richardwu.ca

CS 485/685 Course Notes

Machine Learning: Statistical and Computational Foundations

Shai Ben-David • Winter 2019 • University of Waterloo

Last Revision: February 28, 2019

Table of Contents

Abstract

These notes are intended as a resource for myself; past, present, or future students of this course, and anyone interested in the material. The goal is to provide an end-to-end resource that covers all material discussed in the course displayed in an organized manner. These notes are my interpretation and transcription of the content covered in lectures. The instructor has not verified or confirmed the accuracy of these notes, and any discrepancies, misunderstandings, typos, etc. as these notes relate to course's content is not the responsibility of the instructor. If you spot any errors or would like to contribute, please contact me directly.

1 January 8, 2019

1.1 What is machine learning?

In machine learning, we aim to construct a program that takes as input **experiences** and produces as output **expertise**, or what we have learned from the experience.
We can then apply the **expertise** to produce useful programs such as a spam filter.
An example of learning in nature is **bait shyness**: rats who become sick from eating poisoned bait will become more cautious of food of similar characteristic in the future. Since rats will become more cautious of bait in the future, a delayed poison mechanism (rat is poisoned only 2 days after consuming the bait) is necessary for effective bait by de-associating poison from the bait.
Another example is an experiment called **pigeon superstition** by Skinner (1947): pigeons are starved in a cage with various objects. At random intervals, food is dispersed to satiate the pigeons. Eventually, each pigeon develops a "superstition": they each associate one arbitrary behaviour (e.g. a specific object or a specific movement) that results in food being dispersed.
On the contrary, Garcia (1996) tried a similar experiment to bait shyness with rats where poisoned and un-poisoned bait were identical in characteristic. Whenever a rat approached poisoned bait, a stimulus (e.g. bell ringing, electric shock) was applied to the rat. Surprisingly, the rats did not associate the arbitrary stimulus to the poisoning. This is contrary to the pigeon superstition: this can be explained by evolution (future generations are those that can become aware of poisonous bait) and the fact that rats have **prior knowledge** that poisoning comes from the bait itself, not some arbitrary stimulus.

1.2 Why do we need machine learning?

We desire machines to perform learning because machines can **process lots of data** and are (generally) **fast**.
We desire machines to *learn* because some tasks are simply to complex to hardcode in (e.g. image recognition). Some tasks we do not fully understand how to solve with hardcoded rules. Furthermore, learning allows adaptivity where the machine can constantly learn from new experiences and inputs.

1.3 Types of machine learning

Supervised and unsupervised Machine learning can be generally classified as either **supervised** or **unsupervised**.

Supervised learning takes labelled examples as experience and tries to re-produce these labels on future examples by learning rules. Spam detection may be supervised learning.

Unsupervised learning does not require labelled training data. Examples of unsupervised learning is outlier detection and clustering.

Semi-supervised learning takes as input both labelled and unlabelled data and sits between supervised and unsupervised.

Reinforcement learning also sits between supervised and unsupervised: the machine knows only the rules of the environment and takes actions until a reward (i.e. label) is produced. The machine then learns to label intermediary actions to the final reward produced in the episode (sequence of actions that resulted in the reward).

Passive and active We can also distinguish between **passive** and **active** learning: the former simply takes observed data whereas the latter involves actively performing experiments and interpreting the consequences of the experiments.

Teacher Machine learning can be guided by a "teacher" i.e. how the random sample used as input is generated. Teachers may be **indifferent**, **helpful** or **adversarial**. Helpful teachers produce hints and try to guide the program in the right direction whereas adversarial tries to fool the program.

Batch and online **Batch** learning is learning from a relatively large corpus of data before producing expertise. In contrast **online** learning requires the program to learn as experience is streamed and may result in more mistakes being made.

2 January 10, 2019

2.1 Components of a model

For example sakes, suppose we observe a number of papayas and assign them a score $\in [0, 1]$ for color and hardness. We then label each one as either tasty or not tasty. Using our observations, we would like to predict in the future the tastiness of papayas based on their color and hardness score.

Input The **input** to our learner consists of three parts:

Domain set (X) It is the set of our explanatory variates, in this case $[0, 1] \times [0, 1]$ corresponding to the color score and the hardness score.

Label set (Y) It is the set of our labels: tasty and not tasty.

Training set Our observations i.e. $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subset X \times Y$

Output The **output** of our learner is a **prediction rule** $h : X \to Y$ i.e. the function we learn that maps our papaya scores to a label.

