CS 4540/CS 8803: Algorithmic Theory of Intelligence Fall 2023

Lecture 3: PAC Learning

Instructor: Santosh Vempala *Lecture date: 9/06,9/11,9/13*

1 Learning the Class of Conjunctions

Consider data of the form $\{x_1, \ldots, x_n\} \in \{0,1\}^n$. A *conjunction* is a function on some subset of the variables $\{x_1, x_2 \ldots, x_n, \bar{x}_1, \bar{x}_2 \ldots, \bar{x}_n\}$, where $\bar{x}_i = 1 - x_i$. It maps x to 1 if every variable in the subset is 1, and 0 otherwise. For example:

$$\begin{aligned} h(x) &= x_1 \wedge x_2 \wedge \bar{x}_5 \\ h(x) &= \bar{x}_3 \wedge \bar{x}_4 \\ h(x) &= x_1 \wedge x_2 \wedge \cdots \wedge x_r \end{aligned}$$

The size of the hypothesis class is 3^n; for each i, either x_i is in the conjunction, $\bar{x}_i$ is in the conjunction, or both are absent.

Any such conjunction can be represented as a linear halfspace, so the tools for learning halfspaces seen in previous lectures apply without modification. Simply sum the variables in the conjunction, and set the threshold to be the number of variables. Using the examples above:

$$\begin{aligned} h(x) &= (x_1 + x_2 + 1 - x_5 \geq 3) \\ h(x) &= (1 - x_3 + 1 - x_4 \geq 2) \\ h(x) &= (x_1 + x_2 + \cdots + x_r \geq r) \end{aligned}$$

If $x_i \in \{0,1\}$, these two ways of writing the function are equivalent.

1.1 Learn-Conj

Let $\mathcal{D}$ be a distribution over $\{0,1\}^n$. Assume that each example is chosen independently from this distribution. Consider the following algorithm for learning an unknown conjunction.

Algorithm 1 LearnConj

Maintain two sets of indices $S = \{x_1, x_2 \ldots, x_n\}$,$\bar{S} = \{\bar{x}_1, \bar{x}_2, \bar{x}_n\}$
for each input $x^{(i)}$ **do**
  if $l(x^{(i)}) = 1$ **then**
    For each $j \in [n]$, remove $\bar{x}_j$ from $\bar{S}$ if $x_j^{(i)} = 1$, and remove x_j from S if $x^{(i)} = 0$
  else
    Do nothing
  end if
end for
Output the conjunction of the remaining literals in S and $\bar{S}$.

At any point, the hypothesis $h(x) = \left(\bigwedge_{i \in S} x_j\right) \wedge \left(\bigwedge_{i \in \bar{S}} \bar{x}_j\right)$ is consistent with all examples seen so far.

Theorem 1. *With $m \geq \frac{1}{\epsilon}(n \log 3 + \log \frac{1}{\delta})$ examples, with probability $1 - \delta$,* LearnConj *will be correct on a new example from $\mathcal{D}$ with probability $1 - \epsilon$.*

Proof. Let $h \in H$ be a hypothesis with $Pr_{x \sim \mathcal{D}}(h(x) \neq l(x)) > \epsilon$. The probability that h is not eliminated after m examples is at most $(1 - \epsilon)^m$.

The probability that any such 'bad h' survives after m examples is at most $3^n(1 - \epsilon)^m$.

Setting $3^n(1 - \epsilon)^m < \delta$ and taking the log of both sides, we see that $m \geq \frac{1}{\epsilon}(n \log 3 + \log \frac{1}{\delta})$ suffices.

□

1.2 Learning Decision Lists

A *decision list* is a generalization of the set of conjunctions. It consists of a series of if-else statements, returning true/false at each step. E.g.,

If $x_1 = 1$, return False; else if $x_4 = 0$, return True; else, return False.

Any conjunction can be written as a decision list. E.g., $x_1 \wedge x_2 \wedge \cdots \wedge x_r$ is equivalent to:

If $x_1 = 0$, return False; else if $x_2 = 0$, return False; ... else if $x_r = 0$, return False; else, return True

The number of possible decision lists is at most $4^n n!$ (choosing 0/1 for the check and return statement, and every possible ordering of the indices).

