

Peer evaluation for PS1 due 10/8 in lecture (this Thursday)

- If you didn't pick up 2 "packets" in last lecture, get from me in office hours

PS2 due 10/15 (next Thursday)

Last time:

- Statistical learning framework
- Online-to-batch

P (population): prob. dist. over $X \times Y$
($Y = \{0, 1\}$)

For $f: X \rightarrow \{0, 1\}$,

$$\text{err}(f; P) = \Pr(f(x) \neq y)$$

where $(x, y) \sim P$

(*) IID training data assumption:
training data is iid sample from P

If $\hat{f}$ is produced by running learning alg on training data under Assumption (*), then

$$\text{err}(\hat{f}; P) = \Pr(\hat{f}(x) \neq y \mid \hat{f})$$

where we regard $(x, y) \sim P$ as independent of training data

ONLINE-TD-BATCH

A: online learning alg. for $\mathcal{C}$
B: batch learning alg. for $\mathcal{C}$

Helmhold & Warmuth '95

B input: training data $((x_1, y_1), \dots, (x_T, y_T)) \in (\mathcal{X} \times \{0, 1\})^T$

- Simulate A in online concept learning scenario using (x_t, y_t) in round t .
- Let f_t be the classifier used by A in round t .
- Get $f_1, \dots, f_{T+1}$

Return majority vote predictor $\hat{f}$
based on $f_1, \dots, f_{T+1}$

Theorem If A is online learner for $\mathcal{C}$ w/ M.B. M,
then B ϵ -learns $\mathcal{C}$ in expectation
with $\epsilon = \frac{2M}{T+1}$

E-learning

If # training data $T \geq \frac{2M}{\epsilon} - 1$

Simple Complexity.

then $E[\text{err}(\hat{f}; p)] \leq \epsilon$.

(ϵ, δ) PAC

If # training data $T \geq \frac{2M}{\epsilon\delta} - 1$

then with probability $\geq 1 - \delta$,
 $\text{err}(\hat{f}; p) \leq \epsilon$.

} (ϵ, δ)-PAC learning

Markov's Inequality
E-learning $\Rightarrow$ (ϵ, δ)-PAC

$$\underline{P} f((x_1, y_1), \dots, (x_T, y_T), (x_{T+1}, y_{T+1})) \sim P^{T+1}$$

S_t = first t training examples.

S_T = training data

(x_{T+1}, y_{T+1}) - not used by alg. B ,
but used in analysis

f_t - depends only on S_{t-1}
independent of $(x_t, y_t) \sim P$

$$\begin{aligned} \text{err}(f_t; P) &= \Pr(f_t(x_t) \neq y_t | f_t) \\ &= E[\mathbb{1}\{f_t(x_t) \neq y_t\} | f_t] \end{aligned}$$

$$E[\text{err}(f_t; P)] = E[\mathbb{1}\{f_t(x_t) \neq y_t\}]$$

$$\begin{aligned} \textcircled{1} \quad \sum_{t=1}^{T+1} E[\text{err}(f_t; P)] &= \sum_{t=1}^{T+1} E[\mathbb{1}\{f_t(x_t) \neq y_t\}] \\ &= E\left[\sum_{t=1}^{T+1} \mathbb{1}\{f_t(x_t) \neq y_t\}\right] \\ &\leq M. \end{aligned}$$

$$\begin{aligned} \textcircled{2} \quad \mathbb{1}\{\hat{f}(x_{T+1}) \neq y_{T+1}\} &\leq \frac{2}{T+1} \sum_{t=1}^{T+1} \mathbb{1}\{f_t(x_{T+1}) \neq y_{T+1}\} \\ E[\mathbb{1}\{\hat{f}(x_{T+1}) \neq y_{T+1}\}] &\leq \frac{2}{T+1} \sum_{t=1}^{T+1} E[\mathbb{1}\{f_t(x_{T+1}) \neq y_{T+1}\}] \\ E[\text{err}(\hat{f}; P)] &\quad \underbrace{E[\text{err}(f_t; P)]}_{\substack{\downarrow \\ \leq \frac{2M}{T+1}}} \end{aligned}$$

