Rightarrow$ at most $(n+1)$ vectors

$\Rightarrow$ what makes it PAC-learnable

Let $(X_1, Y_1) \ldots (X_n, Y_n)$ is the treaming data
G is the concept class
$C = (X_1, \ldots X_n)$. The set $G_C = \{(g(X_1) \ldots g(X_n)), g \in G\}$

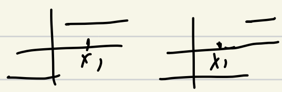
G_c : restriction of G onto C.
$|G_c| \leq 2^n$
If $|G_c| = 2^n$, we will say G shatters C
Remark : C can be an arbitary finite set

Ex $G = \{g_t, t \in \mathbb{R}\}$
$$g_t(x) = \begin{cases} +1, & x \geq t \\ -1, & x < t \end{cases}$$

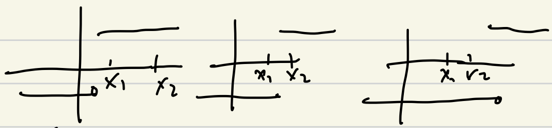


Shift t to get ± 1 for X_1

$C = \{X_1\}$ $\quad G_c = \{g_t(X_1), t \in \mathbb{R}\} = \{+1, -1\}$ shatters



$C = \{X_1, X_2\}$



$(+1, +1)$ $\quad (-1, +1)$ $\quad (-1, -1)$ $\quad \cancel{(+1, -1)}$

$G_c = \{(1,1), (-1,-1), (-1,+1)\}$ no shatter

Def $VC(G)$ – the Vapnik–Chervenkis dimension of G – is the largest d such that $\exists \{X_1, \ldots X_d\}$ that is shattered by G

Remark : $VC(G) = d \iff \begin{cases} \exists \{X_1, \ldots, X_d\} \text{ shattered by } G \\ \text{any set } \{X_1, \ldots X_{d+1}\} \text{ is not shattered by } G. \end{cases}$

<u>Remark</u>: we will prove the "Fundamental theorem of PAC learning"

G is agnostic PAC learnable $\iff VC(G) < \infty$

<u>Ex 2</u> S is $\boxed{\text{infinite}}$

$$G = \{ g_T ,\ T \subseteq S,\ |T| < \infty \}$$

$$g_T(x) = \begin{cases} 1, & x \in T \\ -1, & x \notin T \end{cases}$$

Then $VC(G) = \infty$

<u>Solution</u>: for any $d \geq 1$, we need to find $\{x_1, \dots, x_d\} = C$ s.t. $|G_C| = 2^d$

Let's take any $\{x_1 \dots x_d\}$, $G_C = \{(g_T(x_1) \dots g_T(x_d)),\ T \subseteq S,\ |T| < \infty\}$

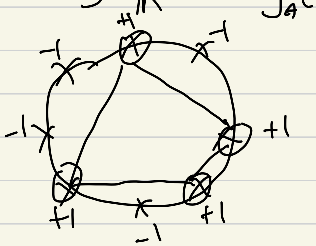
$W = \{+1, -1\}^d$, $J = \{j ; w_j = +1\}$

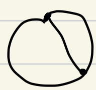
$$(\underset{1}{\underline{+1}}, -1, -1, -1, \underset{4}{+1}, \dots \underset{d}{\underline{+1}})$$

Take $T = \{x_j,\ j \in J\} \Rightarrow (g_T(x_1) \dots g_T(x_d)) = w$

<u>Ex3</u> $G = \{g_A, A \text{ is convex set}\}$

$S = \mathbb{R}^2$ $\quad g_A(x) = \begin{cases} 1, & x \in A \\ -1, & x \notin A \end{cases}$





convex

not convex.

$VC(G) = \infty$

<u>Ex4</u> G is $\boxed{\text{finite}}$ $\quad VC(G) < \infty$

$(x_1 \dots x_d) = C$

$G_C = \{ g(x_1), \dots g(x_d) \} , g \in G \}$

$|G_C| \leq |G|$

If $VC(G) = d \Rightarrow |G_C| = 2^d \quad |G| \geq 2^d \Rightarrow d \leq \lfloor \log_2 |G| \rfloor + 1$

to_integer()

Lemma 1: Let G be a concept class of infinite VC dimension, then G is not PAC-learnable

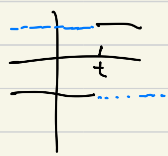
Proof: G is P.A.C-learnable if $\forall \varepsilon, \delta > 0$,
$\exists A$ - an algorithm and $m = m(\varepsilon, \delta)$ m # of traing data must form.
then if $(X_1, Y_1), \ldots (X_n, Y_n)$ is the training data and $n \geq m(\varepsilon, \delta)$ then $P_r(L(\hat{g}) \geq \varepsilon) \leq \delta$, $\hat{g} = A((X_1, Y_1), \ldots (X_n, Y_n))$

$VC(G) = \infty \Rightarrow \forall m \geq 1$, $\exists \{X_1, \ldots, X_m\} \subseteq S$, such that G shatters $\{X_1 \ldots X_m\}$
$\Rightarrow$ for any $z \in \{+1, -1\}^m$, $\exists g \in G$ st $z = (g(X_1), \ldots, g(X_m))$
$G_c = \{+1, -1\}^m$

Take $\varepsilon = \frac{1}{8}$, $\delta = \frac{1}{10}$
Take N arbitarily large. Find $C = \{X_1, \ldots X_N\}$ shattered by G
By no free lunch theorem for any A, $\max_{g^*} Pr(L(\hat{g}) \geq \frac{1}{8}) \geq \frac{1}{8}$

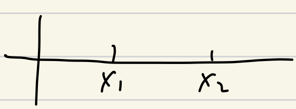
Proved by contradiction.

EX 1: $G = \{g_t^+, g_t^-, t \in \mathbb{R}\}$
$S = \mathbb{R}$, $g_t^+ = \begin{cases} 1 & x \geq t \\ -1 & x < t \end{cases}$ $\quad g_t^- = \begin{cases} 1 & x < t \\ -1 & x \geq t \end{cases}$



Then $VC(G) = 2$ (Prove this)
(a) Find two points $\{X_1, X_2\}$ that are shattered.
(b) Show no set of 3 points are shattered.

(a)

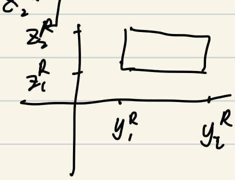


$(+1, +1) \to g_t^+ \; t < X_1$
$(+1, -1) \to g_t^-, t \in (X_1, X_2)$
$(-1, +1) \to g_t^+, t \in (X_1, X_2)$
$(-1, -1) \to g_t^-, t > X_2$

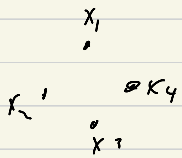
(b) $(-1, +1, -1)$
$(+1, -1, +1)$ nope

Ex 2: $S = \mathbb{R}^2$, $G = \{ g_R,\ R \text{ is axis-aligned Rectangle} \}$

$$g_R = \begin{cases} +1, & x_1 \in [y_1^R, y_2^R],\ x_2 \in [z_1^R, z_2^R] \\ -1, & \text{else} \end{cases}$$



(a)



just need one example

(b) no 5 points are shattered

WLOG, assume that x_1 has the largest coordinate.

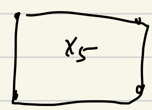
x_2 has the smallest x coordinate

x_3 has the largest y coordinate

x_4 has the smallest y coordinate

x_5 will be inside the rectangle with -1

which is impossible.



Theorem (R. Dudley) {not in textbook}

Let L be a finite-dimensional space of function $f: S \to \mathbb{R}$. Consider

$C_f = \{ \{x : f(x) > 0\}, f \in L \}$ and $\overline{C_f} = \{ \{x : f(x) \geq 0\}, f \in L \}$.

Let $G = \{ I_{C_f} - I_{C_f^c},\ f \in F \}$ $\quad \overline{G} = \{ I_{\overline{C_f}} - I_{\overline{C}_f^c},\ f \in F \}$

Then $VC(G) = VC(\overline{G}) = \dim(L)$

Finite dim: Ex $\quad L = \{ \langle a, x \rangle + b,\ a \in \mathbb{R}^d,\ b \in \mathbb{R} \}$

$a \in \mathbb{R}^d \Rightarrow a = a_1 e_1 + \dots + a_d e_d$

$$e_j = (0, \dots 0, \underset{\substack{\uparrow \\ j}}{1}, 0, \dots, 0)$$

$\langle a, x \rangle + b = a_1 \langle e_1, x \rangle + \dots + a_d \langle e_d, x \rangle + b$

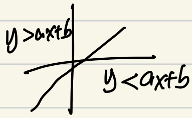
$\dim(L) = d + \underset{\substack{\downarrow \\ \text{constant}}}{1}$

Example Polynomials of degree at most d.

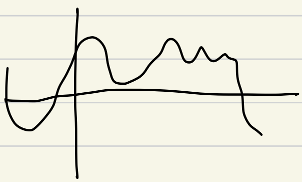
$f(x) = a_1 x^d + a_2 x^{d-1} + \dots + a_d x + a_{d+1}$ $\quad$ $\dim(L) = d+1$

$y = ax+b$ $\quad$ $\{(x,y) : y-ax-b > 0\}$

$y-ax-b = 0$



$a_1 x + a_2 y + b = 0$ $\quad$ $\dim = 3 = VC(G) = VC(\bar{G})$



rectangle formation:
intersection of
4 half subspaces
$\dim \times 4 = 12$ (upper bound?)
$\overset{!!}{3}$

Proof for (R. Dudley): $VC(G) \leq \dim(L)$

let $\dim(L) = d$, We need to show no set of $d+1$ points is shattered by G.

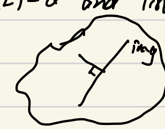
Take $\{x_1, \dots, x_{d+1}\}$, Consider $T(f) = (f(x_1) \dots f(x_{d+1}))$ $\quad d+1$

$T(\alpha f + \beta g) = \alpha T(f) + \beta T(g)$

Note that $\dim(\text{Image}(T)) \leq d$ because $\dim(L) = d$ and linear maps don't increase dimension.

$\exists\, w \in \mathbb{R}^{d+1}$ such that $w \perp \text{Image}(T)$

and $w \neq 0$. Hence, if $w = (w_1, w_2, \dots w_{d+1})$



$\Rightarrow \exists j$ st $w_j < 0$ (if $w_j > 0$, take $-w$ instead)

$A_- = \{1 \leq j \leq n : w_j < 0\}$ $\quad$ $A_+ = \{1 \leq j \leq n, w_j \geq 0\}$

Assume that $\{x_1, \dots x_{d+1}\}$ is shattered by G

Since every $g \in G$ is $\text{sign}(f)$ for some $f \in F$.