Suppose we have an algorithm LearnDL that maintains a decision list h which is consistent with all examples seen so far. Then, the following theorem applies:

Theorem 2. *With $m \geq \frac{c}{\epsilon}(n \log n + \log \frac{1}{\delta})$ examples, with probability $1-\delta$,* LearnDL *will be correct on a new example from $\mathcal{D}$ with probability $1 - \epsilon$.*

Proof. The argument from Theorem 1 does not use any special properties of conjunctions or the algorithm. We can apply the same logic here.

Let $h \in H$ be a hypothesis with $Pr_{x \sim \mathcal{D}}(h(x) \neq l(x)) > \epsilon$. The probability that h is not eliminated after m examples is at most $(1-\epsilon)^m$.

The probability that any such 'bad h' survives after m examples is at most $4^n n!(1-\epsilon)^m$. (This time, substituting the size of the class of decision lists!)

Setting $4^n n!(1-\epsilon)^m < \delta$ and taking the log of both sides, we see that $m \geq O\left(\frac{1}{\epsilon}\right)(n \log n + \log \frac{1}{\delta})$ suffices. □

In general, for an algorithm that learns a consistent hypothesis from a class H, $m \geq \frac{c}{\epsilon}(\log |H| + \log \frac{1}{\delta})$ examples are sufficient to achieve this guarantee.

2 (ϵ, δ)-PAC-Learning

Definition 3. *Given iid samples from an unknown distribution $\mathcal{D}$, labeled by a hypothesis ℓ from a hypothesis class H, an algorithm $\mathcal{A}$ (ϵ, δ)-PAC-learns the hypothesis class H if, with probability $1-\delta$, it produces a hypothesis $h \in H$ that correctly classifies a random sample from $\mathcal{D}$ with probability at least $1 - \epsilon$, i.e.,*

$$Pr_{x \in \mathcal{D}}(h(x) = \ell(x)) \geq 1 - \epsilon.$$

PAC stands for Probably Approximately Correct. Intuitively, it states that the algorithm $\mathcal{A}$ is *probably* successful (producing a good h with probability $1 - \delta$). A good h is *approximately correct* (classifying samples from $\mathcal{D}$ correctly with probability $1 - \epsilon$).

Note that this definition says nothing about the efficiency of the algorithm. Usually the efficiency is judged by time or sample complexity (i.e., the time it takes to process a sample, and the number of samples that it needs).

2.1 PAC-learning Halfspaces

In section 1, we showed a heuristic for learning an arbitrary hypothesis class H. However, the number of samples needed depends on $\log |H|$. For the case of halfspaces, H is infinite, so we must use another method.

Suppose, $\|w^*\|_2 \leq 1$, $\max_x \|x\|_2 \leq 1$, and $\max_x |\langle w^*, x \rangle| \geq \gamma$. At any time, define the set of consistent hypothesis.

$$W = \{w : \|w\|_2 \leq 1, \forall i \langle w, l(x^{(i)}) x^{(i)} \rangle \geq \gamma\}$$

For the next sample $x^{(i+1)}$, divide W into two cases:

$$W_+ = \{w \in W : \langle w, x^{(i+1)} \rangle \geq 0\}$$

$$W_- = \{w \in W : \langle w, x^{(i+1)} \rangle < 0\}$$

By predicting the set with the larger volume, we guarantee that either we are correct, or the size of W is reduced by at least half.