Simple generating model There exists some underlying (unknown) generating process of the population we are interested in (papayas). There is some unknown **probability distribution D over X** and some unknown **labelling rule** $f : X \to Y$.

Together (D, f) describes the generation of papayas.

Success measure We define some metric to measure how well our learner learns the underlying generating model. For example

$$L_{(D,f)}(h) = \Pr_{x \sim D}\left[h(x) \neq f(x)\right]$$

We would like to minimize $L_{(D,f)}(h)$ to find the optimal learner h.

2.2 Empirical Risk Minimization (ERM)

As a first strategy, a learner can employ **Empirical Risk Minimization (ERM)** whereby it picks an h that minimize the errors on the *training sample*.
There is however an issue with this strategy: suppose we learn a rule where we label tasty for papayas with scores that exactly match our tasty papayas' scores and not tasty for everything else. That is

$$h_S(x) = \begin{cases} \text{tasty} & \text{if } (x, \text{tasty}) \in S \\ \text{not tasty} & \text{otherwise} \end{cases}$$

We define the **empirical loss (risk)** over a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$

$$L_S(h) = \frac{|\{i \mid h(x_i) \neq y_i\}|}{m}$$

Note the above strategy give us exactly zero empirical error on our sample set S for any generating process but is obviously not a very robust strategy as it **overfits** to our sample.
Suppose the generating process is such that D is the uniform distribution over $[0, 1] \times [0, 1]$ and let

$$f(x_1, x_2) = \begin{cases} \text{tasty} & \text{if both coordinates in } [0, 1, 0.9] \\ \text{not tasty} & \text{otherwise} \end{cases}$$

Note that the empirical risk is $L_S(h) = 0$, but the risk on our population is $L_{(D,f)}(h) = (0.8)^2 = 0.64$ (since we would be predicting incorrectly for infinitely many points in the region $[0.1, 0.9] \times [0.1, 0.9]$).

2.3 Introducing prior knowledge with inductive bias

For the papaya example above, we could incorporate some prior knowledge such that tasty papayas belong in some rectangular region of color and hardness scores (which we must learn). We could have also assumed the tasty papayas belong in some linear halfspace, or some arbitrary region that we can learn.
More formally, prior knowledge are a set of rules H (set of functions from X to Y) assumed by the learner to contain a good predictor: H is a **hypothesis class**.
Reformulating our previous ERM strategy, ERM_H picks $h \in H$ that minimizes empirical risk over the training set, that is we pick h^* where

$$h^* \in \text{argmin}_{h \in H} L_S(h)$$

Under the following assumptions ERM_H has good success guarantees:

Assumption 1 (Realizability) $\exists h \in H$ such that $L_{(D,f)}(h) = 0$

Assumption 2 S is picked iid by D and labelled by f (i.e. our sample is representative)

3 January 15, 2019

3.1 Finite hypothesis classes

Theorem 3.1. Let H be a **finite set** of predictors (our hypothesis class). Assume our two assumptions from above hold. Then every ERM_H learning rule is guaranteed to converge to a zero-loss predictor as the sample size tends to infinity.

Namely for every ERM_H learner A and every $\epsilon > 0$

$$\Pr_{S\sim X\times f}\left[L_{(D,f)}(A(S)) > \epsilon\right] \to 0$$

as $|S| \to \infty$ (this is exactly convergence in probability where $\Pr\left[L_{(D,f)}(A(S))\right] \to 0$).

Proof. Let B_ϵ denote the set of all hypotheses in H that have error $> \epsilon$ i.e.

$$B_\epsilon = \{h \in H \mid L_{(D,f)}(h) > \epsilon\}$$

Let our set of "misleading" samples be

$$M = \{S \mid |S| = m \text{ and } \exists h \in B_\epsilon \text{ s.t. } L_S(h) = 0\}$$

(samples where we have a zero empirical loss but has $> \epsilon$ loss on the true population: misleading because it tricks us that the $h \in B_\epsilon$ re-constructs f with zero error when in fact it does not).
Note that

$$\Pr_S\left[L_{(D,f)}(A(S)) > \epsilon\right] \leq \Pr\left[S \in M\right]$$

that is the probability that our sample does not perform better than ϵ on our true population is bounded by the probability of picking a misleading sample (this is not just an equality since we also have samples S where $L_{(D,f)}(A(S)) > \epsilon$ and $L_S(A) > 0$).