$\exists f \in F$ st $f(x_j) > 0,\ j \in A_-$, $f(x_j) < 0,\ j \in A_+$

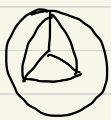
On one hand, since $w \perp \mathrm{Im}(T)$

$$\sum_{j=1}^{d+1} w_j f(x_j) = 0$$

On the other hand, $$\sum_{j=1}^{d+1} w_j f(x_j) = \underbrace{\sum_{j \in A_-} w_j f(x_j)}_{<0} + \underbrace{\sum_{j \in A_+} w_j f(x_j)}_{\leq 0} < 0$$

Contradiction $\Rightarrow \{x_1, \dots, x_{d+1}\}$ cannot be shattered by G.

Example: $B_d(x, r) = \{ y \in \mathbb{R}^d, \ \|y - x\|_2 \leq r \}$



ball

$G = \{ g : \mathbb{R}^d \to \{+1, -1\} \}$, where

$g = g_{x,r}$, and $g_{x,r}(y) = \begin{cases} 1, & y \in B_d(x,r) \\ -1, & \text{else} \end{cases}$

Then $VC(G) \leq d+2$

Express definition of a ball as $f(y) \geq 0$ for $f \in L$, where $\dim(L) = d+2$

norm $= \sqrt{\sum_{j=1}^{d} (y_j - x_j)^2} \leq r \iff \sum_{j=1}^{d} (y_j - x_j)^2 \leq r^2 \iff \sum_{j=1}^{d} (y_j^2 - 2x_j y_j + x_j^2) \leq r^2$

$$\iff \underbrace{-\left(\sum_{j=1}^{d} y_j^2 - 2\sum_{j=1}^{d} x_j y_j + \sum_{j=1}^{d} x_j^2 - r^2 \right)}_{f(y_1, \dots, y_d)} \geq 0$$

$f_1(y_1, \dots, y_d) = \sum_{j=1}^{d} y_j^2 \qquad f_2(y_1, \dots, y_d) = y_1 \qquad \dots \ f_{d+1}(y_1 \dots y_d) = y_d$

$f_{d+2}(y_1, \dots, y_d) = 1$

$f(y_1, \dots, y_d) = -1 \cdot f_1 + 2x_1 f_2 + 2x_2 f_3 + \dots + 2x_d f_{d+1} + (-\sum_{j=1}^{d} x_j^2 + r^2) f_{d+2}$

$\Rightarrow f(y_1, \dots, y_d) \in L$ and $\dim(L) = d+2$

$C = \{x_1, \dots, x_k\}$, G – concept class, $G_C = \{ (g(x_1), \dots, g(x_k)) : g \in G \}$

Assume that $VC(G) = d$. Then $\exists \{x_1, \dots, x_d\} = C_d$ st $|G_{C_d}| = 2^d$

What if $k > d$? what can we say about $|G_C|$? (beyond the fact that $|G_C| \leq 2^k$)

Def (The growth function)

$\tau_G(k) = \max |G_C| \qquad C = \{x_1, \dots, x_k\}$

Lemma (Shelah – Sauer – Perles – Vapnik – Chervonenkis)

Let G be such that $VC(G) = d$.

Then $\tau_G(k) \leq \sum_{j=0}^{d} \binom{k}{j} \qquad \binom{k}{j} = \frac{k!}{j!(k-j)!}$

If $k \le d$, $\tau_G(k) \le \sum_{j=0}^{k} \binom{k}{j} = 2^k$

For $k > d$, $\sum_{j=0}^{d} \binom{k}{j} \le \left(\frac{ek}{d}\right)^d$

Proof: $\binom{k}{j} = \frac{k!}{j!(k-1)!} = \frac{k(k-1)\cdots(k-j+1)}{j!}$

$$= \frac{k}{d}\,\frac{(k-1)}{d}\cdots\frac{(k-j+1)}{d}\cdot\frac{d^j}{j!} < \left(\frac{k}{d}\right)^j \frac{d^j}{j!}$$

$1 < \frac{k}{d}$

$$\le \left(\frac{k}{d}\right)^d \frac{d^j}{j!}$$

$$\sum_{j=0}^{d} \binom{k}{j} \le \sum_{j=0}^{d} \left(\frac{k}{d}\right)^d \frac{d^j}{j!}$$

$$\le \left(\frac{k}{d}\right)^d \underbrace{\sum_{j=0}^{\infty} \frac{d^j}{j!}}_{e^d} = \left(\frac{ek}{d}\right)^d$$

Exercise: $\sum_{j=0}^{d} \binom{k}{j} \ge \left(\frac{k}{d}\right)^d$

Recap

G - class of binary classifiers

$\tau_a(m) = \max_{C=\{x_1 \ldots x_m\} \subset S} |G_C|$

Lemma If $VC(G) = d < \infty$, the $\tau_a(m) \le \left(\frac{me}{d}\right)^d$ for all $m > d$

Example: Let G_1, G_2 be two classes of binary classifiers, $\boxed{\text{Prove intersection is finite}}$

$VC(G_1) = d_1 < \infty$, $VC(G_2) = d_2 < \infty$

Let $C_{G_i} = \{\{x : g(x) = +1\},\ g \in G_i\}$, $i = 1, 2$

$$C_{G_1} \cap C_{G_2} = \{C_1 \cap C_2 : C_1 \in C_{G_1},\ C_2 \in C_{G_2}\}$$

Let G be the set of all classifiers

$$g(x) = \begin{cases} 1, & x \in C \text{ for } C \in C_{G_1} \cap C_{G_2} \\ -1, & \text{else} \end{cases}$$

intersection

Proof: $VC(G) < \infty$

<u>idea:</u> If we can show that $\tau_G(m) = O(m^v)$ for some $v < \infty$, then $VC(G) < \infty$

Fix some $\{x_1 \dots x_m\} = M$

Consider the set $\left| M \cap \{\{x : g(x) = +1\}, g \in C_1\}\right| \leq \tau_{G_1}(m)$

illustrate: $g_t(x) = \begin{cases} +1, & x \geq t \\ -1, & x < t \end{cases}$

$x_1 \quad x_2 \quad x_3$

$M = (x_1, x_2, x_3)$

$M \cap \{\{x : g(x) = +1\} g_t \in G\} = \{(x_1, x_2, x_3), (x_2, x_3), x_3, \phi\}$

$G_m = \{(+1,+1,+1), (-1,+1,+1), (-1,-1,+1), (-1,-1,-1)\}$

$M \cap \{\{x : g_1(x) = +1\}, g_1 \in G\} = M_{G_1}$

$M \cap \{\{x : g_2(x) = +1\}, g_2 \in G_2\} = M_{G_2}$

$C_1 \in M_{G_1}$

$|C_1 \cap M_{G_2}| \leq |M_{G_2}| = \tau_{G_2}(m)$

$\Rightarrow \underset{G_1}{|M_{C_1}} \cap \underset{G_2}{M_{C_2}|} \leq \underset{G_1}{|M_{C_1}|} \cdot \underset{G_2}{|M_{C_2}|} \leq \tau_{G_1}(m) \cdot \tau_{G_2}(m)$

10k

$$\tau_G(m) \leq \tau_{G_1}(m) \cdot \tau_{G_2}(m)$$
$$\leq \left(\frac{me}{d_1}\right)^{d_1} \cdot \left(\frac{me}{d_2}\right)^{d_2} = O(m^{d_1+d_2})$$
$$\Rightarrow VG(G) < \infty$$

<u>Example:</u> $G_1 = \{g_t^+, t \in \mathbb{R}\}$ $\quad$ + $\quad$ − $\quad$ $VC(G_1) = 1$

$G_2 = \{g_t^-, t \in \mathbb{R}\}$ $\quad$ + $\quad$ − $\quad$ $VC(G_2) = 1$

$G = \{g_{[a,b]}, a, b \in \mathbb{R}\}$

$g_{[a,b]}(x) = \begin{cases} +1, & x \in [a,b] \\ -1, & \text{else} \end{cases}$

$\tau_G(m) \leq \tau_{G_1}(m)\, \tau_{G_2}(m)$

$\tau_{G_1}(m) = m+1, \quad \tau_{G_2}(m) = m+1$

$\tau_G : \binom{m+1}{2} = \frac{m(m+1)}{2} \leq (m+1)^2$

−1	+1	+1
−1	+1	+1
−1	−1	+1
−1	−1	−1
+1	+1	−1
+1	−1	−1

$(X_1, Y_1) \ldots (X_n, Y_n)$ is ε-representative if

$$\max_{g\in G}\left|\frac{1}{n}\sum_{1}^{n} I\{Y_j \neq g(X_j)\} - \underset{\substack{\| \\ Pr(Y\neq g(X))}}{L(g)}\right| \leq \varepsilon$$

Theorem: Let G be a concept class (a class of binary classifiers) and let $\tau_G(n)$ be its growth function.

Then $\max_{g\in G}\left|\frac{1}{n}\sum_{1}^{n} I\{Y_j \neq g(X_j)\} - L(g)\right| \leq \dfrac{\sqrt{2\log(2\tau_G(n))}}{\delta\sqrt{n}}$ $\leq \frac{\varepsilon}{2}$

with probability at least $1-\delta$.

It's sufficient to prove that $\underbrace{\mathbb{E}\max_{g\in G}\left|\frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - L(g)\right|}_{Z} \leq \dfrac{\sqrt{2\log(2\tau_G(n))}}{\sqrt{n}}$

It's because

$Pr(Z > t) \leq \frac{\mathbb{E}Z}{t}$ - Markov's inequality.