Note that because the margin is gamma, we only need to reduce W to within a 'cap' of angle γ. The set is

$$V_\gamma = \{w : \|w\|_2 = 1, \langle w, w^* \rangle \geq 1 - \gamma^2\}$$

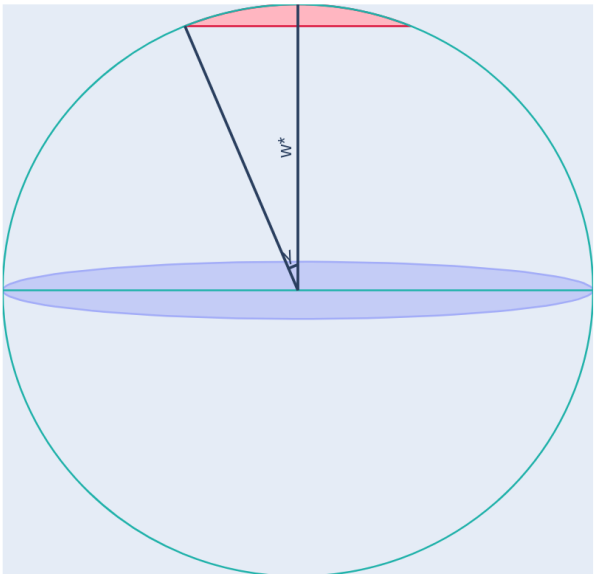

Figure 1: The set of points on the sphere where $\langle w, w^* \rangle \geq 1 - \gamma^2$

This set is illustrated in red in Figure 1. All vectors in V_γ will correctly classify examples that obey the margin.

Theorem 4. *The number of steps required to find a consistent classifier is* $O(n \log \gamma)$.

Proof. The volume of W is reduced by at least half at each step; the initial set is $S_{n-1} = \{w : \|w\| = 1\}$. So, we need to bound $\log_2 \frac{\text{Vol}(S_{n-1})}{\text{Vol}(V_\gamma)}$.

To do this, we can write both sets as an integral over the first coordinate. Note that when we fix $x_1 = t$, then by the equation for the sphere:

$$\sum_{i=1}^{n} x_i^2 = 1 \rightarrow \sum_{i=2}^{n} x_i^2 = 1 - t^2$$

Hence, the cap formed by intersecting the sphere with the plane $x_1 = t$ is a sphere of radius $1 - t^2$

$$\text{vol}(S_{n-1}) = 2 \int_0^1 (1 - t^2)^{(n-2)/2} \text{Vol}(S_{n-2})\, dt$$

$$\text{vol}(V_\gamma) = \int_{\sqrt{1-\gamma^2}}^1 (1 - t^2)^{(n-2)/2} \text{Vol}(S_{n-2})\, dt$$

This integral is not straight-forward to evaluate; fortunately, we only need an asymptotic bound, so we can make some simplifications.

Simplifying this integral

$$\begin{aligned}\int_{\sqrt{1-\gamma^2}}^{1}(1-t^2)^{(n-2)/2}\,dt &\geq \int_{\sqrt{1-\gamma^2}}^{1}\left(\gamma^2\right)^{(n-2)/2}\,dt \\ &= (1-\sqrt{1-\gamma^2})\gamma^{n-2}\,dt \\ &= c^n\gamma^n \text{ for a constant } c\end{aligned}$$

Since $\int_0^1(1-t^2)^{(n-2)/2}\,dt \leq 1$, the ratio is

$$\frac{\text{Vol}(S_{n-2})\int_{\cos(\gamma)}^{1}(1-t^2)^{(n-2)/2}\,dt}{2\text{Vol}(S_{n-2})\int_0^1(1-t^2)^{(n-2)/2}\,dt} \geq c^n\gamma^n$$

Taking the logarithm gives us the theorem statement. □

3 Ways to label points

Definition 5. *For a concept class $\mathcal{H}$ and an integer $m > 0$, let $\mathcal{H}(m)$ be the maximum number of distinct ways a set of m points can be labelled using concepts in $\mathcal{H}$*

Example: Intervals in 1-D
Let $\mathcal{H} = \{[a,b] : a,b \in \mathbb{R}, a < b\}$; the set of all intervals on the real line.

As illustrated in Figure 2, $\mathcal{H}(2) = 4$, since the two points shown in the plot can be labelled in all four ways. However, $\mathcal{H}(3) = 7$, since there is a sequence of labels $(+-+)$ which cannot be achieved with a single interval (but all other labellings are possible!)

