Probability Theory

Intro – Ds:

$\langle P, \mathcal{F}, \Omega \rangle$

$P: \mathcal{F} \to [0,1]$

① $0 \le P \le 1$

② $P(\Omega) = 1$

③ $P(\bigcup A_i) = \sum P(A_i)$

* LTRF $N \to P$

Lemma: $P(A^c)$, $P(A \cup B)$,

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ — Bayes

$P(\cdot|\cdot)$ ① ② ③ ✓

independence

Total probability

RV: $\langle \Omega, \mathcal{F}, P \rangle$ $X: \Omega \to \mathbb{R}$

(C) DF: $F(x) = P(X \le x)$, $\forall x \in \mathbb{R}$

$\text{card}(X) = ?$

$\aleph_1$

CDF: $\exists f: \mathbb{R} \to [0, +\infty]$ — PDF

$F(x) = P(X \le x) = \int_{(-\infty, x)} f \, d\nu$

Th: $P(X = x) = 0$

$f(x) = F'(x)$

$\int_{\mathbb{R}} f = 1$

C

PMF: $X \xrightarrow{f} [0,1]$

$f(x) = P(X = x)$

① transformation

② Expectation: $E(X) = \int x \, dF_X$ / $(X \in L^1(P))$ lemma: $E[E(X|Y)] = E(X)$

Variance: $X \in L^2(P)$ ..... $X \in L^k(P)$

③ $\mathbb{R}^n$: JDF: $F_{\vec{X}}(\vec{x}): \mathbb{R}^m \to [0,1]$

$F = P(\bigcap X_i \le x_i)$

$F_{X|Y}(x|A) = \frac{P(X \le x, Y \in A)}{P(Y \in A)}$

Th: Markov: $X \in L^1(P)$, $X \ge 0$

$P(X \ge \varepsilon) \le \frac{E(X)}{\varepsilon}$, $\forall \varepsilon > 0$

chebychev': $P(|X - E(X)| \ge \varepsilon) \le \frac{V(X)}{\varepsilon^2}$

Lemma: $P(X \in [a,b]) = 1$ & $E(X) = 0$

$\forall \lambda \in \mathbb{R}$, $E[e^{\lambda X}] \le e^{\frac{\lambda^2 (b-a)^2}{8}}$

* TH: $X_1 \dots X_n$ IID F. $P(X_i \in [a,b]) = 1$ *

$P(|\bar{X}_n - E(\bar{X}_n)| \ge \varepsilon) \le 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$, $\forall \varepsilon > 0$

→AP: $CI = (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$, α

?: sub-Gaussian (?) $E[e^{s(X - E(X))}] \le e^{\frac{s^2 \lambda^2}{2}}$, $\forall s$

sub-exponential (?) $E[e^{s(X - E(X))}] \le e^{\frac{s^2 \lambda^2}{2}}$, $|s| \le \frac{1}{\lambda}$

Th: s-G(σ) $\Rightarrow P(\bar{X}_n - E(\bar{X}_n) \ge \varepsilon) \le e^{-\frac{n\varepsilon^2}{2\sigma^2}}$

s-e(λ) $\Rightarrow \sim \le e^{-\frac{n\varepsilon}{2\lambda^2}} \vee e^{-\frac{(\varepsilon+1)n}{2\lambda}}$

Lemma:

$X \sim S\text{-}G(\lambda) \Rightarrow aX \sim (|a|\lambda)$

$X \sim s\text{-}e(\lambda) \Rightarrow aX \sim (|a|\lambda)$

$X \sim S\text{-}G(\lambda) \Rightarrow X \sim S\text{-}\alpha(\lambda)$

$P(X \in [a,b]) = 1 \Rightarrow X \sim s\text{-}G(\frac{b-a}{2})$

$X \sim S\text{-}G(\lambda) \Rightarrow X^2 \sim 8\lambda^2$

X, Y (d) $\sim S\text{-}G(\sigma_i) \Rightarrow X + Y \sim S\text{-}G(\sqrt{\sigma_1^2 + \sigma_2^2})$

* form

Convergence:

$X_n \to X$: $\lim F_n(t) = F(t)$

$X_n \xrightarrow{P} X$: $\lim P(|X_n - X| > \varepsilon) = 0$

$X_n \xrightarrow{a.s.} X$: $P\{\lim X_n(\omega) = X(\omega)\} = 1$

$E[|X_n - X|^p] \to 0$

LLN: $X_i \in L^2(P)$ iid. $E(X_i) = \mu$

$\bar{X}_n \xrightarrow{L^2} \mu$

$\bar{X}_n \xrightarrow{a.s.} \mu$

CLT: $X_i \in L^2(P)$ iid. $E(X_i) = \mu$, $V(X_i) = \sigma^2$

$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} \Rightarrow Z \sim N(0,1)$

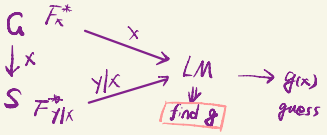
* Beaneffs Inequality

Estimation

Fun: (data)

Stat model: $S = \{f_{X,\theta} : \theta \in \Theta\}$ — pm: θ^k — npm: θ^∞.

