Lecture 24 11/21/2025

PAC Learning

How do we formalize machine learning? Broadly speaking, supervised learning is about coming up with an algorithm or suite of algorithms that given some initial training/labeled data, outputs a "prediction algorithm" that does well on future data that is unlabeled. The most basic but still very useful and important problem is binary classification. Given some data point $\bar{x}$, is $\bar{x}$ a Yes or No

example. Cat / not cat Cat / Dog.

+ / −.

Set up:

1. Data is formalized as vectors $\bar{x} \in \mathbb{R}^d$ or $\bar{x} \in \{0,1\}^d$

2. Labeled data: pairs $(\bar{x}_1, y_1), \ldots, (\bar{x}_n, y_n)$ where y_i is the correct label for data $\bar{x}_i$.

3. Want to obtain a prediction algorithm (could be randomized) that given a future data point $\bar{x}$ outputs $A(\bar{x})$ as the label.

Since arbitrary algorithms are complicated and hard to reason about and interpret and optimize over, traditionally the class of prediction algorithms were Simple.

The term concept class or hypothesis class is used to indicate that we are only interested in finding "prediction" algorithms from this class C.

The second important aspect is the issue of "generalization". How do we capture the fact that future data is not completely different from past data or training data?

The third issues are about amount of training data (sample complexity) and efficiency of learning.

PAC (probably approximately correct) model introduced by Valiant is a nice theoretical model that addresses the above issues.

1. Data/examples come from a set X.
2. A function $c: X \to \{0,1\}$ (we are assuming binary classification) is called a "concept".
3. $\mathcal{C}$ is a collection of concepts called a concept class.
3. $\mathcal{H}$ is a collection of "hypothesis".

Typically we will assume $\mathcal{C} \subseteq \mathcal{H}$.

4. In the full consistency model we will assume that $\exists$ some $c^* : X \to \{0,1\}$ which gives the true label for each example.

5. The examples are drawn from a distribution $\mathcal{D}$.

6. Given a hypothesis $h \in \mathcal{H}$, we define its error with respect to c^* and $\mathcal{D}$ as

$$\Pr_{x \sim \mathcal{D}} [h(x) \neq c^*(x)].$$

Defn: A pair $(\mathcal{C}, \mathcal{D})$ is PAC-learnable if $\exists$ a learning algorithm that given $m = \text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}\right)$ random samples from $\mathcal{D}$ outputs (efficiently) a hypothesis h such that $P_{\mathcal{D}}[\text{err}(h) > \varepsilon] \leq \delta$.

(i) Algorithm is allowed to output an imperfect hypothesis even when there exists a perfect concept.

(ii) Algorithm is allowed to fail with some small probability.

Defn: Given a set of correctly labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$ (here $y_i = c^*(x_i)$) a hypothesis $h: X \to \{0,1\}$ is consistent if $h(x_i) = y_i \quad \forall i \in [n]$

Suppose we have a training sample S and we are able to find a hypothesis h that is consistent with S. Will this be good hypothesis for future data?

Theorem: Let $\mathcal{H}$ be a finite hypothesis class and let $\varepsilon, \delta \in (0,1)$. Suppose S is a sample of size

$$n \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

Then if h is consistent with S then $err_D(h) \leq \varepsilon$ with prob $\geq 1-\delta$.

Proof:

.. Say $h \in H$ is bad if $err_D(h) > \varepsilon$. Let $h_1, h_2, \ldots, h_l$ be the bad hypotheses. What is the probability that h_i is consistent with S?

At most $(1-\varepsilon)^n$. By union bounds if $l(1-\varepsilon)^n < \delta$ then no bad hypothesis will be consistent $\Rightarrow$ output of algorithm will be good hypothesis.

So we need $n > \frac{1}{\varepsilon}\left(\ln l + \ln \frac{1}{\delta}\right)$

and $l \leq |\mathcal{H}|$.

□.

The previous setting required that there is a "correct" concept/hypothesis in the concept class. However, maybe we have only a "good" hypothesis h^* that makes a small error. Say $\text{err}_D(h^*)$. is the best we can hope for.