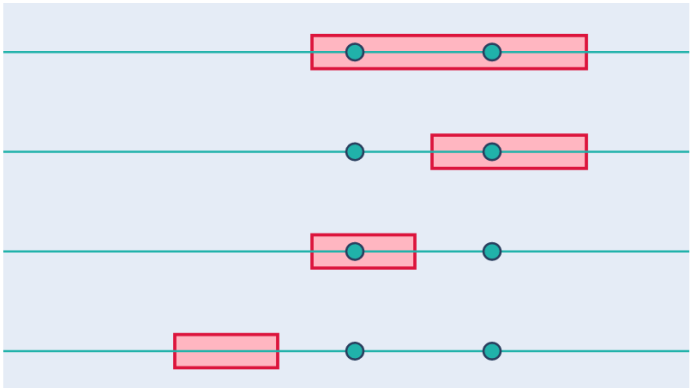

Figure 2: The four ways to label a set of points on the real-line with a single interval

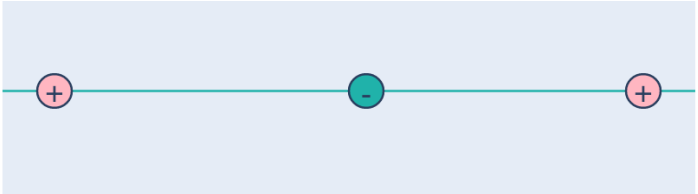

Figure 3: A set of three points with an assignment of labels that can't be achieved by a single interval

Rectangles in 2D
Let $\mathcal{H} = \{[a,b] \times [c,d] : a,b,c,d \in \mathbb{R}, a < b, c < d\}$; the set of all rectangles in $\mathbb{R}^2$.

We have $\mathcal{H}(m) \leq (m+1)^4$. There are $(m+1)^2$ ways to pick the vertical lines, and $(m+1)^2$ ways to pick the horizontal lines.

Halfspaces in $\mathbb{R}^n$
Note that in $\mathbb{R}^2$, defining a separating line requires 2 points. This generalizes to $\mathbb{R}^n$, where it takes n points to define a separating hyperplane. We can choose our labelling by picking n points out of the sample to define the plane.

This gives:

$$\mathcal{H}(m) \leq \binom{m}{n} * 2 \leq m^n \tag{1}$$

3.1 Theorem: Sample Complexity of PAC-Learning

Theorem 6. *Suppose we have a true labeling function $h \in \mathcal{H}$. For (ϵ, δ)-PAC-Learning, we can output any $\tilde{h} \in \mathcal{H}$ that is consistent with m iid examples from the unknown distribution provided that:*

$$m \geq \frac{c}{\epsilon}\left(\log \mathcal{H}(2m) + \log \frac{1}{\delta}\right)$$

Corollary 7. *The sample complexity of (ϵ, δ)-PAC-Learning halfspaces in $\mathbb{R}^n$ is $\frac{c}{\epsilon}\left(n \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)$.*

Proof. To argue this, we can substitute the bound given in Equation 1 into the theorem statement.

We can use the following inequality for binomial coefficients: $\binom{2m}{n} \leq \left(\frac{2me}{n}\right)^n$. Substituting:

$$m \geq \frac{c}{\epsilon}\left(n \log \frac{2m * e}{n} + \log \frac{1}{\delta}\right)$$

To find a value of m that obeys this inequality, we can replace the m on the right hand side with something larger than m.

From the above equation, $m < \frac{c}{\epsilon}(n(\log \frac{n^{10}}{\epsilon}) + \log \frac{1}{\delta})$ (for example). Substituting this for m:

$$\frac{c}{\epsilon}\left(n \log\left(\frac{c}{\epsilon n}(n \log \frac{n^{10}}{\epsilon} + \log \frac{1}{\delta})\right) + \log \frac{1}{\delta}\right) = \frac{c}{\epsilon}\left(n(\log \frac{1}{\epsilon})(1 + o(1)) + \log \frac{1}{\delta}\right)$$

□

Proof of Theorem 6. Let A be the following event:

A: Given m i.i.d. samples from D, there exists an $h \in \mathcal{H}$ which is correct on all samples, but $Pr_D(h(x) \neq l(x)) > \epsilon$.