SLP: $\langle G, S, LM \rangle$:



how produce an approx?

↓ a measure of quality in math ⇒ Stat M: $S = \{F_{X,\theta}\}$

↓ loss function $L(z(x,y), g(z(x,y)))$ 干扰 — Model space: $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda\}$ 特殊的函数 model

$Z = (X,Y) \sim F_{XY}$ — Loss: $L : [0,1] \times \mathbb{R} \to \mathbb{R}$

$R(g) = E[L(Z,g)]$ 的 ↑ — $g_R = \underset{g \in \mathcal{M}}{\operatorname{argmin}} R(g) \Leftrightarrow \lambda_R = \underset{\lambda \in \Lambda}{\operatorname{argmin}} R(g_\lambda)$

Eg: Regression: $(X,Y) \sim F_{XY}^*$, $L(a,b) = (a-b)^2$ ⇒ MSE, $f = \hat{g}$

$R(f) = E[(Y - f(X))^2]$ — $r(x) = E(Y|X=x)$

$= E[(Y - r + r - f(X))^2] = I + II + III$

$= \underbrace{E[(Y - r(X))^2]}_{\text{Var}} + \underbrace{E[r(X) - f(X)]^2}_{\text{bias}}$

$\underset{f(x)}{\operatorname{argmin}} R(f) = r(x) \Leftrightarrow$ MSE 下 — g 为 $r(x)$ 的无偏估计

Eg: Classification ⇒ Y_i is discrete ⇒ 0-1 loss ⇒ $\mathbb{R} \Rightarrow R(h^*)$

Eg: $S = \{P_\alpha(x), \alpha \in \mathbb{R}^d\}$ — JDF — loss: $L(x,\alpha) = -\ln P_\alpha(x)$

$R(\alpha) = E[L(X,\alpha)]$ — Empirical: $\hat{R}(\alpha) = \frac{1}{n}\sum(-\ln P_\alpha(x_i))$ ($\hat{R} = \frac{1}{n}\sum L_i$) — entropy — coding skill (optim)

Eg ⟷ Eg: 有监督 无监督

$f_{XY} = f_{Y|X} \cdot f_X \Rightarrow \ln f_{XY} = \ln f_{Y|X} + \ln f$

$f_{Y|X} = P_\alpha^*(y|x)$

⇒ linear Reg: ~

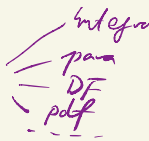
logistic Reg: $P_{(\omega)}(Y|X) = \text{Ber}(p) = \text{Ber}(G(ax+b))$, $G = \frac{1}{1+e^{-x}}$

$P_\alpha(y|x) = p^y(1-p)^{\bar{y}}$ $\hat{R} \Rightarrow$ cross entropy

Point:

parametric:

Point est: guess (point)



⇒ data: $\mathbb{X}_{(\omega)}^{m \times n} : \Omega \to \mathbb{X} \subset \mathbb{R}^{m \times n}$ Matrix

Statistic: $T(x) : \mathbb{X} \to \mathcal{T}$

point esti: $\hat{\theta} : \mathbb{X} \to \Theta$ ↓ sampling distri

⇒ measure quality:

$bias(\hat{\theta}(\mathbb{I})) = E(\hat{\theta}) - \theta$

$sde(\hat{\theta}(x)) = \sqrt{Var(\hat{\theta}_n)}$

$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

prop: $MSE = bias^2 + Var$

→ unbiaed (asym)

$\hat{\theta}_n \xrightarrow{P} \theta$ (consisted)

Non-parametric

empirical DF: $\hat{F}_n(x, \mathbb{I}) = \frac{1}{n}\sum \mathbb{1}_{X_i \le x}$

lemma: $E(\hat{F}_n) = F(x)$

$V(\hat{F}_n) = \frac{F(1-F)}{n}$

DKW inequality: $X_i \overset{iid}{\sim} F$. $\forall \varepsilon > 0$:

$P(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \le 2e^{-2n\varepsilon^2}$