$Pr(Z > \underbrace{\tfrac{\mathbb{E}Z}{\delta}}_{t}) \leq \frac{\mathbb{E}Z}{\mathbb{E}Z}\delta = \delta$

For (agnostic) PAC-learnability, we need

(a) An algorithm $\mathcal{A}$

(b) $n(\varepsilon, \delta)$

s.t. given $(X_1, Y_1) \ldots (X_n, Y_n)$ with $n \geq n(\varepsilon, \delta)$, $\mathcal{A}$ outputs $\hat{g}_n$ st

$Pr(L(\hat{g}_n) > \varepsilon) \leq \delta$

In our case, if $\mathcal{A}$ is ERM, we know that $\frac{\varepsilon}{2}$-representative sample yields a classifier $L(\hat{g}_n) \leq \varepsilon$

Doing some algebra, we get $n \geq K\frac{V}{\delta^2\varepsilon^2}\log\left(\frac{V}{\delta^2\varepsilon^2}\right)$, K = constant

Symmetrization inequality:

let $\sigma_1, \dots, \sigma_n$ be random signs. ie iid random variables st $Pr(\sigma=1) = Pr(\sigma=-1) = \frac{1}{2}$

$$\mathbb{E} \max_{g \in G} \left| \frac{1}{n} \sum_{1}^{n} I\{Y_j \neq g(X_j)\} - L(g) \right| \leq 2 \mathbb{E} \max_{g \in G} \left| \frac{1}{n} \sum_{1}^{n} \sigma_j I\{Y_j \neq g(X_j)\} \right|$$

$$\downarrow$$

inner product $\langle (\sigma_1, \dots \sigma_n), (I\{Y_1 \neq g(X_1)\} \dots I\{Y_n \neq g(X_n)\}) \rangle$

like random noise

Theorem G - a class of binary classifiers

$$\max_{g \in G} |L_n(g) - L(g)| \leq \frac{4}{\delta} \sqrt{\frac{\log(2 T_a(n))}{n}}$$

with probability $\geq 1-\delta$ over the choice of the sample $(X_1, Y_1) \dots (X_n, Y_n)$

In other words, if

sample size n as a function of ε, δ

$$\underset{n}{\overset{}{n(\varepsilon, \delta)}} \geq 100 \frac{V(G)}{\delta^2 \varepsilon^2} \log\left(\frac{V}{\delta^2 \varepsilon^2}\right)$$

$\Rightarrow \hat{g}_n$ produced by ERM when given a sample of size $n(\varepsilon, \delta)$ satisfies

$$Pr\left(L(\hat{g}_n) > \min_{g \in G} L(g) + \varepsilon\right) \leq \delta$$ $\rightarrow$ defn of PAC-learnability

It suffices to show that

$$\mathbb{E} \max_{g \in G} |L_n(g) - L(g)| \leq \frac{4}{\sqrt{n}} \sqrt{\log(2 T_a(n))}$$

Symmetrization inequality:

$$\mathbb{E} \max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - L(g) \right| \leq 2 \mathbb{E} \max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} \sigma_j I\{Y_j \neq g(X_j)\} \right|$$

$\sigma_1, \dots \sigma_n$ iid random signs independent from $(X_1, Y_1), \dots (X_n, Y_n)$, i.e. $Pr(\sigma_i = 1) = P_r(\sigma_i = -1) = \frac{1}{2}$

Note that $\mathbb{E} \max_{g \in G} \left| \frac{1}{n} \sum_{1}^{n} \sigma_j I\{Y_j \neq g(X_j)\} \right|$

$$\mathbb{E}_{(X_j,Y_j)_{j=1}^n} \underbrace{\mathbb{E}_{\sigma_1 \dots \sigma_n} \max_{g \in G} \left| \frac{1}{n} \sum_1^n \sigma_j \, \overbrace{I\{Y_j \neq g(X_j)\}}^{t_j \in \{0,1\}} \right|}_{\text{focus on this.}}$$

$t = (t_1, \dots, t_n)$

$$\mathbb{E}_{\sigma_1 \dots \sigma_n} \max_{t \in T} \left| \frac{1}{n} \sum_1^n \sigma_j t_j \right|$$

Remark: $\left| \{ (I\{Y_1 \neq g(X_1)\}, \dots, I\{Y_n \neq g(X_n)\}), g \in G \} \right|$

$= |G_c|, \quad C = \{X_1, \dots, X_n\}$

$\quad \; \| $

$\{(g(X_1), \dots, g(X_n)), g \in G\}$

$\quad -1 \qquad 1$

The number of such vectors is at most $T_G(n)$!

Lemma: Let $t^{(1)} \dots t^{(k)} \in \mathbb{R}^n$

Then $\mathbb{E} \max_{j=1 \dots k} \left| \frac{1}{n} \sum_1^n \sigma_i t_i^{(j)} \right| \leq 2 \max_{j=1 \dots k} \frac{\|t^{(j)}\|_2}{\sqrt{n}} \sqrt{\frac{\log(2k)}{n}}$

Exercise $\mathbb{E} \left| \frac{1}{n} \sum_1^n \sigma_j t_j \right| \leq \frac{1}{\sqrt{n}} \frac{\|t\|_2}{\sqrt{n}}$

Proof: let $f(\lambda) = \mathbb{E} e^{\lambda \sigma_1} \leq e^{\frac{\lambda^2}{2}}$

Indeed, $\mathbb{E} e^{\lambda \sigma_i} = e^{\lambda \cdot 1} \frac{1}{2} + e^{\lambda(-1)} \frac{1}{2}$

$= \frac{1}{2}(e^{\lambda} + e^{-\lambda}) = \frac{1}{2}\left(1 + \cancel{\lambda} + \frac{\lambda^2}{2} + \dots + \frac{\lambda^k}{k!} + \dots + 1 - \cancel{\lambda} + \frac{\lambda^2}{2} + \dots (-1)^k \frac{\lambda^k}{k!} + \dots\right)$

$= \frac{1}{2} \cdot 2 \sum_{k \geq 0} \frac{\lambda^{2k}}{(2k)!} = 1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \dots$

$\lambda^{2k} = (\lambda^2)^k \qquad (2k)! = \underbrace{1 \cdot 2 \cdots k}_{k!} \underset{\geq 2}{(k+1)} \underset{\geq 2}{(k+2)} \cdots \underset{\geq 2}{(2k)} \geq 2^k \cdot k!$

$$\sum_{k \geq 0} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k \geq 0} \frac{(\lambda^2)^k}{k! \, 2^k} = \sum_{k \geq 0} \frac{(\frac{\lambda^2}{2})^k}{k!} = e^{\frac{\lambda^2}{2}} \qquad \text{for } k \geq 1$$

MGF (Moment Generating Function) of $\frac{1}{n}\sum_{1}^{n}\sigma_j t_j$:

$$\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{1}^{n}\sigma_j t_j)} = \mathbb{E}\left(e^{\frac{\lambda}{n}\sigma_1 t_1}\cdots e^{\frac{\lambda}{n}\sigma_n t_n}\right)$$

$$= \mathbb{E}e^{\frac{\lambda}{n}\sigma_1 t_1} \times \cdots \times \mathbb{E}e^{\frac{\lambda}{n}\sigma_n t_n}$$

$$\leq e^{\frac{\lambda^2 t_1^2}{2n^2}}\cdots e^{\frac{\lambda^2 t_n^2}{2n^2}} = e^{\frac{\lambda^2}{2n}\cdot\frac{\sum_1^n t_j^2}{n}} = e^{\frac{\lambda^2}{2n}\frac{\|t\|^2}{n}}$$

Next, $x \mapsto e^{\lambda x}$ is convex, i.e. $e^{\lambda(\sum_{j=1}^{k}\alpha_j x_j)} \leq \alpha_1 e^{\lambda x_1} + \cdots + \alpha_n e^{\lambda x_n}$

$\alpha_1 \dots \alpha_k \geq 0$

$\sum_1^k \alpha_j = 1$

In other words, $\underbrace{e^{\lambda\mathbb{E}Z} \leq \mathbb{E}e^{\lambda Z}}_{\text{Jensen's inequality}}$ $\quad Pr(X = x_j) = \alpha_j$

$$Z = \max_{j=1\dots k}\left|\frac{1}{n}\sum_{1}^{n}\sigma_i t_i^{(j)}\right| \qquad |a| = \max(a, -a)$$

$$e^{\lambda\mathbb{E}Z} \leq \mathbb{E}e^{\lambda Z} = \mathbb{E}\max_{j=1\dots k}\underbrace{\left(\underset{\geq 0}{e^{\frac{\lambda}{n}\sum_1^n\sigma_i t_i^{(j)}}}, \underset{\geq 0}{e^{-\frac{\lambda}{n}\sum_1^n\sigma_i t_i^{(j)}}}\right)}_{2k \text{ random variables}}$$

$$\mathbb{E}\sum_{j=1}^{k}\left(e^{\frac{\lambda}{n}\sum_1^n\sigma_i t_i^{(j)}} + e^{-\frac{\lambda}{n}\sum_1^n\sigma_i t_i^{(j)}}\right) \leq 2\sum_{j=1}^{k}\exp\left(\frac{\lambda^2}{n^2}\frac{\|t^{(j)}\|_2^2}{2}\right)$$

$$\leq 2k\exp\left(\frac{\lambda^2}{n^2}\max_{j=1\dots k}\frac{\|t^{(j)}\|_2^2}{2}\right)$$

$$e^{\lambda\mathbb{E}Z} \leq 2ke^{\frac{\lambda^2}{n^2}\max_j\frac{\|t^{(j)}\|_2^2}{2}}$$

Take log : $\cancel{\lambda}\mathbb{E}Z \leq \frac{\log(2k)}{\lambda} + \frac{\lambda}{2n^2}\max_{1\dots k}\|t^{(j)}\|_2^2$

True for any $\lambda > 0$

$$h(\lambda) = \frac{\log(2k)}{\lambda} + \frac{\lambda}{2n^2}\max_{j=1\dots k}\|t^{(j)}\|_2^2$$

$$h'(\lambda) = 0 \iff \lambda_*^2 = \frac{\log(2k)}{2\frac{1}{n}\max\frac{\|t^{(j)}\|_2^2}{n}}$$

$$h(\lambda_*) = \sqrt{2}\sqrt{\frac{\log(2k)}{2}} \max_{j=1,\dots,n} \frac{\|t^{(j)}\|_2}{\sqrt{n}}$$

$$\mathbb{E}_{\sigma_1 \dots \sigma_n} \max_{g \in G} \left| \frac{1}{n} \sum_{1}^{n} \sigma_j \cdot I\{Y_j \neq g(X_j)\} \right| \leq \sqrt{2}\sqrt{\frac{\log(2T_G(n))}{n}} \cdot 1$$

$t^{(j)} = (I\{Y_1 \neq g(X_1)\}, \dots I\{Y_n \neq g(X_n)\})$
$\|t\|_2^{(j)} = \sqrt{n}$

Symmetrization inequality

Let G be a class of binary classifiers, and $\sigma_1, \dots \sigma_n$ are iid Random signs, i.e. $Pr(\sigma_1 = 1) = Pr(\sigma = -1) = \frac{1}{2}$

Then $\mathbb{E}\max_{g \in G} |L_n(g) - L(g)| = \mathbb{E}\max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - Pr(Y \neq g(X)) \right|$

$$\leq 2\,\mathbb{E}\max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} \sigma_j I\{Y_j \neq g(X_j)\} \right|$$

Proof: Let $(X_1', Y_1'), \dots (X_n', \dots Y_n')$ - an independent copy of $(X_1, Y_1) \dots (X_n, Y_n)$

Note that $L(g) = \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^{n} I\{Y_j' \neq g(X_j')\}\right)$

Therefore, $\mathbb{E}\max_{g \in G} |L_n(g) - L(g)| = \mathbb{E}\max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(Y_j)\} - \mathbb{E}_{(X',Y')} \frac{1}{n} \sum_{j=1}^{n} I\{Y_j' \neq g(X')\} \right|$

$$\leq \mathbb{E}_{(X,Y)} \max_{g \in G} \mathbb{E}_{(X',Y')} \left| \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - I\{Y_j' \neq g(X_j')\} \right|$$

$|\mathbb{E} Z| \leq \mathbb{E}|Z|$

$|\sum_i a_i| \leq \sum_i |a_i|$

$\max \mathbb{E} Z_i \leq \mathbb{E}\max_i Z_i$

$$\leq \mathbb{E}\max_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} - I\{Y_j' \neq g(X_j')\} \right|$$

Note that we can "switch" (X_j, Y_j) with (X_j', Y_j') for any j without changing the expectation.