Lemma 3.1. We claim

$$\Pr_{|S|=m}\left[S \in M\right] \leq |H|(1-\epsilon)^m$$

Proof. Consider any $h \in B$ (where obviously $h \neq f$). There exists a "disagreement" region D where for any x, $h(x) \neq f(x)$ (either $h(x) = +, f(x) = -$ or $h(x) = -, f(x) = +$).
For our sample S of size m, we know that $S \subseteq D^c$ (our sample cannot be in the disagreement region since $L_S(h) = 0$ i.e. our sample is perfect; empirical loss is zero so it must agree with f).
Note that since $L_s(h) > \epsilon$ (i.e. h disagrees with f on a region of proportion at least ϵ, our D), then the region where h and f agree is at most of proportion $1 - \epsilon$ (i.e. D^c).
Choose m sample points iid from D^c is thus

$$\Pr_S\left[L_S(h) = 0\right] \leq (1-\epsilon)^m$$

□

Lemma 3.2 (Union bound)**.** Given two set of events A, B we know $P(A \cup B) \leq P(A) + P(B)$.

Note that $\Pr[S \in M] = \Pr[\text{for some } h \in B, L_S(h) = 0]$ is the union of all misleading hypotheses $h \in B$, thus

$$\begin{aligned}\Pr\left[\text{for some } h \in B, L_S(h) = 0\right] &\leq \sum_{h\in B} \Pr(L_S(h) = 0) \\ &= |B|(1-\epsilon)^m \\ &< |H|(1-\epsilon)^m\end{aligned}$$

Note that $1 - \epsilon \leq e^{-\epsilon}$ thus $\Pr[S \in M] \leq |H|e^{-\epsilon m}$ which goes to 0 as $m \to \infty$ as desired. □

4 January 17, 2019

4.1 Probably Approximately Correct (PAC) learning

In our previous theorem with ERM_H with inductive bias we showed it could do well but only under **strong assumptions**.
Our goal is to prove similar guarantees but with more realistic/relaxed assumptions. Specifically, we would like to relax our assumption that there exists a *detrministic* f that generates the true distribution D (the labels Y) over domain X.
That is, the relax model given domain of instances X and label set Y, the data is generated by a probability distribution D over $X \times Y$. We denote our set of predictors $h : X \to Y$ as H.
Our relaxed model defines the empirical loss as

$$L_S(h) = \frac{|i \mid h(x_i) \neq y_i|}{|S|}$$

and the true loss over our probability distribution D over $X \times Y$

$$L_D(h) = P(\alpha(x, y) \mid h(x) \neq y)$$

Definition 4.1 (PAC learnable (Leslie Valiant 1984))**.** A *hypothesis class* H is **PAC learnable** if there exists a function $m_H(\epsilon, \delta) : (0, 1) \times (0, 1) \to \mathbb{N}$ and a learner A (map from labelled samples to functions $h : X \to Y$) such that for every $\epsilon, \delta \in (0, 1)$ for every distribution D over X and every labelling function $f \in H$, if $m' \geq m(\epsilon, \delta)$ and a labelled sample $S = \{(x_1, f(x_1)), \ldots, (x_m, f(x'_m))\}$ generated iid according to D and labelled by f, then

$$\Pr_{S \sim (D^m, f)} \left[L_{(D,f)}(A(S)) > \epsilon \right] < \delta$$

That is: the error is bounded by ϵ (approximately), and the probability of error is bounded by δ (probably) for some large enough sample size.

Remark 4.1. The number of required samples is determined regardless of D and f.

Some weaknesses of this definition:

- Realizability assumption ($h \in H$ such that $L_{(D,f)}(h) = 0$): the learner has strong prior knowledge.
- The labelling rule is *deterministic*: the label of any x is fully determined by X.
- The training distribution and test distribution are the sample: this may be unobtainable in some cases.

4.2 Finite hypothesis H is PAC learnable

Theorem 4.1. Any finite H is PAC learnable.

Proof. Recall if H is finite then

$$\Pr_{S \sim (D^m, f)} \left[L_{(D,f)}(A(S)) > \epsilon \right] < |H| e^{-m\epsilon}$$

Our claim holds if $|H| e^{-m\epsilon} \leq \epsilon$. Solving for m

$$m \geq \frac{\ln(|H|) + \ln\left(\frac{1}{\delta}\right)}{\epsilon}$$

□

4.3 Real intervals on real domain is PAC learnable

Theorem 4.2. Let $X = \mathbb{R}$ and H is the class of all real intervals where $H_{int} = \{h_{(a,b)} \mid a \leq b\}$ where

$$h_{(a,b)}(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

then H_{int} is PAC learnable.