⇒ CI: $(\hat{F} - \varepsilon, \hat{F} + \varepsilon)$

RV – Generation

UPRNG : $\frac{N_n(A)}{n} \to \int_A dx$

pseudo-random on $\mathcal{U}$: $\frac{N_n(a)}{n} \to \frac{1}{M}$

$\longrightarrow$ how to make a UPRNG

① Congruential generator : (a, b, M), $D(x) = (ax+b) \bmod M$

$\Rightarrow$ period ? pseudo-random

Lemma : CG : D on $\{0 \dots M-1\}$ with period M

$\Rightarrow$ $\mu_i = D(\mu_{i-1})$ is pseudo-random

? how to have M as period

Th : Hull-Dobell D : (a, b, M) has period M.

iff : $\gcd(b, M) = 1$

$a-1 \bmod p = 0$ for every prime p that divide M

$a-1 \bmod 4 = 0$ if $M \bmod 4$

? how to generate any distribution.

Alg : Accept-Reject Sampler.

In : $f(x)$: target

$g(x)$: $f \le M g(x)$

Out : $(x_0 \dots x_{n-1}) \sim f$

X^* from g :

do

iter : t

$x = g(\cdot) \Rightarrow r(x) = \frac{f(x)}{M g(x)}$

$X_{t+1} = x \in U_{[0,1]} \le r(x)$

while : length $= l$

Markov chains (Stochastic process)

Stochastic process : $(X_\alpha \mid \alpha \in A)$ $(A = \mathbb{N})$

$\Rightarrow$ Markov chain : $P(X_{t+1} = x \mid \mathcal{F}_t) = P(X_{t+1} = x \mid X_t)$

$\longrightarrow$ homogeneous : $P(X_{t+1} \mid X_t) = P(X_{s+1} \mid X_s)$

$\longrightarrow$ transition matrix : $X^{n-1} \times X \to [0,1]$ $P_{xy} = P(X_1 = y \mid X_0 = x)$

Irreducible : $S_i \leftrightarrow S_j$ $(S_i \to S_j \Rightarrow P(X_t = j \mid X_0 = i) > 0)$

Stationary distri : $\pi P = \pi$, $\sum_{x \in \mathcal{X}} \pi(x) = 1$

Reversible : $\pi(x) P(x,y) = \pi(y) P(y,x)$

i-th state $\to$ other state = stay in i-th

Question : ① Now in i-state, n step $\to$? - state.

$\star$ P^n

② the expectation step from a first to b

$E_a = P_{aa} E_a + P_{ab} E_b + P_{ac} E_c$

$\vdots$

$E_c = \dots$

E_a : $a \to b$ steps

E_b :

③ period : $i \to i$: $\gcd$ $\gcd(K)$

④ steady state vector : $s = \lim p^{(n)}$

s is P's 1 eigen's vector

$p^{(n)} = p^{(0)} P^n$

if $\lim P^n = S$. P is normal $\Rightarrow s_i = s_j$

$P^n_{ij} > 0$

Pattern Recognization

* when $Y \in \{0 \dots k\}$: class
$\mathcal{M} = \{g_n(x) : g_n(x) \in \{0 \dots k\}\}$
0-1 loss $L(z_{(x,y)}, u) = \begin{cases} 1 & \text{if } u \neq z \\ 0 & \text{if } u = z \end{cases}$
$R(\gamma) = \mathbb{E}(L(g_n(x), y))$
$\min R \Leftrightarrow g^* = \arg\min_{g \in \mathcal{M}} R(g)$

* linear classification : $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$
$g_{w,c}(x) = \text{sign}(\langle w, x \rangle + c)$ $w \in \mathbb{R}^d$, $c \in \mathbb{R}$
$L = \frac{-\text{sign}(\langle w, x_i \rangle + c) y_i + 1}{2}$ → cross-entropy
redef: $\tilde{w} = (w, c)$, $\tilde{x} = (x, 1)$
$\Rightarrow \langle \tilde{w}, \tilde{x} \rangle = \langle w, x \rangle + c = \langle \tilde{w}, \tilde{x} \rangle$
$\Rightarrow \langle \tilde{w}, \tilde{x} \rangle = 0 \Rightarrow$ super-plane crossing $(0, \dots)$

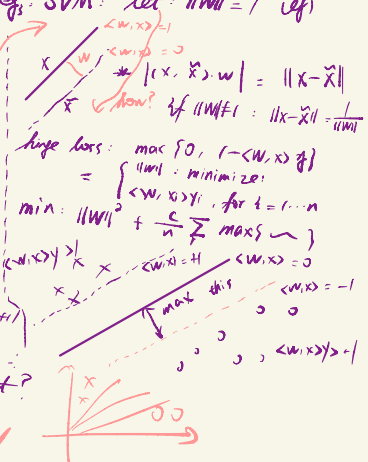
Alg 1: Perceptron Alg.
① $w = (0 \dots 0)_{d+1}$
② if $\exists \langle w, x_i \rangle y_i < 0$
$w \leftarrow w + x_i y_i$

