

PS 2 out, due 10/15

Peer Eval for PS 1 - due 10/8

- 2 "packets" per student (who submitted PS 1 successfully)

Last time: finished Online concept learning!

Today: start Statistical learning

- Statistical framework for machine learning
- PAC criterion
- Online-to-batch conversion

STATISTICAL LEARNING

Goal: solve problems using computer programs or other automated mechanisms

- Sort a list of numbers
- Decide if a graph is connected
- Decide if an email is spam

Machine learning approach

- Get collection of labeled examples, i.e. input/output pairs, (Training data)
- Apply learning alg. to training data to get desired computer program $\hat{f}: X \rightarrow Y$
- Hope: $\hat{f}$ generalizes to cases NOT in training data
- Evaluation: Check performance of $\hat{f}$ on another collection of labeled examples (test data)

Statistical framework

Assert there's a population of input/output pairs that we care about — probability dist. over $X \times Y$

P

Suppose we have $\hat{f}: X \rightarrow Y$ (a.k.a. classifier, predictor)

$(x, y) \sim P$ (random example from P)

Joint dist. of $(x, y, \hat{f}(x))$ tells us the behavior / performance of $\hat{f}$

Simplification: just look at $\mathbb{1}\{\hat{f}(x) \neq y\}$

whose dist. is characterized by

$$E[\mathbb{1}\{\hat{f}(x) \neq y\}] = \Pr(\hat{f}(x) \neq y)$$

→ called the (population) error rate

$$\text{err}(\hat{f}; P)$$

IID = independent, identically distributed (according to P)

IID training data assumption

$$S = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)) \sim P^m$$

So $\hat{f}$ from learning alg applied to S
is a "random" function

and $\text{err}(\hat{f}; P)$ is also a random variable

$$P \quad \Pr[\hat{f}(x) \neq y | S]$$

where (x, y) is also drawn from P
independent of S .

$$((x_1, y_1), \dots, (x_m, y_m), (x, y)) \sim P^{m+1}$$

① $E[\text{err}(\hat{f}; P)]$

② 95th quantile of $\text{err}(\hat{f}; P)$:

smallest ϵ s.t.

$$\Pr(\text{err}(\hat{f}; P) \leq \epsilon) \geq 0.95$$

Learning a concept class $\mathcal{C}$

Assumption on P : " $\mathcal{C}$ -realizability assumption"

$$\exists c^* \in \mathcal{C} \text{ s.t. } \Pr(c^*(x) = y) = 1 \\ \text{where } (x, y) \sim P$$

Say learner A (ϵ, δ) -PAC learns $\mathcal{C}$ if
it gets IID training data from some P satisfying
 $\mathcal{C}$ -realizability,
and returns $\hat{f}$ s.t.

$$\Pr(\text{err}(\hat{f}; P) \leq \epsilon) \geq 1 - \delta$$

(Comes from Valiant 1984)

We'll say learner A ϵ -learns $\mathcal{C}$ in expectation
if

$$\underline{E[\text{err}(\hat{f}; P)] \leq \epsilon}$$

$$\rightarrow E[\Pr(\hat{f}(x) \neq y | \hat{f})] \leq \epsilon$$

We'll call the # training data needed to
 (ϵ, δ) -PAC learn $\mathcal{C}$

(or ϵ -learn $\mathcal{C}$ in expec.)

the "sample complexity" of the learner.

↑
could depend on $\epsilon, \delta, \mathcal{C}$

Online-to-batch conversion

Batch Learning

$S \rightarrow \text{Alg} \rightarrow \hat{f}$ learning is done!

Use $\hat{f}$ in the wild

Online learning: learning / predicting / evaluation never stops

If $\exists$ online learner for $\mathcal{C}$

does there exist a PAC learner for $\mathcal{C}$? Yes

A: online learner for $\mathcal{C}$

B: batch learner for $\mathcal{C}$

- simulate A in online scenario
- keep track of what A does -

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 has mistake bound M then B ϵ -learns in expectation

$$\text{w/ } \epsilon = \frac{2M}{T+1}$$

Corollary If A has mistake bound M then B (ϵ, δ) -PAC learn if

$$T \geq \frac{2M}{\epsilon \delta}$$

PF

Markov's ineq.

If Z is non-neg r.v. and $t > 0$

$$\Pr(Z > t) \leq \frac{E(Z)}{t}$$

$$\text{PF } E(Z) \geq E(\mathbb{1}\{Z > t\}) \cdot t$$

$$\downarrow \quad \downarrow$$
$$E(Z) \geq \Pr(Z > t) \cdot t$$

$\square$

$$E(\text{err}(\hat{f}; P)) \leq \frac{2M}{T+1} \text{ by Theorem.}$$

$$\Pr(\text{err}(\hat{f}; P) > \epsilon) \leq \frac{E(\text{err}(\hat{f}; P))}{\epsilon} \leq \frac{2M}{T+1} \cdot \frac{1}{\epsilon} \leq \delta$$

$$\text{when } T \geq \frac{2M}{\epsilon \delta}$$

$\square$