To achieve $(\epsilon, \delta)-$PAC learning, we want $Pr(A) < \delta$. To bound $Pr(A)$, we will define a different event.

Draw $2m$ samples from D, and divide them into two sets S_1, S_2, with $|S_1| = |S_2| = m$. Let B be the following event.

B: There exists an $h \in \mathcal{H}$ such that h makes no errors on S_1, but h makes at least $\frac{\epsilon m}{2}$ errors on S_2.

Claim: $Pr(B) = Pr(A)Pr(B \mid A) > \frac{1}{2}Pr(A)$.
Proof: Assume that event A is true; i.e., , $Pr_{x \sim D}(h(x) \neq l(x)) > \epsilon$. Since each $x \in S_2$ is independently from D, the expected number of mistakes on S_2 is at least ϵm.

Let M_{S2} be the number of mistakes h makes on S_2. So, $\mathrm{E}[M_{S2}] > \epsilon m$

Event B occurs if $M_{S2} > \epsilon m/2$. Let's bound the probability that B *does not* occur using the Chernoff bound.

$$Pr\left(M_{S2} < \frac{\epsilon m}{2}\right) < Pr\left(M_{S2} < \frac{1}{2}\,\mathrm{E}[M_{S2}]\right) \leq e^{-\frac{\epsilon m}{8}}$$

This is less than $1/2$ when $m > 8/\epsilon$.

So, $Pr(B \mid A) > \frac{1}{2}$, giving the claim.

Claim: $Pr(B) \leq \mathcal{H}(2m)2^{-\epsilon m/2}$

Proof: Fix $2m$ points, their true labels, and a hypothesis h which makes at least $\epsilon m/2$ errors on this set. Randomly assign the points to S_1 and S_2; i.e., $S_1 = \{a_1, \ldots, a_m\}$, $S_2 = \{b_1, \ldots, b_m\}$.

For B to hold, each point where h makes a mistake must be assigned to S_2. The probability of this is at most $2^{-\epsilon m/2}$, since each sample has a $1/2$ probability of being assigned to S_2.

The number of possible labelings of $2m$ points is $\mathcal{H}(2m)$, by definition. Thus, taking the union bound over all possible labels, we have:

$$Pr(B) \leq \mathcal{H}(2m)2^{-\epsilon m/2}$$

This proves the claim. Now, we have $Pr(A) < 2Pr(B)$, and $Pr(B) \leq \mathcal{H}(2m)2^{-\epsilon m/2}$. Setting $\mathcal{H}(2m)2^{-\epsilon m/2} < \delta/2$ and taking the logarithm of both sides, $\log \mathcal{H}(2m) - \frac{\epsilon m}{2} < \log(\delta/2)$. Solving for m:

$$m > \frac{2}{\epsilon}\left(\log \mathcal{H}(2m) + \log \frac{2}{\delta}\right)$$

For this m, $Pr(A) < \delta$. This proves the theorem.

□

3.2 (ϵ, δ)-PAC Agnostic Learning

Previously, we have assumed that there exists an h which is consistent on all examples from D. In practice, this is often an unrealistic assumption. In this section, we define a notion of PAC-learning in the case where every hypothesis in $\mathcal{H}$ has some error.

Definition 8. *Let $\mathcal{H}$ be a hypothesis class, and let $\epsilon_0 = \inf_{h \in \mathcal{H}} Pr_D(h(x) \neq \ell(x))$ be the minimum achievable error within $\mathcal{H}$.*

Given independent samples from D, an algorithm $\mathcal{A}$ (ϵ, δ)-PAC agnostic learns the hypothesis class H if, with probability $1 - \delta$, it produces a hypothesis $h \in H$ that correctly classifies a random sample from $\mathcal{D}$ with probability at most ϵ more than the minimum achievable error, i.e.,

$$Pr_{x \in \mathcal{D}}(h(x) \neq \ell(x)) \leq \epsilon_0 + \epsilon.$$

The proof of sample complexity for this case is very similar to the non-agnostic case. Let $err_D(h)$ denote the fraction of D on which h makes a mistake (disagrees from a fixed labeling function) and for a subset of labeled examples S, let $err_S(h)$ denote the fraction of S on which h differs from the fixed target. We now bound the sample complexity of simultaneously estimating the error of every hypothesis in $\mathcal{H}$.

