MATH/COMP 562

Probability Review & Notation

PAC Learning

Ref

Mohri Ch 2

Shalev-Schwartz

2.3.1 & Ch 3

2022

Notation: Prob. vs Measure theory

Measure Theory	Prob	ML
Measu. Space $(X, \mathcal{M}, \mu)$	Sample space $(\Omega, \mathcal{B}, P)$	$\Omega = X$ no $\mathcal{B}$ mentioned $P \Rightarrow p(x)$ or $\mathcal{D}(x)$.
Measurable Set	= Event	= never. But x_i sampled i.i.d. from $p(x)$
Measurable function $f(x)$	Random Variable X	$S^m = \{x_1, \ldots, x_m\}$ dataset
Integral of f $\int_X f(x) d\mu(x)$	Expectation / Expected Value / Mean of X $\mathbb{E}[X]$	random variable: X or f
probability measure	Distribution $P(x)$ $p(x)$ $\mathcal{D}(x)$	Typical distribution $p(x)$ generates data $\hat{p}(x) = \frac{1}{m}\delta_{x_i}$ "empirical distribution"

Note As $m \to \infty$ $\hat{p}_m \to p$ in some sense. — not a function; yes as distribution.

discret r.variable X

$\Omega = \{\alpha_1, \alpha_2, \dots, \alpha_6\}$

$p(\alpha_i) = 1/6$ roll dice.

$X(\alpha_i) = i$

$S^m = \{1,1,2,2,3,3,4,5,5,5,6,6\}$

$\mathbb{E}[X] = \frac{1+2+\dots+6}{6} = 21$

$Var(X) = \dots$

$\hat{p}(i) = \begin{cases} 2/12 & i = 1,2,3,6 \\ 1/12 & i = 4 \\ 3/12 & i = 5 \end{cases}$

---

continuous r.v.

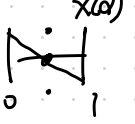
$\Omega = [0,1]$

$\alpha \in \Omega$

Borel sets, generated by intervals (a,b)

$p(S = (a,b)) = b-a$ uniform

$X(\alpha) = \frac{1}{2} - \alpha$



$\mu = \mathbb{E}[X] = \frac{1}{2}$

$Var(X) = \mathbb{E}[(X-\mu)^2]$

$= \int \alpha^2 dp(\alpha)$

$= \frac{\alpha^3}{3}\Big|_0^1 = \frac{1}{3}$

Convergence in distribution

Discrete $\hat{p}_m(x) = \frac{1}{600} \begin{cases} 110 & x=1 \\ 96 & \\ \vdots & \\ 131 & x=6 \end{cases}$ $m = 600$

Continuous $\mathcal{X} = [0,1]$ $(a,b) \subset [0,1]$

$p(x) = dx$ usual measr $p([a,b]) = b-a$

$\hat{p}_m(x) = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}$

Given (a,b) $\hat{p}_m(a,b) \to p(a,b)$. for all $(a,b) \subset [0,1]$

ML Notation

$\mathcal{X}$ domain
X r.v.

$\mathcal{X} = [0,1]^d$ vector images.

$S^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$ dataset of cat images & labels

$y_i = \begin{cases} +1 & \text{if cat pict. cute} \\ 0 & \text{o.w.} \end{cases} = y(x_i)$ x_i in database

$y(x) =$ true label of cuteness.

$\hat{p}(x) = \frac{1}{m} \delta_{x_i}$ empirical distribution.

Trained Model / function / hypothesis

f or $h = A(S^m)$

f ↑ model, h ↑ hypothesis, A ↖ training alg

$p(x) =$ prob density on $[0,1]^d$ of cat images

$f(x) \in \{0,1\}$ random variable. (because depends on S^m)

Expected loss $L(f) = \mathbb{E}(\ell_{0\text{-}1}(f, y)) = \int_{\mathcal{X}} \ell_{0\text{-}1}(f(x), y(x)) \, dp(x)$

Mohri Ch2

X domain
$Y = \{0, 1\}$

$y : X \to Y$
target unknown

$h : X \to Y$

$h \in \mathcal{H} = \{h \mid h : X \to Y\}$
hypothesis class.

$$l_{0-1}(y, y') = \begin{cases} 1 & y \neq y' \\ 0 & y = y \end{cases}$$

Generalization error $R(h) = L_{D,c}(h)$ or Risk or Test error

$$= \mathop{\mathbb{E}}_{x \sim D}\left[l_{0-1}(h(x), c(x))\right] = \mathop{\mathbb{P}}_{x \sim D}\left[h(x) \neq c(x)\right]$$

Empirical / training error

$$S = S_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \qquad y_i = y(x_i)$$

$$\hat{R}_S(h) = \hat{L}_S(h) = \frac{1}{m}\sum_{i=1}^{m} l_{0-1}(h(x_i), y(x_i))$$

$$\left(= \mathop{\mathbb{E}}_{x \sim \hat{D}}\left[l_{0-1}(h(x), y(x)\right] \right)$$

Defn: probably approximately correct (PAC) learning

how many samples

a concept class C, is PAC learnable if for every $c \in C$ there exist a learning algorithm such that, given any $\varepsilon > 0$ (error), $\delta > 0$ (tolerance) there exists $m = m(\varepsilon, \delta)$ s.t. if you have m samples (generated i.i.d. from D)

$$\mathbb{P}_{S \sim D^m}\left[L_D(h_S) \leq \varepsilon \right] \geq 1 - \delta$$

- $S \sim D^m$: over different S_m
- $L_D(h_S)$: gen loss
- ε: error
- $1 - \delta$: prob tol