Equivalently, for any fixed $\varepsilon_1 \dots \varepsilon_n \in \{+1, -1\}^n$

$$\leq \mathbb{E} \max_{g\in G} \left| \frac{1}{n}\sum_{j=1}^{n} q_j \left(I\{Y_j \neq g(X_j)\} - I\{Y_j' \neq g(X_j')\} \right) \right| = \frac{1}{2^n} \sum_{(q_1 \ldots q_n)\in\{-1,+1\}^n} \mathbb{E} \max_{g\in G} \left| \frac{1}{n}\sum_{j=1}^{n} q_j \left(I\{Y_j \neq g(X_j)\} - I\{Y_j' \neq g(X_j')\} \right) \right|$$

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_n} \mathbb{E} \max_{g\in G} \left| \frac{1}{n}\sum_{j=1}^{n} \sigma_j \left(I\{Y_j \neq g(X_j)\} - I\{Y_j' \neq g(X_j')\} \right) \right|$$

$|a-b| \leq |a| + |b|$

$$\leq 2\, \mathbb{E} \max_{g\in G} \left| \frac{1}{n}\sum_{j=1}^{n} \sigma_j I\{Y_j \neq g(X_j)\} \right|$$

END

The Fundamental Theorem of PAC Learning.

Let G be a class of binary classifiers. Then the following conditions are equivalent:

(a) G is agnostic PAC learnable via the ERM algorithm.

(b) G is PAC learnable via the ERM algorithm.

(c) G has the "uniform convergence" property: $\forall \varepsilon, \delta > 0, \exists n(\varepsilon,\delta)$ s.t. $\forall n \geq n(\varepsilon,\delta)$
$\max_{g\in G} |L_n(g) - L(g)| \leq \varepsilon$ with probability at least $1-\delta$.

(d) G has finite VC dimension.

Proof: (d) $\Rightarrow$ (c)

Moreover, we have shown that $n(\varepsilon,\delta) \leq \text{Constant}\ \frac{VC(G)}{\delta^2\varepsilon^2} \log\left(\frac{e\,VC(G)}{\delta^2\varepsilon^2}\right)$

(c) $\Rightarrow$ (a) (a) $\Rightarrow$ (b) (b) $\Rightarrow$ (d)

Remark

$$C_1 \frac{V + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \leq n(\varepsilon,\delta) \leq C_2 \frac{V + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$$

End of theory.

Practical stuff

Learning beyond binary classification.

- What if there are 3 or more classes that the objects of interest should be classified into?
- What if the "label" Y takes values in $\mathbb{R}$?

- Our theory remains valid modulo minor changes.

"One vs All"   Cat vs Dog or Rabbit
yes → Cat   no → Dog or Rabbit

"1 vs 1": if we have k possible labels, $\{1, \dots, k\}$, consider $\binom{k}{2}$ binary classification problems "X is in class i or X is in class j". Pick the label that gets most "+1" votes.

- We will talk about general "prediction" problems: predict the "response" Y based on the "predictor" X. Prediction is performed via some function $g \in G$.

Example Multi-label classification
X is a test paper, $Y \in \{A, B, C, D, F\}$

Example general regression problem
X = hours spent on social media / week, Y = GPA $\in [1,4]$

- Need to generalization the notion of the loss function, denoted $\ell(y, g(x))$
  e.g. in binary classification, $\ell(y, g(x)) = I\{y \neq g(x)\}$
  In multi-label classification, it can be $\ell(y, g(x)) = I\{y \neq g(x)\}$
  We can also choose

$$\ell(y, g(x)) = \begin{cases} 0, & y = g(x) \\ 1, & y \neq g(x) \text{ and } y = 0 \\ 100, & y \neq g(x) \text{ and } y = 1 \end{cases}$$

- The goal remains as before; minimize $\mathbb{E}\, l(Y, g(X))$ over $g \in G$.

- *Example* Regression problem: $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$

$$l(y, g(x)) = (y - g(x))^2 \quad \leftarrow \text{why squared} \quad (MLE)$$

$$G = \{ \langle w, x \rangle + b,\ w \in \mathbb{R}^d,\ b \in \mathbb{R} \}$$

- *Ex* Assume that $X \in \mathbb{R}$, assume that $Y_j = \alpha X_j + \beta + \varepsilon_j$, $\alpha, \beta \in \mathbb{R}$

  and ε_j is (a) $N(0, \sigma^2)$

  (b) Laplace distribution with density $p(x) = \frac{1}{2} e^{-|x|}$, $x \in \mathbb{R}$

  $\varepsilon_1, \dots, \varepsilon_n$ are independent. Show that the MLE of α, β minimizes

  (a) $\frac{1}{n} \sum_{j=1}^{n} (Y_j - a x_j - b)^2$ over $a, b \in \mathbb{R}$

  (b) $\frac{1}{n} \sum_{1}^{n} | Y_j - a x_j - b |$

Question of practical importance:

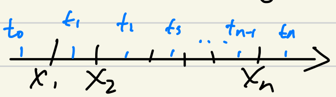
Can we implement ERM methods that have strong theoretical guarantees?

Example $S = \mathbb{R}$, $T = \{+1, -1\}$, $(X, Y) \in S \times T$

$$G = \{ g_t^+, g_t^- \}, \quad g_t^+(x) = \begin{cases} +1, & x \geq t \\ -1, & x < t \end{cases}, \quad g_t^-(x) = \begin{cases} -1, & x \geq t \\ +1, & x < t \end{cases}$$

(a) *Realizable scenario*

$\hat{g}_n$ minimizes $\frac{1}{n} \sum_{j=1}^{n} I\{ Y_j \neq g(X_j) \}$ over $g \in G$



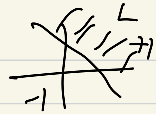
$X_{(j)}$ is the jth smallest among $X_1, \dots, X_n$

$O(n \log n)$ to sort. $O(\log n)$ to find t_*

Compare $L(g^+_{t_0}), \dots, L(g^+_{t_n})$

$L(g^-_{t_0}), \dots, L(g^-_{t_n})$

In agnostic learning framework, we only need to compare the empirical risks of at least $2(n+1)$ classifiers.



What about linear separators in dimension 2?
Specifically, let $S = \mathbb{R}^2$, $T = \{+1, -1\}$
$G = \{g_L,\ L \to$ a half-plane$\}$ $\quad g_L(x) = \begin{cases} 1, & x \in L \\ -1, & x \notin L \end{cases}$

Let's take a look at the more general problem:
$S = \mathbb{R}^d$, $T = \{+1, -1\}$
The <u>hyperplane</u> is a set of points $\{x \in \mathbb{R}^d : \langle w, x\rangle + b = 0,\ w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$

The half-spaces are given by
$L = \{x \in \mathbb{R}^d : \langle w, x\rangle + b \geq 0$

$\tilde{x} = (x, 1)$, $\tilde{w} = (w, b)$, $\langle w, x\rangle + b = \langle \tilde{w}, \tilde{x}\rangle$

<u>Realizability</u> there exist w_* s.t. $y_j = \operatorname{sign}(\langle w_*, x_j\rangle)$

$\Leftrightarrow y_j \langle w_*, x_j\rangle > 0$

Let $\gamma = \min_{1 \ldots n} y_j \langle w_*, x_j\rangle \Rightarrow y_j \langle \frac{w_*}{\gamma}, x_j\rangle \geq 1 \ \forall j = 1 \ldots n$

Denote $\tilde{w} = \frac{w_*}{\gamma} \Rightarrow y_j \langle \tilde{w}, x_j\rangle \geq 1$ for all j.

Can we find such $\tilde{w}$?

Perceptron Algorithm (Frank Rosenblatt)
Given $(x_1, y_1), \ldots, (x_n, y_n)$, let $w_0 = \underbrace{(0, \ldots, 0)}_{d}$
for $t = 1, 2, \ldots$
  if $\exists\, 1 \leq j \leq n$ $\ y_j \langle w^{(t)}, x_j\rangle \leq 0$
  then $w^{(t+1)} = w^{(t)} + y_j x_j$ else return $w^{(t)}$

~~Perceptron as the pseudo-gradient descent method~~

$\exists w_* \quad \langle w_*, x_j \rangle > 0 \iff y_j = +1$

$\iff y_j \langle w_*, x_j \rangle > 0, \quad 1 \le j \le n$

$\gamma = \min_j y_j \langle w_*, x_j \rangle \Rightarrow \min_j y_j \langle \frac{w_*}{\gamma}, x_j \rangle = 1$

Goal: find a vector w s.t. $y_j \langle w, x_j \rangle \ge 1$ for all j

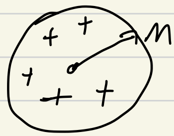
Perceptron Algorithm (Frank Rosenblatt)

Given $(x_1, y_1), \ldots, (x_n, y_n)$, let $w_0 = \underbrace{(0, \ldots, 0)}_{d}$

for $t = 1, 2, \ldots$

if $\exists\, 1 \le j \le n \quad y_j \langle w^{(t)}, x_j \rangle \le 0$

then $w^{(t+1)} = w^{(t)} + y_j x_j$ else return $w^{(t)}$



Th Assume that $\max_j \|x_j\| \le M$

Then the Perceptron algorithm stops after at most $\left(\frac{M}{\gamma}\right)^2$ iterations

Pf: Let $\tilde{w} = \frac{w_*}{\gamma}$ be s.t. $y_j \langle \tilde{w}, x_j \rangle \ge 1$

Consider $\langle w_{t+1}, \tilde{w} \rangle = \langle w_t + y_j x_j, \tilde{w} \rangle = \langle w_t, \tilde{w} \rangle + \underbrace{y_j \langle x_j, \tilde{w} \rangle}_{\ge 1} \ge \langle w_t, \tilde{w} \rangle + 1$

$\Rightarrow \langle w_{t+1}, \tilde{w} \rangle \ge t + 1$

At the same time, $\|w_{t+1}\|^2 = \|w_t + y_j x_j\|^2 = \langle w_t + y_j x_j, w_t + y_j x_j \rangle$

$= \|w_t\|^2 + \|x_j\|^2 + \underbrace{2 y_j \langle w_t, x_j \rangle}_{\le 0} \le \|w_t\|^2 + M^2 \le (t+1) M^2$

Combine 2 inequalities, $(t+1) \le \langle w_{t+1}, \tilde{w} \rangle \le \|w_{t+1}\| \cdot \underbrace{\|\tilde{w}\|}_{\frac{1}{\gamma}} \le M\sqrt{t+1} \cdot \frac{1}{\gamma}$

$\Rightarrow t+1 \le \frac{M}{\gamma}\sqrt{t+1} \quad \Rightarrow t+1 \le \left(\frac{M}{\gamma}\right)^2$

Perceptron as the $\boxed{\text{pseudo}}$-gradient descent method.