In the training phase we obtain a sample S and we find a hypothesis h s.t. $\text{err}(h,S)$ is as small as possible. Will this hypothesis generalize?

Yes.

Theorem: Let H be a finite hypothesis class and let ε and $\delta \in (0,1)$. Suppose S is a training sample from D with $n = |S|$ such that

$$n \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

Then $\forall h \in H$

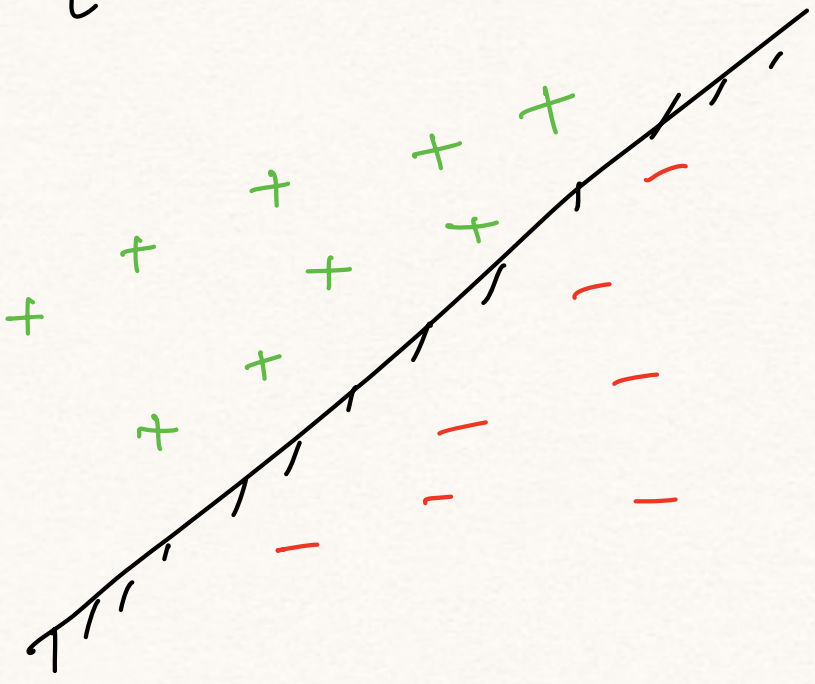
$$|err_S(h) - err_D(h)| \leq \varepsilon.$$

Proof: Needs additive Chernoff bound.

□.

VC-dimension based bounds

Consider the setting where $X \subseteq \mathbb{R}^d$ and $\mathcal{H} = \{h \mid h \text{ is a half space in } \mathbb{R}^d\}$.



In this setting $\mathcal{H}$ is infinite.

However if $|X| = n$ then there are only $n^{O(d)}$ interesting half spaces so we can use previous bounds but then we will have dependence on $\log n$ which is undesirable.

Via the ε-net and ε-sample theorems it is not hard to show the following theorem.

Theorem: Let $\mathcal{H}$ be a family of hypothesis with VC-dim $\leq d$. Let $\varepsilon, \delta \in (0,1)$. Suppose S is a sample of size $\geq \frac{C}{\varepsilon}\left(d \log \frac{d}{\varepsilon} + \log \frac{1}{\delta}\right)$. Then, with prob $\geq 1-\delta$ any consistent $h \in \mathcal{H}$ has $\text{err}_D(h) \leq \varepsilon$.

If $n \geq \frac{C}{\varepsilon^2}\left(d \log \frac{d}{\varepsilon} + \log \frac{1}{\delta}\right)$ then $\forall h \in \mathcal{H}$, $\left|\text{err}_D(h) - \text{err}_S(h)\right| \leq \varepsilon$.

Examples:

Disjunctions

$X = \{0,1\}^n$ for some n.

i.e. all bit strings of length n.

$\mathcal{C}$ = is the class of disjunctions over boolean variables $z_1, z_2, \ldots, z_n$.

For instance $z_1 + z_5 + \bar{z}_7$

$\bar{z}_1 + z_7 + \bar{z}_{10} + z_{12}$

$|\mathcal{C}| = 3^n$ why?