Learn / measure & averages

E.G coin toss

$\varepsilon = 0.02$

$\delta = 0.05$

want $\left| \frac{\#\text{heads}}{m} - \text{prob(heads)} \right| < \varepsilon$

with prob $\geq 1 - \delta$

E.G electron survey

52% vote for X

$\pm 3\%$ (19/20)

$\varepsilon = 0.03$ $\quad \delta = 0.05$

MATH Background

MATH 1 Union Bound

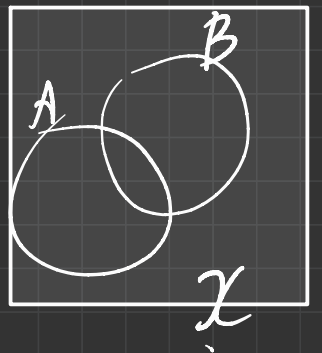
Given distribution $\mathcal{D}$ on $\mathcal{X}$

$A \subset \mathcal{X}$ $B \subset \mathcal{X}$

$p(\cdot)$

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

$$\mathcal{D}\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} \mathcal{D}(A_i)$$



MATH 2

proof

$$\textcircled{*} \quad (1-\varepsilon)^m \leq \exp(-m\varepsilon) \qquad \varepsilon \in [0,1)$$

$1 - x \leq e^{-x}$ $\quad x \in (0,1)$

$(1-x)^m \leq e^{-mx}$ $\quad x \in (0,1)$

$x \Rightarrow \varepsilon$ $\quad (1-\varepsilon)^m \leq e^{-m\varepsilon}$

$\left(e^{-x} = 1 - x + \frac{x^2}{2} \cdots \right.$

$\left. 1 - x \leq e^{-x} \right)$

$y(x) = \begin{cases} 1 & x \le t \\ -1 & o.w. \end{cases}$

IDEA of proof.

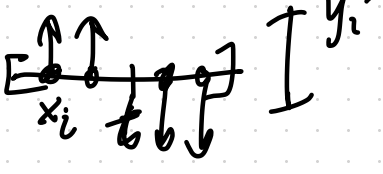
Simpler problem

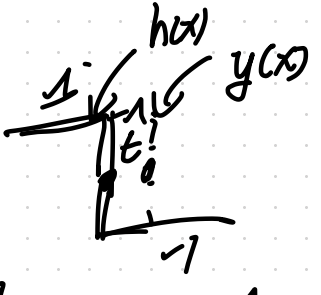
$C = \{$ intervals $[0,t] \subset [0,1]\}$

PAC learnable:

$h(x) = \begin{cases} 1 & x \le \hat{t} \\ -1 & o.w. \end{cases}$

$\hat{t} = \max(x_i \mid y_i = +1)$



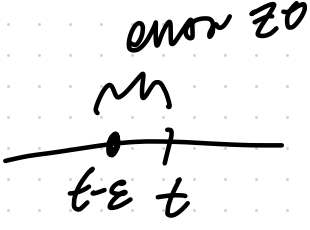


$R(h) = L(h)$

$= \mathbb{E}[l_{0-1}(h(x), y(x))] = t - \hat{t}$

Want $\mathbb{P}_{S\sim D^m}[R(h_S)\le\varepsilon]\ge 1-\delta$

Fix $\varepsilon>0$. (Assum $t>\varepsilon$ ow. fine.)



For a single sample x_i

prob miss error zone

$= 1-\varepsilon$

For m sample, prob <u>all</u> miss $= \mathbb{P}(R(h_S)>\varepsilon]$

$= (1-\varepsilon)^m$

$\le \exp(-m\varepsilon)$ by Math 2

<u>Thus</u> want $\mathbb{P}(\ )\le\delta \Rightarrow e^{-m\varepsilon}\le\delta$

$\Rightarrow -m\varepsilon \le \log\delta \Rightarrow$ $m\ge\frac{1}{\varepsilon}\log\frac{1}{\delta}$ ✓

Equivalently Gen. bound:

$$\delta = \exp(-m\varepsilon)$$

so with prob $\geq 1-\delta$

$$R(h_S) \leq \frac{1}{m} \log \frac{1}{\delta}$$

✓

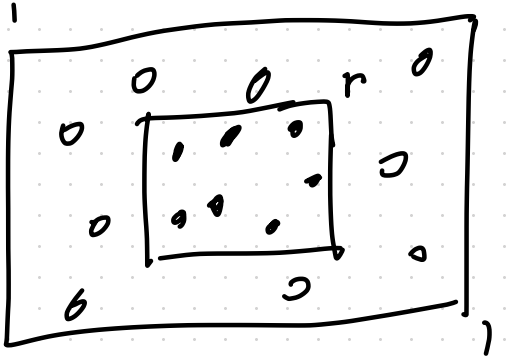
HW

EK 2.4

Learning Axis-Aligned Rectangles

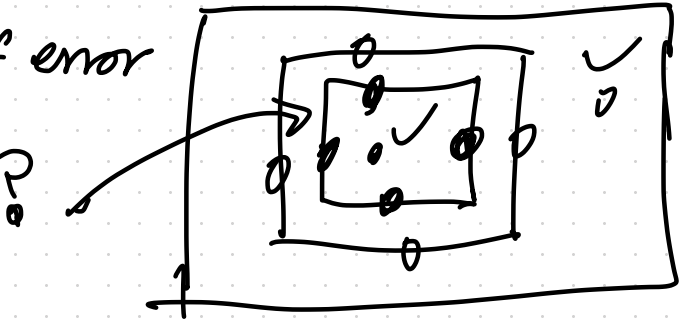
C = {all axis aligned rect}

$$y(x) = \begin{cases} 1 & x \in C \\ 0 & x \notin C \end{cases}$$



Alg Define h = smallest rectangle contains positive examples.

Window of error



prove (Math) final error of uncertain → 0 as $m \to \infty$ with high prob.