PAC Learning:

PAC = Probably Approximately Correct

Definition (PAC): A function class $\mathcal{H}$ is called PAC learnable if for every $\epsilon, \delta \in (0,1)$, there exists a number $n_0(\epsilon, \delta)$ such that the following holds:

There exist a learning algorithm that for any distribution D over $X \times Y$, by using a set S of $n_0(\epsilon, \delta)$ data points, we obtain a predictor function h_S^* such that with probability $1-\delta$ we have

$$\min_{h \in \mathcal{H}} L_D(h) \leq L_D(h_S^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

$\mathcal{H} \longrightarrow \begin{cases} \text{Algorithm} \rightarrow h_S^* \\ n_0(\epsilon, \delta) \end{cases} \Rightarrow \mathcal{H}$ is PAC learnable

---

Let's start with the simplest possible hypothesis class and see what PAC-learning means with that class.

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$$

$\min_{h \in \mathcal{H}} L_D(h)$ $\longrightarrow$ $\min_{h \in \mathcal{H}} L_S(h)$

$\Downarrow$

h_S^* is one of the functions in $\mathcal{H}$ with smallest empirical error.

$\min_{h \in \mathcal{H}} L_S(h) = \min\{L_S(h_1), \ldots, L_S(h_m)\}$

What we'd like to find in this case is a number $n_0(\epsilon, \delta)$ such that for any distribution D over data we have:

with probability $1-\delta$:

$$L_D(h_S^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon$$

In other words we're asking how large the number of training data points should be s.t. we have w.p. $1-\delta$ that $L_D(h_S^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon$.

In order to find $n_0(\epsilon, \delta)$ we need to answer two main questions.

(1) Given a fixed function $h: X \to Y$, how many traing data point should we have such that:

with probability $1-\delta$:

$$|L_S(h) - L_D(h)| < \varepsilon$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x_i) \neq y_i\} - E_D[\mathbb{1}\{h(x) \neq y\}] \right| < \epsilon$$

what should be $|S| = n_0(\epsilon, \delta)$?

To answer this question we'll use a very important probabilistic tool which is call the Hoeffding's inequality. This inequality has very useful in a variety of applications in data science (both in terms of theory

and algorithm design).

Hoeffding's Inequality:

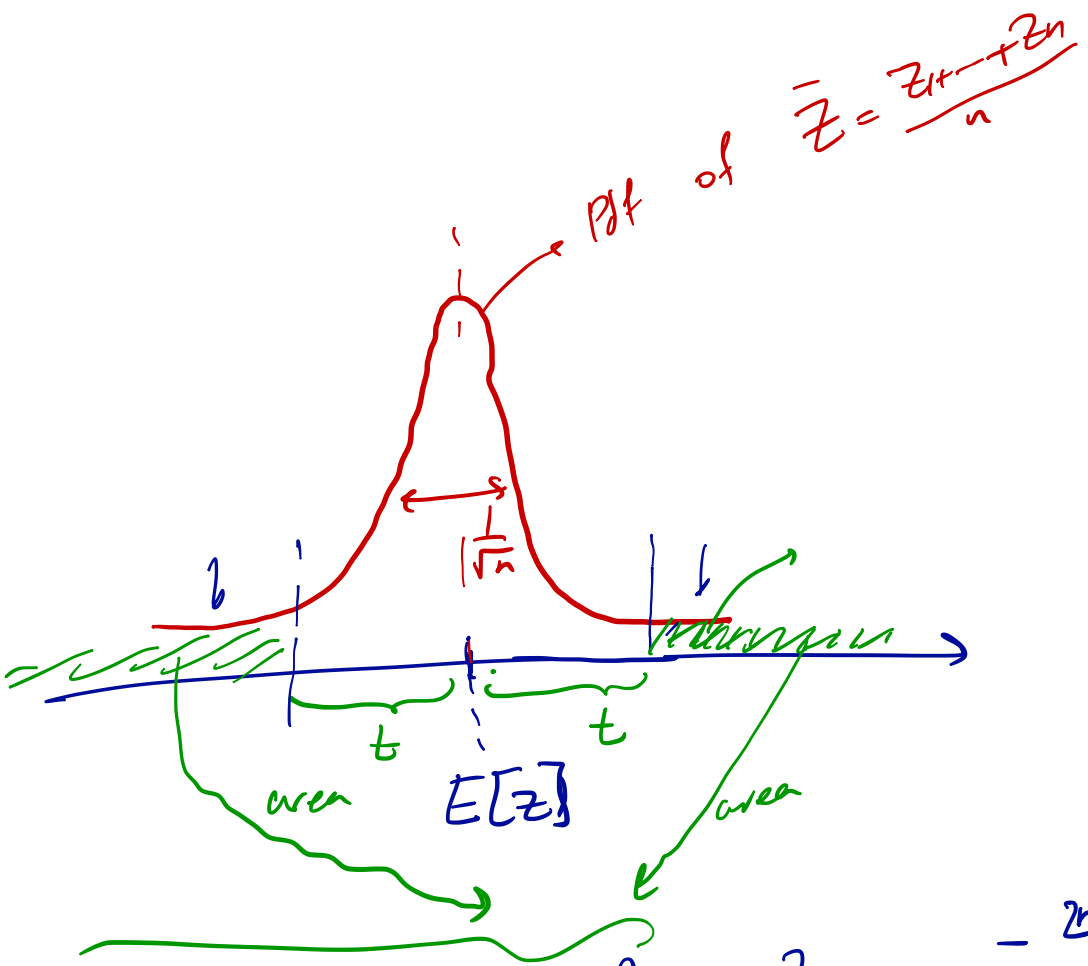
Let $Z_1, Z_2, \ldots, Z_n$ be n iid random variables that take value in the unit interval (i.e. $Z_i \in [0,1]$).