Question: find w s.t. $y_j \langle w, x_j \rangle > 0 \ \forall j$

Know: $\exists \tilde{w}$ s.t. $y_j \langle \tilde{w}, x_j \rangle \geq 1 \ \forall j$

We can "find" $\tilde{w}$ by minimizing $F(x) = \frac{1}{2} \| x - \tilde{w} \|_2^2$ $\quad \nabla F(x) = x - \tilde{w}$

To minimize any differentiable function F, we can use the gradient descent method:

let $x_0 = 0$, for $t = 1, 2, \dots T$, $x_{t+1} = x_t - h \nabla F(x_t)$ $\quad \nabla F(\underbrace{x_1, \dots x_d}_{x}) = \left(\frac{\partial F}{\partial x_1}(x) \dots \frac{\partial F}{\partial x_d}(x) \right)$

Pseudo-gradient descent: instead of using $\nabla F(x_t)$, assume that we can find v_t such that $\langle \nabla F(x_t), v_t \rangle \geq \delta > 0$. Then define $x_{t+1} = x_t - h \cdot v_t$

$v_t = -y_j x_j$, where $y_j \langle w_t, x_j \rangle < 0$ for the perceptron.

$$\Rightarrow \langle \nabla F(w_t), -h y_j x_j \rangle = \overbrace{-h y_j \langle w_t, x_j \rangle}^{>0} + \underbrace{h y_j \langle \tilde{w}, x_j \rangle}_{\geq h} \geq h$$

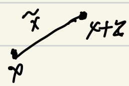
Convergence of GD

Assume that $\forall x, y \in \mathbb{R}^d$

$\| \nabla F(x) - \nabla F(y) \| \leq L \| x - y \|_2$ $\quad$ E.g. if $F(x) = \frac{1}{2} \| x - \tilde{w} \|^2$, then $\nabla F(x) - \nabla F(y) = x - y \Rightarrow L = 1$

Taylor's expansion:

$F(x+z) = F(x) + \langle \nabla F(\tilde{x}), z \rangle$, where $\tilde{x}$ is a point on an interval connecting x and $x+z$



$$\Leftrightarrow F(x+z) - F(x) = \langle \nabla F(x), z \rangle + \langle \nabla F(\tilde{x}) - \nabla F(x), z \rangle$$

Let $x_{t+1} = x_t \underbrace{- h \nabla F(x_t)}_{z}$ $\quad F(x_{t+1}) - F(x_t) = \langle \nabla F(x_t), -h \nabla F(x_t) \rangle + \langle \nabla F(\tilde{x}) - \nabla F(x_t), -h \nabla F(x_t) \rangle$

$$= -h \| \nabla F(x_t) \|^2 + \| \nabla F(\tilde{x}) - \nabla F(x_t) \| \cdot h \| \nabla F(x_t) \|$$

$$\leq -h \| \nabla F(x_t) \|^2 + L \| \tilde{x} - x_t \| \cdot h \| \nabla F(x_t) \|$$

$$\leq -h \| \nabla F(x_t) \|^2 + L \| \underbrace{x_{t+1} - x_t}_{h \| \nabla F(x_t) \|} \| \cdot h \| \nabla F(x_t) \|$$

$$\leq -h \| \nabla F(x_t) \| + L h^2 \| \nabla F(x_t) \|^2$$

If $h \leq \frac{1}{2L}$, then RHS $\leq -\frac{h}{2} \| \nabla F(x_t) \|^2$

$$\sum_{t=0}^{T} F(x_{t+1}) - F(x_t) = F(x_{t+1}) - F(x_0) \leq -\frac{h}{2} \sum_{j=0}^{T} \| \nabla F(x_j) \|^2$$

$$\Rightarrow \nabla F(x_t) \to 0 \quad \Rightarrow x_t \xrightarrow{t \to \infty} x_* \quad \text{s.t. } \nabla F(x_*) = 0$$

$\boxed{w_{t+1} = w_t + h y_j x_j}$

One the one hand,

$$F(w_{t+1}) = F(w_t) + \langle \nabla F(w_t), \underbrace{w_{t+1} - w_t}_{= h y_j x_j} \rangle + \langle \nabla F(\tilde{w}) - \nabla F(w_t), w_{t+1} - w_t \rangle$$

$\tilde{w} \in [w_t, w_{t+1}]$

$$\Rightarrow F(w_{t+1}) - F(w_t) \le h \underbrace{\langle \nabla F(w_t), y_j x_j \rangle}_{\ge h} \quad L \|w_{t+1} - w_t\|_2^2$$

$$= -h + h^2 \underbrace{\|x_j\|_2^2}_{\le M} \le -h + h^2 M^2$$

Take the sum for $t = 0, \dots, T$

$$F(w_{T+1}) - F(w_T) + F(w_T) - F(w_{T-1}) + \cdots$$

$$= F(w_{T+1}) - F(w_0) \le -hT + T h^2 M^2$$

Since the number of steps of perceptron does not depend on h

We have that

$$\frac{1}{2}\|w_{T+1} - w_*\|^2 - \frac{1}{2}\|w_*\|^2 \le -hT + T h^2 M^2$$

we know that $\|w_*\| \le \frac{1}{\gamma}$

$$\cancel{h} T \le T h^{\cancel{2}} M^2 + \frac{1}{2} \frac{\|w_*\|^2}{h} - \frac{1}{2}\|w_{T+1} - w_*\|^2$$

$$\le T h M^2 + \frac{1}{2h}\|w_*\|^2 \quad \forall h > 0$$

Optimize over $h \Rightarrow \sqrt{T} \le \left(\sqrt{2} + \frac{1}{\sqrt{2}}\right) \sqrt{T}\|w_*\| \cdot M^2$

$$T \le \left(\sqrt{2} + \frac{1}{\sqrt{2}}\right) \|w_*\|^2 \cdot M^2 = \left(\sqrt{2} + \frac{1}{\sqrt{2}}\right) \left(\frac{M}{\gamma}\right)^2$$

Logistic Regression (an example of a "generalized linear model")

It is an example of a discriminative model: namely, it specifies the form of $P(Y|X=x)$

Here, we will assume that $x \in \mathbb{R}^d$, $Y \in \{0,1\}$

Assume that $\Pr(Y=1|X=x) = p(x)$ – function of x

Remark if $p(x) > \frac{1}{2}$ => the best guess is $Y=1$, otherwise $Y=0$.

Note that (X_1, Y_1) is the observed data, then $\mathcal{L}(p(x)) = p(x_1)^{Y_1}(1-p(x_1))^{1-Y_1}$

If the training data is $(X_1, Y_1), \dots, (X_n, Y_n)$

then $\mathcal{L}(p(x)) = \prod_{j=1}^{n} p(x_j)^{Y_j}(1-p(x_j))^{1-Y_j}$

$$p(x)^Y = e^{Y \log p(x)} \qquad \Rightarrow e^{\sum_{j=1}^{n} Y_j \log p(x_j) + \sum_{j=1}^{n}(1-Y_j)\log(1-p(x_j))}$$

$$\log \mathcal{L}(p(x)) = \sum_{j=1}^{n} Y_j \underbrace{\log \frac{p(x_j)}{1-p(x_j)}}_{\text{"log odds ratio"}} + \sum_{j=1}^{n} \left(\log(1-p(x_j))\right)$$

Main assumption: $\log \frac{p(x)}{1-p(x)} = \langle w, x \rangle + b = \langle \tilde{w}, \tilde{x} \rangle$ $\quad \tilde{x} = (x, 1) \in \mathbb{R}^{d+1}$, $\tilde{w} = (w, b) \in \mathbb{R}^{d+1}$

$$\log \mathcal{L} = \sum_{j=1}^{n} Y_j \langle \tilde{w}_j, \tilde{x}_j \rangle + \sum_{j=1}^{n} \log(1 - p(x_j))$$

$$\frac{p(\tilde{x})}{1-p(\tilde{x})} = e^{\langle \tilde{w}, \tilde{x} \rangle} \Rightarrow p(\tilde{x}) = \frac{e^{\langle \tilde{w}, \tilde{x} \rangle}}{1 + e^{\langle \tilde{w}, \tilde{x} \rangle}}$$

$$\Leftrightarrow 1 - p(\tilde{x}) = \frac{1}{1 + e^{\langle \tilde{w}, \tilde{x} \rangle}}$$

Therefore, maximizing $\log \mathcal{L}(p)$ is equivalent to maximizing

$$\sum_{j=1}^{n} Y_j \langle \tilde{w}_j, \tilde{x}_j \rangle - \sum_{j=1}^{n} \log\left(1 + e^{\langle \tilde{w}, \tilde{x} \rangle}\right)$$

$$\Leftrightarrow \sum_{j=1}^{n} \log\left(1 + e^{\langle \tilde{w}_j, \tilde{x}_j \rangle}\right) - \sum_{1}^{n} Y_j \langle \tilde{w}_j, \tilde{x}_j \rangle \qquad \Rightarrow \text{convex}$$

minimize over $\tilde{w} \in \mathbb{R}^{d+1}$ $\qquad$ => it has a unique minimizer $\hat{w}$

Remark: a sufficient condition for F to be convex is that the eigenvalues of the Hessian have to be nonnegative

Let $\hat{w}$ be the optimal solution $\Rightarrow \hat{p}(x) = \dfrac{e^{\langle \hat{w}, x\rangle}}{1+e^{\langle \hat{w}, x\rangle}}$

$$\hat{p}(x) \geq \frac{1}{2} \iff e^{\langle \hat{w}, x\rangle} \geq 1$$
$$\iff \langle \hat{w}, x\rangle \geq 0$$

Boosting

$G = \{g : S \to \{+1, -1\}\}$

Example will be on BB

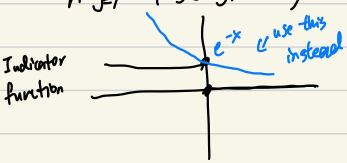
Q: What if we look at classifiers of the form $\text{sign}(\alpha g_1 + (1-\alpha) g_2)$, $g_1, g_2 \in G$? $\alpha \in (0,1)$

Recall that our goal is to find some g (a binary classifier) s.t.
$P_r(Y \neq g(x))$ is small

Note that $P_r(Y \neq g(x)) = Pr(Y g(x) < 0) = \mathbb{E}\, I\{Y g(x) < 0\}$

Any function $f : S \to \mathbb{R}$ can be transformed into a binary classifier
$g_f : \text{sign}(f) = \begin{cases} +1, & f \geq 0 \\ -1, & f < 0 \end{cases}$

Problem: given a class G of function $g : S \to \mathbb{R}$, minimize the empirical risk
$\frac{1}{n}\sum_{j=1}^{n} I\{Y_j g(x_j) < 0\}$



(1) $f(x) = e^{-x}$ is convex
$f''(x) = e^{-x} > 0$

(2) $e^{-x} \geq I\{x \leq 0\}$

Instead, consider the problem $\frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(x_j)} \longrightarrow$ minimize over $g \in G$

The function $z \mapsto e^{-z}$ is convex, so this can often be done numerically.