Claim: Given a set of examples S (x_i, y_i) $i \in [L]$ where $x_i \in \{0,1\}^n$ and $y_i \in \{0,1\}$ there is an

efficient algorithm that outputs a disjunction $c \in \mathcal{C}$ that is consistent with S if there is one.

Proof:

x_1	$(0,1,1,0,1)$	0	y_1
x_2	$(0,0,1,0,1)$	1	y_2
x_3	$(1,1,0,0,1)$	0	y_3
x_4	$(1,1,0,0,0)$	1	y_4
x_5	$(1,1,1,0,0)$	1	y_5

Start with $c = z_1 + \bar{z}_1 + \cdots z_n + \bar{z}_n$

which means that all strings are accepted.

Let S_0 be examples with $y_i = 0$

and S_1 be examples with $y_i = 1$

for each $i \in S_0$ do

  for $j = 1$ to n do

    if $X_{i,j} = 0$

      remove $\bar{z}_j$ from c.

    else

      remove z_j from c.

for each i in S_1 do

- flag = FAIL
- for $j = 1$ to n do

    if $X_{i,j} = 1$ and $z_j \in c$

      flag = DIC

      Break.

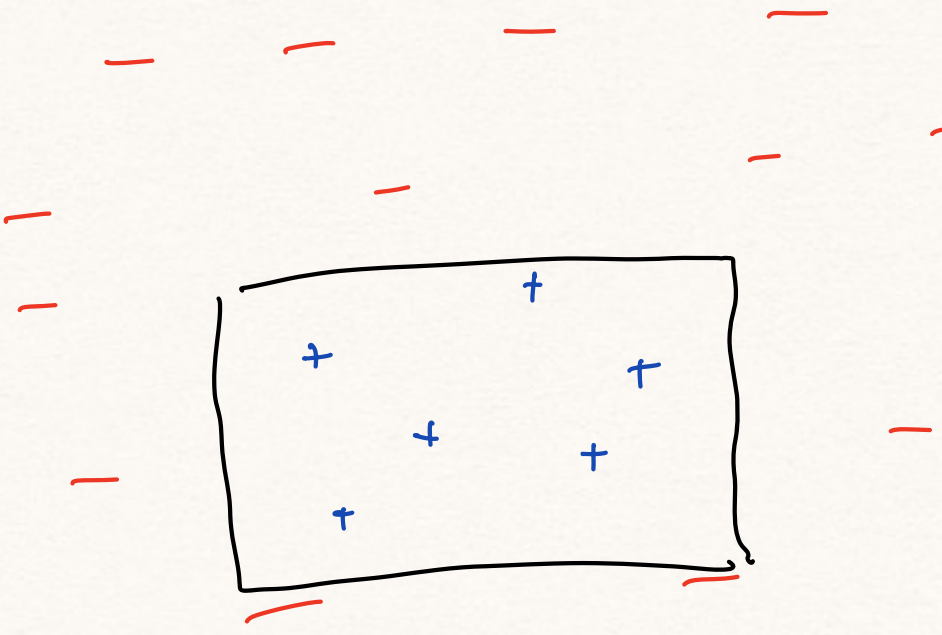
- If flag = FAIL output no consistent hypotheses.

end for.

Output c.

Example: Points in $\mathbb{R}^2$

$\mathcal{C}$ = {all axis parallel rectangles}



Checking if $\exists$ a consistent concept is simple. Find the smallest rectangle that contains all the + ve examples and see if there is any neg example.

Example: $P \subseteq \mathbb{R}^d$

$\mathcal{C} = \{h \mid h \text{ is a half space in } \mathbb{R}^d\}$

Suppose S is a set of samples and S_0 is -ve examples $y_i = 0$ and S_1 is +ve examples $y_i = 1$

Can write an LP to find $a_1, a_2, \ldots, a_d, b$

$$\sum_{j=1}^{d} a_j x_{ij} - b > 0 \quad \text{if } i \in S_1$$

$$\sum_{j=1}^{d} a_j x_{ij} - b \leq 0 \quad \text{if } i \in S_0.$$

LP is feasible iff $\exists$ half space $a\bar{x} - b = 0$ that

Separates +ve from -ve examples.