Then:

$$\Pr\left\{ \left| \frac{\sum_{i=1}^{n} Z_i}{n} - E[Z_i] \right| \geq t \right\} \leq 2e^{-2nt^2}$$

$Z_1, Z_2, \ldots, Z_n$ iid $Z \longrightarrow \sigma^2 = \mathrm{Var}(Z)$

$$\frac{Z_1 + Z_2 + \cdots + Z_n}{n} \sim E[Z] + \frac{1}{\sqrt{n}} N(0, \sigma^2)$$



$$Pr\{|\bar{Z} - E[Z]| > t\} \leq 2e^{-2nt^2}$$

e.g. $n = 1000$, $t = 0.5$

$$e^{-2 \cdot 1000 \cdot (0.5)^2} = e^{-500} \longrightarrow 0$$

$H \longrightarrow$ PAC learnable:

(1) $n_0(\epsilon, \delta)$

(2) learning algorithm that takes as input a set of training data points S, and outputs h_S^*.

for any data distribution D, as long as $|S| \geq n_0(\epsilon, \delta)$ then with probability $1-\delta$ we have:

$$L_D(h_S^*) \leq \min_{h \in H} L_D(h) + \epsilon$$

Example:

$\mathcal{H} = \{h_1, \dots, h_m\}$

(1) specify what $n_0(\epsilon, \delta)$ is.

(2) Algorithm: ERM $\mathcal{H}$:

$$h_S^* = \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

what is $n_0(\epsilon, \delta)$, s.t.

w.p. $1-\delta$

$$L_D(h_S^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon$$

Two questions:

(1) Given a fixed function h, how many data points, S, do we need such that

w.p. $1-\delta$

$$|L_S(h) - L_D(h)| < \varepsilon$$

$$L_S(h) = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbb{1}\{h(x_i) \neq y_i\}$$

$\Updownarrow$ equivalent

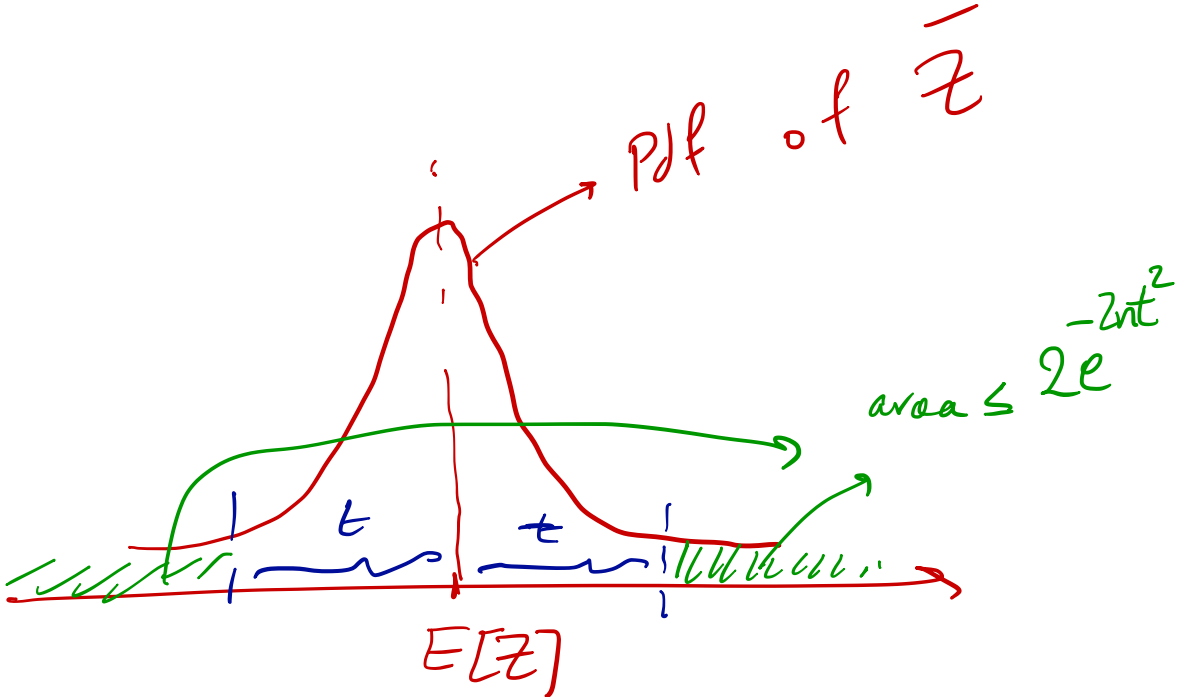
$$\Pr\{|L_S(h) - L_D(h)| \geq \varepsilon\} \leq \delta$$