Question since $\forall g \in G$, $\frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(x_j)} \xrightarrow{P} \mathbb{E}\, e^{-Y g(x)}$
it's natural to ask which g minimizes $\mathbb{E}e^{-Yg(x)}$ over all $g : S \to \mathbb{R}$

Baye's classifier

Reminder : The minimum of $\mathbb{E}\, I\{Y g(X) < 0\}$ is achieved for $g(x) = \text{sign}(\mathbb{E}(Y|X=x))$

Theorem: Let $\tilde{g}$ minimize $\mathbb{E} e^{-Yg(X)}$. Then $\text{sign}(\tilde{g}) = g^*$

Proof : Assume that X takes values $x_1, \dots, x_k$. (Discrete)

$$\mathbb{E} e^{-Yg(X)} = \sum_{j=1}^{k} \mathbb{E}[e^{-Yg(X)} | X = x_k] \Pr(X = x_k)$$

$$\mathbb{E}[e^{-Yg(X)} | X=x_k] = e^{1 \cdot g(x_k)} \Pr(Y=1|X=x_k) + e^{-1 \cdot g(x_k)} \Pr(Y=-1|X=x_k)$$

We know that $\Pr(Y=1|X=x_k) = \frac{1+\eta(x_k)}{2}$ $\quad \Pr(Y=-1|X=x_k) = \frac{1-\eta(x_k)}{2}$

let $g(x_k) = t$

where $\eta(x_k) = \mathbb{E}[Y|X=x_k]$ $\quad \Pr(Y=t|X=x_k) = \frac{1+t\eta(x_k)}{2}$ $\quad t = 1$ or -1

Therefore, it suffices to minimise $F(t) = e^{-t} \frac{1+\eta(x_k)}{2} + e^{t} \frac{1-\eta(x_k)}{2}$ over $t \in \mathbb{R}$

$F(t) > 0$, $F''(t) = F(t) > 0$, F is convex

$$F'(t) = -\frac{1}{e^t} \frac{1+\eta(x_k)}{2} + e^t \frac{1-\eta(x_k)}{2} = 0$$

$$= -(1+\eta(x_k)) + e^{2t}(1-\eta(x_k)) = 0 \Rightarrow e^{2t} = \frac{1+\eta(x_k)}{1-\eta(x_k)}, \quad t = \frac{1}{2}\log\frac{1+\eta(x_k)}{1-\eta(x_k)}$$

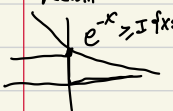
$\mathcal{F}$ - "base class" of binary classifiers (e.g. threshold classifiers)

$$G = \{\sum_{j=1}^{k} \alpha_j f_j,\ k \geq 1,\ \alpha_1, \dots, \alpha_k \geq 0,\ f_1, \dots f_k \in F\}$$

$\sum_{j}^{k} \alpha_j = 1$

Any $g \in G$ can be transformed into a binary classifier via $g \to \text{sign}(g)$

Recall that for any binary classifier h, $I\{Y \neq h(X)\} = I\{Yh(X) < 0\}$



$e^{-x} \geq I\{x \leq 0\}$

$\mathbb{E}\, I\{Y g(X) < 0\}$ is minimized for $g_*(x) = \text{sign}(\mathbb{E}(Y|X=x))$

Proof

Th The expression $\mathbb{E} e^{-Yg(X)}$ is minimized for $\tilde{g}(x) = \frac{1}{2}\log\frac{1+\eta(x)}{1-\eta(x)}$, where $\eta(x) = \mathbb{E}(Y|X=x)$

$$\text{sign}\left(\frac{1+\eta(x)}{1-\eta(x)}\right) = 1 \iff \frac{1+\eta(x)}{1-\eta(x)} \geq 1 \iff \eta(x) \geq 0 = \text{sign}(\eta(x))$$

$\Rightarrow$ we recover the Bayes classifier!

Summary : minimizing $\mathbb{E} e^{-Yg(X)}$ over all functions g gives us a Bayes classifier.

$\Rightarrow$ it makes sense to look at the "empirical" version of this problem,

$\frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(X_j)}$ where $(X_1, Y_1) \dots (X_n, Y_n)$ is the training data.

Let G be a class of function, and let us consider minimizing $\frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(X_j)}$ over $g \in G$

Definition : We will say that a class F of binary classifier satisfies the following: for any $n \geq 1$, any $(x_1, y_1), \dots, (x_n, y_n)$, any nonnegative weights $w_1, \dots, w_n$ s.t. $\underbrace{\sum_1^n w_j = 1}_{\text{like probability}}$, $\exists f \in F$ s.t. $\sum_1^n w_j I\{y_j \neq f(x_j)\} \leq \frac{1}{2}$

Remark: If $f \in F \iff -f \in F \Rightarrow$ then F satisfies the weak learnability assumption.

if loss $f > \frac{1}{2}$, we pick $-f$ which is $< \frac{1}{2}$ satisfied

$$\frac{1}{n}\sum_1^n e^{-y_j g(x_j)} \longrightarrow \text{minimize over } g \in G$$

Using the notion of pseudogradient descent

Assume that, at iteration t, we have $g_t \in G$

$$\frac{1}{n}\sum_1^n e^{-y_j[g_t(x_j) + \alpha f(x_j)]}$$

Goal: find $\alpha \in \mathbb{R}$, $f \in F$ that make this expression as small as possible

The function f can be viewed as a "proxy" to the gradient.

The methods of this type are referred to as "steepest descent" methods. umm ok.

$$= \frac{1}{n}\sum_1^n e^{-y_j g(x_j)} e^{-y_j \alpha f(x_j)}$$

Let $w_j = \frac{1}{n} e^{-y_j g(x_j)} > 0$

Then, we are trying to minimize $\sum_{j=1}^n w_j e^{-\alpha y_j f(x_j)}$ over $f \in F$, $\alpha \in \mathbb{R}$

If $\tilde{w}_j = \frac{w_j}{\sum_1^n w_j}$ so that $\sum_1^n \tilde{w} = 1$, then we need to minimize $\sum_1^n \tilde{w}_j e^{-\alpha y_j f(x_j)}$

Note that $\sum_1^n \tilde{w}_j e^{-\alpha y_j f(x_j)} = \sum_1^n \tilde{w}_j e^{-\alpha} I\{y_j = f(x_j)\} + \sum_1^n \tilde{w}_j e^{\alpha} I\{y_j \neq f(x_j)\}$

$$\pm \sum_1^n \tilde{w}_j e^{-\alpha} I\{y_j \neq f(x_j)\} = e^{-\alpha}\sum_1^n \tilde{w}_j + (e^{\alpha} - e^{-\alpha})\sum_1^n \tilde{w}_j I\{y_j \neq f(x_j)\}$$

$$= e^{-\alpha} + (e^{\alpha} - e^{-\alpha})\sum_1^n \tilde{w}_j I\{y_j \neq f(x_j)\}$$

To minimize this expression, (1) minimize $\sum_1^n \tilde{w}_j I\{y_j \neq f(x_j)\}$ WRT $f \in F$

(2) minimize wrt α

Let $e_{n,\tilde{w}}(f) = \sum_1^n \underbrace{\tilde{w}_j}_{\Pr(X = x_j)} I\{y_j \neq f(x_j)\}$

Weak Learnability $\Rightarrow \forall \{\tilde{w}_j\}_{j=1}^n$, $\exists f \in F$ s.t. $e_{n,\tilde{w}}(f) \leq \frac{1}{2}$

Next, the minimum $\alpha \to e^{-\alpha} + (e^{\alpha} - e^{-\alpha}) e_{n,\tilde{w}}(f)$ is achieved

$$\hat{\alpha} = \frac{1}{2} \log \frac{1 - e_{n,\tilde{w}}(f)}{e_{n,\tilde{w}}(f)}$$

AdaBoost

Initialize $w_j^{(1)} = \frac{1}{n}$, $j = 1, \dots n$, $g_0 = 0$ for $t = 1, \dots, T$ do

(i) Call the weak Learner

(ii) WL outputs f_t s.t. $e_{n,w^{(t)}}(f_t) \leq \frac{1}{2}$

(iii) $\alpha_t = \frac{1}{2} \log \frac{1 - e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)} \geq 0$

(iv) Update the weights

$$w_j^{(t+1)} = \frac{w_j^{(t)} e^{-Y_j \alpha_t f_t(X_j)}}{Z_t}$$

Z_t - normalizing factor

$$Z_t = \sum_1^n w_j^{(t)} e^{-Y_j \alpha_t f_t(X_j)}$$

Output $\hat{g}_T = \frac{\sum_{t=1}^T \alpha_t f_t}{\sum_{t=1} \alpha_t}$

Theorem Assume that for any probability $w_1, \dots w_n$, the WL finds f st $\sum_1^n w_j I\{Y_j \neq f(X_j)\} \leq \frac{1}{2} - \gamma$ for some $\gamma > 0$. Then the training error of AdaBoost satisfies

$$\frac{1}{n} \sum_1^n I\{Y_j \neq \mathrm{sign}(\hat{g}_T(X_j))\} \leq e^{-2T\gamma^2}$$

Proof: $\frac{1}{n} \sum_1^n I\{Y_j \neq \mathrm{sign}(\hat{g}_T(X_j))\} = \frac{1}{n} \sum_1^n I\{Y_j \hat{g}_T(X_j) < 0\} \leq \frac{1}{n} \sum_1^n e^{-Y_j \hat{g}_T(X_j)} = \frac{1}{n} \sum_1^n e^{-Y_j \sum_1^T \alpha_t f_t(X_j)}$

Note that $w_j^{(t+1)} = \frac{1}{n} \frac{e^{Y_j \sum_1^t \alpha_i f_i(X_j)}}{\prod_1^t Z_j}$

$$e^{-Y_j \sum_1^T \alpha_i f_i(X_j)} = n \prod_1^T Z_j w_j^{(T+1)}$$

$$Z_t = \sum_1^n w_j^{(t)} e^{-\alpha_t Y_j f_t(X_j)}$$
$$= e^{-\alpha_t} + (e^{\alpha_t} - e^{-\alpha_t}) \sum_1^n w_j^{(t)} I\{Y_j \neq f_t(X_j)\}$$

Plug in $\alpha_t = \frac{1}{2} \log \frac{1 - e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}$

$e_{n,w^{(t)}}(f_t) = \sum_1^n w_j^{(t)} I\{Y_j \neq f_t(X_j)\}$

$$Z_t = 2\sqrt{\underbrace{e_{n,w^{(t)}}(f_t)}_{\leq \frac{1}{2} - \gamma}(1 - e_{n,w^{(t)}}(f_t))} \leq 2\sqrt{(\tfrac{1}{2} - \gamma)(\tfrac{1}{2} + \gamma)}$$

Regression and Linear Regression

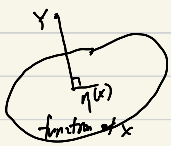
Assume that Y can take values beyond $\{-1,+1\}$ (or $\{0,1\}$), specifically assume that $Y \in \mathbb{R}$

- Y will be called the "response variable".
- The goal is to predict Y based on the observation X
  - $X \in \mathbb{R}^d$, the coordinates of X are called "features".
  - X is also called the "predictor variable".