Th: most $r^2 \|w\|^2$ updates, $r = \max \|x_i\|$
if $\exists w^*$ s.t. $x_i y_i \geq 1$, to find a w:
$w \cdot x_i y_i \geq 0$, $\forall i$

why: $\nabla_w (\langle w_i, x_i \rangle y_i)$ dim $(x_i) = d+1$
$= \vec{x}_i y_i$

But for non-linear divisible set?
→ The kernel trick.
if: $w \leftarrow w + x_i y_i$ ($w = 0$ initial)
$\Rightarrow w = \sum_i c_i x_i$ ($c_i \in \{-1, 1\}$)
$\langle w, x_k \rangle = \sum_i c_i \langle x_i, x_k \rangle$
We have: $\phi \to z_i = \phi(x_i)$
$\langle \tilde{w}, z_k \rangle = \sum_i^n \tilde{c}_i \langle \phi(x_i), \phi(x_k) \rangle$
$\Rightarrow$ function $k(a, b)$, $a, b \in \mathbb{R}^d$
$\tilde{k}_{ik} = K(x_i, x_k)$ $\phi : \mathbb{R}^k \to \mathbb{R}^{od}$
$\Leftrightarrow k(a, b) = \langle \phi(a), \phi(b) \rangle$
Ex: $k(a, b) = \langle a, b \rangle$
$k(a, b) = (r \langle a, b \rangle + m)^r$ ...
$\Rightarrow$ lemma : k

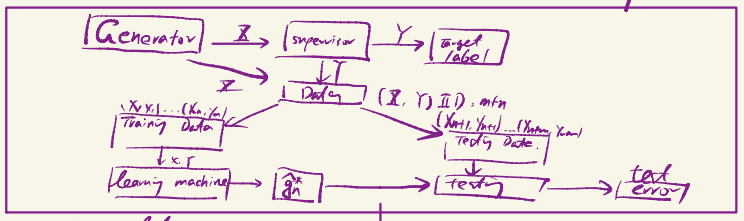
Alg 2: SVM: let $\|w\| = 1$ (if)



hinge loss: $\max\{0, 1 - \langle w, x \rangle y\}$
$= \begin{cases} \|w\| : \text{minimize!} \\ \langle w, x \rangle y_i > 1, \text{ for } i = 1 \dots n \end{cases}$
min: $\|w\|^2 + \frac{c}{n} \sum \max\{\dots\}$

margin = $\frac{y^* \langle w, x \rangle}{\|w\|}$

Training and Testing



LM : minimizes
$\hat{g}_n^* = \arg\min_{g \in \mathcal{M}} \frac{1}{n} \sum L(g, y_i)$

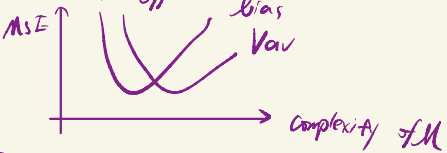
test error : $\frac{1}{m} \sum^m L(\hat{g}_n, y)$
test metric : make decision (cost fn)
Understandable (MAE, 0-1)
we find $\hat{g}_n$ fixed

mean ? posterior : once trained
prior : test error
argmin

Bias & Variance :

$\mathbb{E}[(y - \hat{g}(x))^2 | X = x] = \text{Noise} + \text{bias}^2 + \text{Variance}$

trade-off :



0 - Training error

Precision : $P(y = 1 | \hat{g} = 1)$
Recall : $P(\hat{g} = 1 | y = 1)$

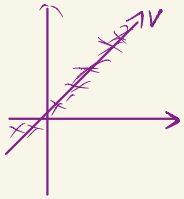
Conventions : L : 0-1 loss
$L(\hat{g}_n^*, y_n) \sim Ber(p)$

Theoretical Guarantee:
$P(|\hat{R}(\hat{g}) - R(\hat{g})| > \varepsilon) \leq 2e^{-cn\varepsilon^2}$
if: loss is bounded even if they're diff

Linear classification:
$P(R(\hat{g}) > \inf_{g \in \mathcal{M}} R(g) + \varepsilon) \leq e^{-n\varepsilon^2}(2(n+1)^d)e^{-\frac{n\varepsilon^2}{2}}$ how to draw $\mathcal{M}$
$\hat{R}_n(\hat{g}) = \min_{g \in \mathcal{M}_n} \hat{R}_n(g)$