Recall that Hoeffding's inequality was given as follows:

For any iid random variables $Z_1, Z_2, \dots, Z_n$, s.t. $Z_i \in [0,1]$,

we have

$$\Pr\left\{ \left| \overline{\frac{1}{n}\sum_{i=1}^{n} Z_i} - \overline{E[Z]} \right| > t \right\} \leq 2e^{-2nt^2}$$



$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}\{h(x_i) \neq y_i\}}_{Z_i}$$

$(|S| = n)$

$$L_D(h) = \underbrace{\mathop{E}_{(x,y)\sim D}\left[\mathbb{1}\{h(x) \neq y\}\right]}_{E[Z]}$$

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} Z_i = \bar{Z}$$

$$Z_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{o.w.} \end{cases} \rightarrow Z_i \in [0,1]$$

$$E[Z_i] = \mathop{E}_{(x_i,y_i)\sim D}\left[\mathbb{1}\{h(x_i) \neq y_i\}\right]$$

$$= L_D(h)$$

Hoeffding:

$$P_r\left\{\left|\underbrace{L_S(h)}_{\frac{1}{n}\sum Z_i} - \underbrace{L_D(h)}_{E[Z]}\right| > \epsilon\right\} \quad \text{for any } \epsilon > 0 \qquad (1)$$

$$\leq 2e^{-2n\epsilon^2}$$

Recall that we're looking for the smallest n s.t. (2)

$$P_r\{|L_S(h) - L_D(h)| > \epsilon\} \leq \delta$$

using (1), and in order to guarantee (2), we can let:

$$2e^{-2n\epsilon^2} \leq \delta \qquad \left(\log\frac{2}{\delta} = -\log\frac{\delta}{2}\right)$$

$$\Rightarrow \boxed{n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}}$$

So the final statement is:

Given any ϵ, δ, as long as the number of training data points is larger than $n_1 = \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$, we have

$$\Pr_{(x,y)\sim D}\left\{ |L_S(h) - L_D(h)| > \epsilon \right\} \leq \delta.$$

(for a fixed function h)

(2) Let's now assume that we have m functions $h_1, h_2, \ldots, h_m$. What is the smallest value $n_0(\epsilon, \delta)$ such that

with probability $1-\delta$: (3)

$$\forall\, i \in \{1, \ldots, m\} \;:\; |L_S(h_i) - L_D(h_i)| < \epsilon .$$

To answer this question, we're going to write an equivalent formulation of relation (3):

Let A_i be the event that

$$|L_S(h_i) - L_D(h_i)| > \epsilon .$$

Then (3) is equivalent to:

$$Pr\{A_1 \cup A_2 \cup A_3 \cdots \cup A_m\} \leq \delta . \quad (4)$$

Remember that we are searching for the smallest value of n such that (4) holds.

$$Pr\{A_1 \cup A_2 \cup \cdots \cup A_m\} \leq \delta$$

To answer this, we're going to use the Union bound:

$$Pr\{A \cup B\} \leq Pr\{A\} + Pr\{B\}$$

$$Pr\{A_1 \cup A_2 \cup \cdots \cup A_m\} \leq Pr\{A_1\} + Pr\{A_2\} + \cdots + Pr\{A_m\}.$$

---

bad event #i : A_i :

$$|L_S(h_i) - L_D(h_i)| > \epsilon$$

We'd like to make sure that none of the bad events, A_i, would take place.