Example Predict the final exam grade $= Y$ base on $\overbrace{(\text{midterm grade}, hw1, hw2)}^{\text{predictor}}$ features

Reminder: The condition expectation of Y given $X=x$, denoted $\eta(x)$, minimize $\mathbb{E}_{Y|X=x}(Y-z)^2$

In other words, $\eta(x)=\mathbb{E}(Y|X=x)$ is the best functional approximation of Y as a function of X



Mathematically.

$\eta(x)$ minimizes $\mathbb{E}(Y-f(x))^2$ over all functions f.

Given the training data $(X_1,Y_1),\dots,(X_n,Y_n)$, we consider the problem of minimizing $\frac{1}{n}\sum_{1}^{n}(Y_j-f(X_j))^2$ over $f \in F$

Question: Let $\hat{f}_n$ be the solution of the problem. What is $\mathbb{E}(Y-\hat{f}_n(X))^2$?

Note that
$$\mathbb{E}(Y-\hat{f}_n(X))^2 = \mathbb{E}(Y-\eta(X)+\eta(X)-\hat{f}_n(X))^2$$
$$= \mathbb{E}(Y-\eta(X))^2 + \mathbb{E}(\eta(X)-\hat{f}_n(X))^2 + 2\underbrace{\mathbb{E}(Y-\eta(X))(\eta(X)-\hat{f}_n(X))}_{0}$$
orthogonal
$$= \mathbb{E}(Y-\eta(X))^2 + \mathbb{E}(\eta(X)-f_n(X))^2$$
$$\le \mathbb{E}(Y-\eta(X))^2 + \mathbb{E}(\eta(X)-\bar{f}(X))^2 \leftarrow \text{approximation error}$$
$$+\mathbb{E}(\bar{f}(X)-\hat{f}_n(X))^2 \leftarrow \text{training error}$$

$\Downarrow$

$\mathbb{E}\,\mathbb{E}\left[(Y-\eta(X))(\eta(X)-\hat{f}_n(X)) \mid X\right]$

$\mathbb{E}(Y|X)=\eta(X)$

app error

$$\mathbb{E}(\eta(x) - \bar{f}(x)) = 0$$

Comparison to Mathematical statistics.

$Y = \alpha X + \beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$

MLE of (α, β) is given by the solution of

$\frac{1}{n}\sum_{1}^{n}(Y_j - \alpha' X_j - \beta')^2 \ \longrightarrow$ minimize over α', β'.

<u>Fact</u>: If (Y, X) has bivariate normal distribution then.

$$\mathbb{E}(Y \mid X = x) = \alpha x + b \ !$$

<u>Error decomposition</u> in linear regression

$$\mathbb{E}(Y - g(x))^2 = \mathbb{E}(Y - \eta(x))^2 + \underbrace{\mathbb{E}(\bar{g}(x) - \eta(x))^2}_{\text{app error of } G} + \mathbb{E}(g(x) - \bar{g}(x))^2$$

In statistics, a common assumption is that (X, Y) has multivariate normal distribution. In this case, $\eta(x) = \langle w_*, x \rangle + b_*$ is a linear function of x, and $\underbrace{Y - \eta(x)}_{\varepsilon}$ is normally distributed. $Y = \langle w_*, x \rangle + b_* + \varepsilon$

- Allows to do inference: build confidence intervals / test statist hypothesis.

Solution of linear regression problem

$G = \langle g_{w,b}(x) = \langle w, x \rangle + b \rangle$

Goal. find $\hat{w}, \hat{b}$ that minimize

$\frac{1}{n}\sum_{j=1}^{n}(Y_j - \langle w, x_j \rangle - b)^2$ over $w \in \mathbb{R}^d, b \in \mathbb{R}$

Simplify: $\tilde{X}_j = (X_j, 1) \in \mathbb{R}^{d+1}$

$\tilde{w} = (w, b) \ \mathbb{R}^{d+1}$

$\langle \tilde{w}, \tilde{x}_j \rangle = \langle w, x_j \rangle + b$

Let $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbb{X} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix}$

$$\begin{pmatrix} Y_1 - \langle \tilde{x}_1, \tilde{w} \rangle \\ Y_2 - \langle \tilde{x}_2, w \rangle \\ \vdots \\ Y_n - \langle \tilde{x}_n, w \rangle \end{pmatrix} = \vec{Y} - \mathbb{X}\tilde{w}$$

Then $\frac{1}{n}\sum_j^n (Y_j - \langle\tilde{\omega}, \tilde{x}_j\rangle)^2 = \frac{1}{n}\|\vec{Y} - \mathbb{X}\tilde{w}\|_2^2 = F(w)$

If $\mathbb{X} = (x^{(1)} | x^{(2)} | \dots | x^{(d+1)})$

$H(w) = \mathbb{X}w$, $\nabla H(w) = \mathbb{X}$

$\nabla F(w) = -2\mathbb{X}^T(Y - \mathbb{X}\tilde{w}) = 0$

$(\mathbb{X}^T\mathbb{X})\tilde{w} = X^TY$ (normal equations)

If $(\mathbb{X}^T\mathbb{X})$ is invertible

$\mathbb{X} \in \mathbb{R}^{n \times (d+1)}$, $n > d+1$

$\mathbb{X}^T\mathbb{X} \in \mathbb{R}^{d+1 \times d+1}$

$\therefore \hat{w} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$

Continue:

$\hat{w} = (X^TX)^{-1}X^TY$ if X^TX is invertible.

$\Leftrightarrow \underbrace{(X^Tx)}_{A}\hat{w} = \underbrace{x^TY}_{b}$

$A\hat{w} = b$

$X^TX = (X^TX)^T \Rightarrow X^TX = V\Lambda_x V^T$ where $V = (V_1|\dots|V_p)$ is a matrix of eigen vectors.

$\Lambda_x^{-1} = \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_p^{-1} \end{pmatrix}$, $\lambda_1 \dots \lambda_p$ - eigenvalues.

$V\Lambda_x V^T\hat{w} = X^TY$ Since $V^TV = I_p$, $\Lambda_x V^T\hat{w} = V^TX^TY$ $\quad V^T\hat{w} = \Lambda_x^{-1}(X^TY)$

The Ridge Regression

Let $\lambda > 0$ - the "regularization parameter"

$\frac{1}{n}\|\vec{Y} - \mathbb{X}w\|_2^2 + \underbrace{\lambda\|w\|_2^2}_{\text{regularization/penalty term}} \to$ minimize over $w \in \mathbb{R}^{d+1}$

Tikhonov regularization

$F(w) = \frac{1}{n}\|\vec{Y} - \mathbb{X}w\|_2^2 + \lambda\|w\|_2^2$

$\nabla F(w) = -\frac{2}{n}\mathbb{X}^T(Y - \mathbb{X}w) + 2\lambda w$

$\nabla F(w) = 0 \Leftrightarrow \frac{2}{n}\mathbb{X}^TY - 2(X^TX)w + \lambda w$

$\tilde{w}$ solves the system

$$\mathcal{X}^T\mathcal{X}w + \lambda I \cdot w = \mathcal{X}^T Y \iff (\mathcal{X}^T\mathcal{X} + \lambda I)w = \mathcal{X}^T Y$$

$$\hat{w} = (\mathcal{X}^T\mathcal{X} + \lambda I)^{-1}\mathcal{X}^T Y$$

If $\mathcal{X}^T\mathcal{X} = V\Lambda_x V^T$, then $\mathcal{X}^T\mathcal{X} + \lambda I = V(\Lambda_x + \lambda I)V^T$

Numerical Instability problem disappears, but need to pay attention to $\lambda\|w\|_2^2$.

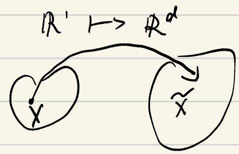
<u>Polynomial Regression</u> (linear regression for polynomial functions).

Assume that $(X,Y) \in \mathbb{R}\times\mathbb{R}$

$$G = \{p(x) = a_0 + a_1 x + \cdots + a_d x^d,\ a_0 \cdots a_d \in \mathbb{R}\}$$

feature space

<u>Idea</u>: create a mapping $x \mapsto \underbrace{(x, x^2, x^3, \ldots, x^d)}_{\tilde{x}}$



$$(X_1,Y_1),\ldots,(X_n,Y_n) \to (\tilde{X}_1, Y_1),\ldots,(\tilde{X}_n, Y_n)$$

$$\langle w, \tilde{x}\rangle = w_1 x + w_2 x^2 + \cdots + w_d x^d$$

Linear regression problem corresponds to solving $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{d} w_j x^j)^2$

Also the idea of SVM

<u>Non-learnability of Linear Regression</u>

Let $(X,Y) \in \mathbb{R}\times\mathbb{R}$, and $G = \{f_w(x) = wx,\ w \in \mathbb{R}\}$

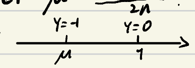
What does it mean for G to be "learnable"?

It means that $\exists$ an algorithm A, s.t. for any distribution over (X,Y), and $\varepsilon, \delta > 0$, $\exists\, n = n(\varepsilon,\delta)$, such that for all $n \geq n(\varepsilon,\delta)$, $A((X_1,Y_1)\ldots(X_n,Y_n))$ outputs $\hat{w}_n$ s.t. $\mathbb{E}(Y - \hat{w}_n X)^2 \leq \min_{w\in\mathbb{R}} \mathbb{E}(Y - wX)^2 + \varepsilon$ with probability $\geq 1-\delta$.

<u>Example</u> let $\varepsilon = 0.01$, $\delta = 0.5$, $n \geq n(\varepsilon,\delta)$

let $\mu = \frac{\log(\frac{100}{99})}{2n}$. Consider two distributions. $\dot{P}_1$ $\dot{P}_2$

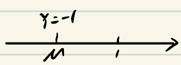
P_1:



$Pr((X,Y) = (1,0)) = \mu$

$Pr((X,Y) = (\mu,-1)) = 1-\mu$

P_2:



$Pr((X,Y) = (\mu,-1)) = 1$

For P_1, $Pr(X_1 = X_2 = \dots X_n = \mu) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$

Since $1-\mu \geq e^{-2\mu} = 1 - 2\mu + \frac{(2\mu)^2}{2} = 1-2\mu+2\mu^2$

$1-\mu \geq 1-2\mu+2\mu^2$ when $\mu \approx 0$

For P_2, $Pr(X_1 = \dots = X_n = \mu) = 1$

We don't know whether the observation comes from which distr

$\Rightarrow$ A will produce the same output $\hat{w}_n$ regardless of the distribution.