$\square$

Maybe better online-to-batch

1. Run alg. B k times on k independent training data sets ($k = \log_2 \frac{2}{\delta}$)

each data set i of size $\frac{8M}{\epsilon} = \frac{2M}{(\epsilon/2)(1/2)}$

enough for $(\frac{\epsilon}{2}, \frac{1}{2})$ -PAC

Get $\hat{f}^{(1)}, \dots, \hat{f}^{(k)}$

2. On yet another iid sample from P find the $\hat{f}^{(i)}$ with fewest mistakes on and return it
call this $\hat{f}$

"validation data"

with probability at least $1 - 2^{-k} = 1 - \delta/2$

at least one of the $\hat{f}^{(i)}$ has

$$\text{err}(\hat{f}^{(i)}; P) \leq \epsilon/2$$

Probability bounds on large deviation events

Let r.v. z have mean $\mu = E(z)$

What's the chance that z deviates from μ by a lot?

Markov's inequality

If z is non-negative, $t > 0$

$$\Pr(z \geq t) \leq \frac{E(z)}{t}$$

Apply Markov's inequality to the ~~magnitude~~ ^{square} of $z - \mu$

$$\Pr((z - \mu)^2 > t^2) \leq \frac{E[(z - \mu)^2]}{t^2} = \frac{\text{Var}(z)}{t^2}$$

$$\Pr(|z - \mu| > t)$$

Chebyshev's inequality

Example: $z = \sum_{i=1}^n x_i$ where $x_1, \dots, x_n$ are iid $\{0, 1\}$ -valued r.v.'s
w/ $\Pr(x_i = 1) = p$.

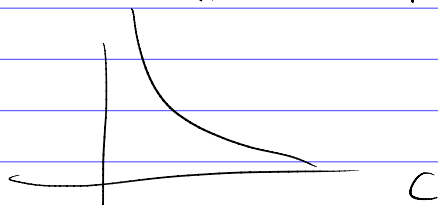
(Bernoulli(p))

($z \sim \text{Binomial}(n, p)$)

$$E(z) = np, \quad \text{Var}(z) = np(1-p) \leq n/4$$

$$\Pr(|z - np| > c\sqrt{n}) \leq \frac{n/4}{c^2 n} = \frac{1}{4c^2}$$

goes to zero at rate $1/c^2$.



This is much worse than reality ...

Chebyshev's Ineq. d/n fully exploit independence of $x_1, \dots, x_n$.

Logarithm moment generating function $\equiv$ cumulant generating function

$$K_Z(\lambda) = \ln E[\exp(\lambda z)]$$

moment generating function for z

- $K_Z(0) = 0$

- $K_Z(\lambda)$ might be infinite ..., but

If $\exists \lambda_1 < 0$ & $\lambda_2 > 0$

st. $K_Z(\lambda_1) < \infty$

$K_Z(\lambda_2) < \infty$

then • $K_Z(\lambda) < \infty \quad \forall \lambda \in [\lambda_1, \lambda_2]$

- K_Z infinitely-differentiable on $[\lambda_1, \lambda_2]$

- Graph of K_Z on $[\lambda_1, \lambda_2]$ characterizes dist. of z .

- $K'_Z(0) = E(z)$

- $K''_Z(0) = \text{var}(z)$

$$K_Z(\lambda) = \ln E[e^{\lambda z}]$$

$$K'_Z(\lambda) = \frac{1}{E[e^{\lambda z}]} E[z e^{\lambda z}] = E[z_\lambda]$$

where $z_\lambda \sim p_\lambda$

suppose $z \sim p(z)$

$$p_\lambda(z) = \frac{p(z) e^{\lambda z}}{\int p(z) e^{\lambda z} dz}$$

exponential
"tilting by λz "

$$\int p(z) e^{\lambda z} dz$$

$E[e^{\lambda z}]$

$$K''_Z(\lambda) = \text{var}(z_\lambda) \geq 0$$

$\Rightarrow K_Z$ is convex.

We'll use Markov's ineq. w/ $\exp(\lambda z)$
For $t > E(z)$, $\lambda > 0$

$$\Pr(z > t) = \Pr(e^{\lambda z} > e^{\lambda t})$$

$$\leq \frac{E[e^{\lambda z}]}{e^{\lambda t}} \quad \text{by Markov.}$$

$$= \exp(-\lambda t + \ln E[e^{\lambda z}])$$

$\therefore$

$$\Pr(z > t) \leq \exp\left(-\sup_{\lambda > 0} \{t\lambda - K_z(\lambda)\}\right)$$