So in order to guarantee

$$Pr\{A_1 \cup A_2 \cup \cdots \cup A_m\} \leq \delta \quad (5)$$

it is sufficient to guarantee that

$$Pr\{A_1\} + Pr\{A_2\} + \cdots + Pr\{A_m\} \leq \delta \quad (6)$$

Note that if (6) holds, then (5) would also hold as a result of the union bound.

Now, to guarantee (6), it is sufficient to choose n large enough s.t. for every h_i

We have

$$Pr\{A_i\} \leq \frac{\delta}{m} \qquad (7)$$

$$\Downarrow$$

$$Pr\{A_1\} + Pr\{A_2\} + \cdots + Pr\{A_m\} \leq \delta$$

$$\Downarrow$$

$$Pr\{A_1 \cup A_2 \cup \cdots \cup A_m\} \leq \delta$$

Now given our answer to question 1, in order to guarantee that $Pr\{A_i\} \leq \frac{\delta}{m}$ we need to

Choose:

if $n \geq n_1(\epsilon, \frac{\delta}{m}) = \underbrace{\frac{1}{2\epsilon^2} \log \frac{2m}{\delta}}_{n_0(\epsilon,\delta)}$

then

$$Pr\{A_i\}$$

$$= Pr\{ |L_S(h_i) - L_D(h_i)| > \epsilon \} \leq \frac{\delta}{m}$$

$\left(n_0(\epsilon,\delta) = n_1(\epsilon, \frac{\delta}{m}) \right)$

Hence,

Statement: If the number of training data points, $|S|$, is larger than $n_0(\epsilon,\delta) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$, then

with probability $1-\delta$ we have

$\forall i \in \{1,\dots,m\}$: $|L_S(h_i) - L_D(h_i)| < \epsilon$.

What we've shown is that we need

$n_0(\epsilon, \delta) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$ data points to guarantee that for all the m functions $h \in \mathcal{H}$ the value of $L_S(h)$ is close to the value of $L_D(h)$:

w.p $1-\delta$:

$$\forall h \in \mathcal{H} : |L_S(h) - L_D(h)| < \epsilon.$$

Now, Let h^*_S be the minimizer of ERM over the class $\mathcal{H}$:

$$h^*_S : \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

($L_S(h^*_S)$ is the smallest among $\mathcal{H}$)

Let $h^*_{true} : \operatorname{argmin}_{h \in \mathcal{H}} L_D(h)$

We will show that if the number of training data points is at least $n_0(\epsilon, \delta)$ then with $1-\delta$ we have

$$L_D(h^*_S) - \min_{h \in \mathcal{H}} L_D(h) \leq 2\epsilon.$$

Hence, $\mathcal{H}$ is PAC-learnable with $n_0(\epsilon, \delta)$ data points.

Let's see why (8) holds.

$$L_D(h^*_S) - L_D(h^*_{true})$$

$$= \underbrace{L_D(h^*_S) - L_S(h^*_S)}_{\leq \epsilon} + \underbrace{L_S(h^*_S) - L_S(h_{true})}_{\leq 0} + L_S(h^*_{true}) - L_D(h^*_{true})$$

(h^*_S is the minimizer of $L_S(\cdot)$)

$$\leq \epsilon + 0 + \epsilon \leq 2\epsilon.$$

Hence, if the number of training data points is at least $n_0(\epsilon, \delta) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$

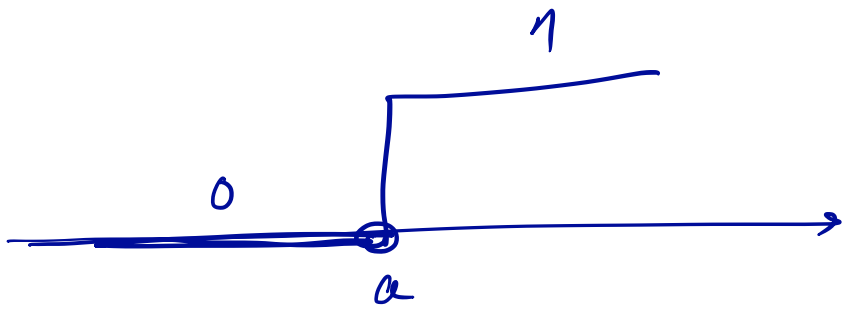
then w.p. $1-\delta$ we have

$$0 \leq L_D(h_S^*) - \min_{h \in \mathcal{H}} L_D(h) \leq 2\epsilon.$$

h_S^*: minimizer of ERM over $\mathcal{H}$.

When $\mathcal{H} = \{h_1, \dots, h_m\}$.

$h_a(x) =$



$$H = \{ h_a(x) \;,\; a \in (-\infty, \infty) \}$$

infinitely many functions

One important consequece of PAC-learnebility is that it's sufficient to work with a finite data set and we don't lose any thing in terms of generalization