Theorem 9. *On a sample S of size m drawn iid from an unknown distribution D, we have that with probability at least $1 - \delta$, for every $h \in \mathcal{H}$,*

$$|err_S(h) - err_D(h)| \leq \epsilon$$

provided

$$m \geq \frac{c}{\epsilon^2}\left(\log \mathcal{H}(2m) + \log \frac{1}{\delta}\right)$$

The theorem says that with a sample of size m satisfying the above lower bound, we can estimate the error of *every* hypothesis in $\mathcal{H}$, and hence we can output any low error hypothesis on m points to guarantee low-error (at most ϵ more) on the full distribution. Note the difference from Theorem 6: previously the bound was proportional to $1/\epsilon$, now it is $1/\epsilon^2$.

Proof of Theorem 9. Let A be the following event:

A: Given m i.i.d. samples from D, there exists an $h \in \mathcal{H}$ which has error at most ϵ_0 on m samples, but $Pr_D(h(x) \neq l(x)) > \epsilon + \epsilon_0$.

To prove the theorem, we want to show that $Pr(A) < \delta$. To bound $Pr(A)$, we will define a different event.

Draw $2m$ samples from D, and divide them into two sets S_1, S_2, with $|S_1| = |S_2| = m$. Let B be the following event.

B: There exists an $h \in \mathcal{H}$ such that h makes at most $\epsilon_0 m$ errors on S_1, but h makes at least $\frac{\epsilon m}{2} + \epsilon_0 m$ errors on S_2.

Claim: $Pr(B) = Pr(A)Pr(B \mid A) > \frac{1}{2}Pr(A)$.
Proof: Analogous to Theorem 6
Claim: $Pr(B) \leq \mathcal{H}(2m)e^{\epsilon^2 m/8}$
Proof: Fix $2m$ points, their true labels, and a hypothesis h which makes at least $2\epsilon_0 m + \epsilon m/2$ errors on this set. Randomly assign the points to S_1 and S_2; i.e., $S_1 = \{a_1, \ldots, a_m\}$, $S_2 = \{b_1, \ldots, b_m\}$.

For B to hold, the number of mistakes assigned to S_1 must be less than the number of mistakes assigned to S_2 by $\epsilon m/2$. The probability that this occurs can be bounded by Hoeffding's inequality, by $e^{\epsilon^2 m/8}$.

The number of possible labelings of $2m$ points is $\mathcal{H}(2m)$, by definition. Thus, taking the union bound over all possible labels, we have:

$$Pr(B) \leq \mathcal{H}(2m)e^{-\epsilon^2 m/8}$$

This proves the claim. Now, we have $Pr(A) < 2Pr(B)$, and $Pr(B) \leq \mathcal{H}(2m)2^{-\epsilon m/2}$. We can set $\mathcal{H}(2m)e^{-\epsilon^2 m/2} < \delta/2$, solve for m to achieve the bound in the theorem statement.

□

4 Useful Concentration Inequalities

In the above proofs, we used two concentration inequalities for the sums of independent, bounded random variables. We state them here:

Theorem 10 (Chernoff Bound)**.** *Let $X = \sum_{i=1}^{n} X_i$ be a sum of independent random indicators ($X_i \in \{0, 1\}$) with mean μ. Then, for any $\delta \geq 0$:*

$$\mathbb{P}(X > (1+\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2+\delta}\right)$$

$$\mathbb{P}(X < (1-\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2}\right)$$

Theorem 11 (Hoeffding's Inequality)**.** *Let $X = \sum_{i=1}^{n} X_i$ be a sum of independent bounded random variables ($a_i \leq X_i \leq b_i$) with mean μ. Then, for any $t \geq 0$:*

$$\mathbb{P}(|X - \mu| > t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$