(i) $|\hat{w}_n| < \frac{1}{2\mu}$, then $\mathbb{E}_{P_2}(Y-\hat{w}_n X)^2 = \mathbb{E}(-1-\hat{w}_n \mu)^2$ $\quad |\hat{w}_n \mu| < \frac{1}{2}$

$$\geq (1-\tfrac{1}{2})^2 = \tfrac{1}{4}$$

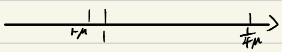
But $\min_w \mathbb{E}_{P_2}(Y-wX)^2 = 0$, for $w = -\frac{1}{\mu}$

(ii) $|\hat{w}_n| \geq \frac{1}{2\mu}$ Consider P_1: $\mathbb{E}_{P_1}(Y-\hat{w}_n X)^2 = \mu(0-\hat{w}_n \cdot 1)^2 + (1-\mu)(\cancel{-1-\hat{w}_n \mu})^2$

$$\geq \mu \cdot w_n^2 \geq \frac{1}{4\mu}$$

$$\mu = \frac{\log(\frac{100}{99})}{2n}$$

But $\min_w \mathbb{E}_{P_1}(Y-wX)^2 \leq \mathbb{E}_{P_1}(Y-0\cdot X)^2 = 1-\mu$



$\Rightarrow \frac{1}{4\mu} - (1-\mu) > \varepsilon = 0.01$ for n large enough

Remark: (a) To make the problem learnable, we need to assume that

(i) $\|x\|_2 \leq M$ $\quad$ (textbook details)

(ii) $\|w\|_2 \leq R$

(b) Compare this to Gaussian linear regression:

$Y = \alpha X + \varepsilon$, X, ε are independent, normally distributed, the no assumption on α is required.

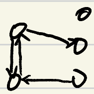
Artificial Neural Nets

· Feedforward neural networks

(V, E) - a graph

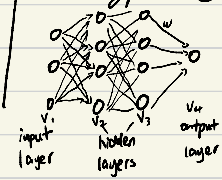
V: a set of vertices or nodes

E: edges



A typical graph $V = \bigcup_{t=0}^{T} V_t$

T - the depth of a network



Each edge in E connects a vertice in V_t to a vertice in V_{t+1} for some t

Nodes correspond to "artificial neurons".

Each neuron is modeled by an "activation function" $\sigma: \mathbb{R} \to \mathbb{R}$, such as

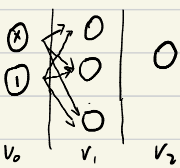
(a) $\sigma(x) = I\{x \geq 0\}$

(b) $\sigma(x) = \frac{1}{1+e^{-x}}$ (sigmoid)

(c) $\sigma(x) = \max(0, x)$ (ReLU, rectified linear units).

Let $O_{t,i}(x)$ be the output of neuron i in level t when given input $x \in \mathbb{R}^d$

By design, $O_{0,j}(x) = x_j$, $O_{0,d+1}(x) = 1$. The input to $V_{t+1,j}$ (j-th neuron in layer $t+1$)



$$a_{t+1,j} = \sum_{r(v_{t,r}, v_{t+1,j}) \in E} W(\underbrace{(t,r),(t+1,j)}_{\text{edge}}) O_{t,r}(x)$$

$O_{0,1}(x) = x$

$O_{0,2}(x) = 1$

$a_{1,1} = W((0,1),(1,1)) \cdot x + W((0,2),(1,1)) \cdot 1$

$a_{1,2} = W((0,1),(1,2)) \cdot x + W((0,2),(1,2)) \cdot 1$

$a_{1,3} = W((0,1),(1,3)) \cdot x + W((0,2),(1,3)) \cdot 1$

Apply activation function: $O_{1,1} = \sigma(a_{1,1})$, $O_{1,2} = \sigma(a_{1,2})$, $O_{1,3} = \sigma(a_{1,3})$

$a_{2,1} = W((1,1),(2,1)) \cdot O_{1,1} + W((1,2),(2,1)) \cdot O_{1,2} + W((1,3),(2,1)) \cdot O_{1,3}$

Explicitly: $a_{2,1} = W((1,1),(2,1)) \sigma(W((0,1),(1,1)) \cdot x + W((0,2),(1,1)) \cdot 1)$

$+ W((1,2),(2,1)) \sigma(W((0,1),(1,2)) \cdot x + W((0,2),(1,2)) \cdot 1)$

$+ W((1,3),(2,1)) \sigma(W((0,1),(1,3)) \cdot x + W((0,2),(1,3)) \cdot 1)$

$G_{V,E,\sigma} = \{ g_{V,E,\sigma,w} \ , \ w : E \to \mathbb{R} \}$ Graph representation

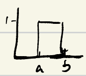
Question: how expressive can these classes be?

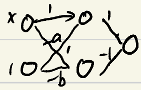
Th. Let $f : [0,1] \to \mathbb{R}$ that is continuous.

Then $\forall \varepsilon > 0$, $\exists (V,E)$ and weights $w \in \mathbb{R}^{|E|}$, such that $| g_{V,E,\sigma,w}(x) - f(x) | \le \varepsilon$, $\forall x \in [0,1]$.

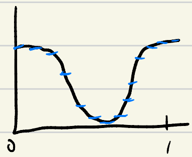
We will take $\sigma(x) = I\{x \geq 0\}$.

$I\{x \in (a,b)\} = \sigma(x-a) - \sigma(x-b)$





$\forall \varepsilon > 0$, $\exists k$ and $0 < x_1 < \cdots < x_k < 1$ s.t. $\forall j \le k$, $\left| f(x) - f\left(\frac{x_j + x_{j-1}}{2}\right) \right| \le \varepsilon$ for $x \in [x_{j-1}, x_j)$

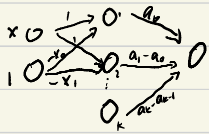


$\tilde{f}(x) = \sum_{j=1}^{k} a_j \, I\{x \in [x_{j-1}, x_j)\}$. Here, $a_j = f\left(\frac{x_j + x_{j-1}}{2}\right)$

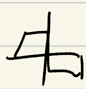
Therefore, $\tilde{f}(x) = \sum_{j=1}^{k} a_j (\sigma(x - x_{j-1}) - \sigma(x - x_j))$

$= \sum_{j=1}^{k} a_j \sigma(x - x_{j-1}) - \sum_{j=1}^{k} a_j \sigma(x - x_j)$

$= a_0 \sigma(x - x_0) + \sum_{j=1}^{k-1} (a_{j+1} - a_j) \sigma(x - x_j) - a_k \sigma(x - x_k)$



1 hidden layer nn can approximate and continuous function.

final: given  , approximate it with nn

Gradient Descent of NN

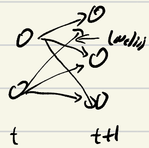
$G = \{ g_{V,E,\sigma,w},\ w \in \mathbb{R}^{|E|} \}$

Goal: $\min F(w) = \frac{1}{2} \sum_{j=1}^{n} (Y_j - g_w(X_j))^2$ over $w \in \mathbb{R}^{|E|}$

(V, E), $V = \bigcup_{t=0}^{T} v_j$. $V_t = (v_{t,1}, \dots, v_{t,k_t})$, $W_t \in \mathbb{R}^{k_{t+1} \times k_t}$

$(W_t)_{i,j}$ = weight on the edge b/w $v_{t+1,i}$ and $v_{t,j}$

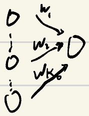
$W = (W_0, \dots, W_{T-1})$, $F(w) = \frac{1}{2} \sum_{i}^{n} (Y_i - g_w(X_i))^2$



Pick some t, $i \le k_{t+1}$, $j \le k_t$

$\frac{\partial}{\partial (W_t)_{i,j}} F(w) = \sum_{j=1}^{n} (g_w(X_j) - Y_j) \frac{\partial}{\partial (W_t)_{i,j}} g_w(X_j)$

Ex. (a) $T = 1$ => no hidden layers

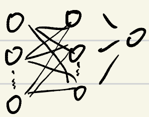


$g_w(x) = \sigma(W_0 \cdot X) = \sigma(W_0 \cdot O_0)$ input, output 0 layer

$\frac{\partial}{\partial (W_0)_i} g_w(x) = \sigma'(W_0 \cdot O_0)(O_0)_j$

$\nabla_W g_w(x) = \sigma'(W_0 \cdot O_0) O_0$

Ex 2:



1 hidden layer

O_0 – output of layer 0

$a_1 = W_0 O_0$ – inputs of layer 1

$O_1 = \sigma(W_0 O_0) = \sigma(a_1) = \begin{pmatrix} \sigma(\langle W_{0,1}, O_0 \rangle) \\ \vdots \\ \sigma(\langle W_{0,k_1}, O_0 \rangle) \end{pmatrix}$

$a_2 = W_1 O_1$

$O_2 = \sigma(a_2) = \sigma(W_1 O_1) = \sigma(W_1 \sigma(W_0 O_0))$

$\nabla g_{W_1} = \sigma'(W_1 \sigma(W_0 O_0)) \cdot O_1$

Differentiate w.r.t. W_0 $\nabla g_{W_0} = \sigma'(W_1 \sigma(O_0 \vec{W_0})) W_1 \sigma'(O_0 \vec{W_0}) O_0$, where $\sigma'(O_0 \vec{W_0}) = \begin{pmatrix} \sigma'(\langle (W_1)_1, O_0 \rangle) & & \\ & \ddots & \\ & & \sigma'(\langle W_{k_t}, O_0 \rangle) \end{pmatrix}$

Remark $W_t \in \mathbb{R}^{k_t \times k_{t-1}}$, k_t = # of nodes in layer t

$\begin{pmatrix} 1 \\ 2 \\ \vdots \\ k_t \end{pmatrix} \to \begin{pmatrix} 1 & 2 & \cdots & k_t \end{pmatrix}$

$O_{t-1} \in \mathbb{R}^{k_{t-1}} \to \begin{pmatrix} O_1^T & 0 & & 0 & 0 \\ & & & & 0 \\ 0 & O_2^T & & & 0 \\ \vdots & & \ddots & & \\ 0 & & & 0 & O_{t-1}^T \end{pmatrix} = O_{t-1}$

Then $W_t O_{t-1} = O_{t-1} w_t$

In general, $\nabla_{W_0} \sigma(W_{T-1}\, \sigma(W_{T-2}\, (\dots \sigma(W_0 O_0)\,))$

$$= \sigma'(W_{T-1} O_{T-1}) W_{T-1} \sigma'(W_{T-2} O_{T-2}) \times \cdots \times W_1 \sigma'(W_0 O_0) O_0$$