VC theory : $\mu_n(A) = \frac{1}{n} \sum_i^n \mathbb{1}_{x_i \in A}$
$V(A) = P(x, y) \in A)$
$= P(L(y, \phi(x)) = 1)$
$P(R(\hat{g}) > \inf_{\phi \in \mathcal{M}} R(\phi) + \varepsilon) < 8 \left(\frac{ne}{k}\right)^k e^{-\frac{n\varepsilon^2}{32}}$
k : VC-dimension, complexity

Dimensionality Reduction



$y_i = x_i V$, $\|V\| = 1$

$\{x_i\}$ assume $\bar{y}_n = 0$ or $\tilde{x}_i = x_i - \bar{x}_n$

$\frac{1}{n}\sum (y_i - \bar{y}_n)^2 = \frac{1}{n}\sum (x_i V)^2$

Goal: select V such that max (Var)

$V_1 = \underset{\|V\|=1}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} (x_i V)^2$ * not unique?

→ first singular vector

$\sigma_1 = \sqrt{\sum (x_i V)^2}$ first singular value

else: Greedy algorithm:

$V_2 = \underset{\substack{\|V\|=1 \\ V_2 \perp V}}{\operatorname{argmax}} \frac{1}{n}\sum (x_i V)^2$, σ_2

$V_k = \{v_1 \cdots v_k\}$ best fit k-subplace of A one measure

$r \le \min(m, n)$

---- $V \perp V_1, V_2, \cdots V_r$ $(r < m = \dim(x_i))$

* Linear Algebra: $A = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}_{n \times m}$ $AV = \begin{pmatrix} x_1 V \\ \vdots \\ x_n V \end{pmatrix}$ $\|AV\|^2 = \sum (x_i V)^2$, $V_1 = \underset{\|V\|=1}{\operatorname{argmax}} \|AV\|^2$

once: $AV = 0$ for all $V \perp V_1 \cdots V_k$. chose a base for the null space of A

Th:

SVD

$V_1 \cdots V_m$ is a base in $\mathbb{R}^m$

$x_i = \sum_{j=1}^{m} (x_i V_j) V_j$

$\Rightarrow A = \sum_{j=1}^{m} A V_j V_j^T$, $A = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $AV_j V_j^T = \begin{pmatrix} (x_1 V_j) V_j \\ \vdots \\ x_n V_j V_j \end{pmatrix}$

$U_j = \frac{AV_j}{\sigma_j}$ if $\sigma_j > 0$.

U_j is the left singular vector

$\Rightarrow A = \sum_{j=1}^{m} A V_j V_j^T = \sum_{j=1}^{m} \sigma_i U_j V_j^T = UDV^T$ → left

$U = (u_1 | u_2 | \cdots | u_m)_{n \times m}$ $D = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{pmatrix}_{m \times m}$ $V = (V_1 | V_2 \cdots V_m)$

Low rank approx.

$A_k = \sum_{j}^{k} \sigma_j U_j V_j^T$, $\sigma_j > 0$

$\text{Rank}(A_k) = k$

→ error: $\tilde{x}_i = (A - A_k)_i$

$\|A - A_k\|_{Fro}^2 = \sum (a_{ij} - (a_k)_{ij})^2$

$= \sum_{j}^{m} \|\tilde{x}_i\|_2^2 = \sum_{j}^{m}\sum_{i}^{n} |\tilde{x}_i V_j|^2 = \sum_{j=k+1}^{m} \sigma_j^2$

$(V_1 \cdots V_m)$ all orthogonal eigenvectors

$(\lambda_1 \cdots \lambda_m)$ eigenvalues of A^TA

① $\|AV\|^2 = \langle AV, AV \rangle = \langle A^TAV, V \rangle$

$\tilde{A} = A^TA$, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m \ge 0$

$\Rightarrow V = \sum_{i}^{m} a_i V_i$ $= \sum_{i}^{m} a_i^2 \lambda_i$

$\|V\| = \sqrt{a_1^2 + \cdots + a_m^2}$

⇒ Lemma: A nxm, V_1 first singular vec. of A, σ_1: first singular value; ⇒ V_1 is eigenvector of A^TA & $\max_{\|V\|=1} |AV| = |AV_1| = \sqrt{\lambda_1} = \sigma(A)$

④ power method:

$A^TA = V D^2 V^T \Rightarrow A^T A V_i = \sigma_i^2 V_i$

Ex: $UU^T = I_m$

⑤ PCA:

$AV = VD$