Proof. Suppose we have a sample generated by $f \in H_{int}$ where all our positive examples and only positive examples lie within an interval of $X = \mathbb{R}$ (since $f \in H_{int}$ so it labels positive examples only in a real interval), e.g.

$$\ldots \quad 0 \quad 0 \quad 1 \quad 1 \qquad 1 \quad \ldots$$

where $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Our learner A could be such that $A(S) \in H_{int}$ where

$$a(S) = \min\{x_i \mid (x_i, 1) \in S\}$$
$$b(S) = \max\{x_i \mid (x_i, 1) \in S\}$$

We show that this rule A is a successful PAC learner for H_{int}.
Given ϵ, m let us upper bound the probability that an m-size sample will lead A to out h with $> \epsilon$ error.
Denote $B_\epsilon = \{h \in H \mid L_{(D,f)}(h) > \epsilon\}$ (bad hypotheses) and $M = \{S \mid A(S) \in B\}$ (set of misleading samples).
Note that a sample $S \in M$ is misleading if our minimum and maximum positive samples cover a "small" region of the actual interval specified $f \in H_{int}$ our arbitrary labelling function.
That is

$$\Pr(S \in M) = \Pr\big(S \text{ does not hit intervals } [\min(f), min(h)] \text{ or } [\max(h), \max(f)]\big)$$

S is a misleading sample only if S does not hit either the interval from $\min(f)$ to $\min(f) + weight_D(\epsilon/2)$ or interval form $\max(f) - weight_D(\epsilon/2)$ to $\max(f)$ (where $weight_D(\epsilon/2)$ is defined as the region R immediately to the right/left where $\Pr_D(x \in R) = \epsilon/2$).
That is

$$\Pr(S \in M) \leq \big(1 - \frac{\epsilon}{2}\big)^m + \big(1 - \frac{\epsilon}{2}\big)^m$$

where each sample misses both intervals of weight/probability $\frac{\epsilon}{2}$.
Therefore

$$\Pr_{S \sim (D^m, f)} \big[L_{(D,f)}(A(S)) > \epsilon\big] \leq 2\big(1 - \frac{\epsilon}{2}\big)^m$$

for some $m(\epsilon, \delta)$ such that $2\big(1 - \frac{\epsilon}{2}\big)^m < \delta$, thus H_{int} is PAC learnable. □

5 January 22, 2019

5.1 Agnostic PAC learning (more general learning model)

Our previous definition of PAC learnable is still too unrealistic. Namely we are going to:

1. Remove the *realizability* assumption: we do not require there exist $h \in H$ such that $L_D(h) = 0$ (although we do not necessarily remove the requirement that the labelling rule is in H)
2. Remove the deterministic labelling requirement (allow smae x to show up with different labels)

We then assume there exists some probability distribution D over "abstract" set Z whereby data is generated iid.
Our new notion of loss is we are given some function $l(\text{hypothesis}, z \in Z)$ which is real-valued.
For example we may have for email spam detection:

Example 5.1. Let $Z = X \times \{0,1\}$ where X is the set of emails and 0/1 denotes not spam or spam.
Let $h : X \to \{0,1\}$ (our hypothesis function).
Let

$$l(h,(x,y)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{if } h(x) = y \end{cases}$$

Given ta training sample $S = (z_1, \ldots, z_m)$ and a predictor (hypothesis function) h the **empirical loss** of h on S is now

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$$

Also we define the **true loss** of h as (for true distribution D over Z)

$$L_D(h) = \mathbb{E}_{Z \sim D}(l(h,z))$$

Remark 5.1. In our example with $Z = X \times \{0,1\}$ and $l_{0,1}$ our new definition of empirical and true loss $L_S(h)$ and $L_D(h)$ are equivalent to our definitions in PAC learnable.

However our individual loss function could be arbitrarily defined:

Example 5.2. Suppose we want to predict tomorrow's temperature from today's measurements.
Let $Z = \big(\text{today's measurements} \times \text{tomorrow's temp}\big)$.
Let $h : \text{today's measurements} \to [-50, +50]$.
We define loss as $l(h,(x,y)) = |h(x) - y|$ (L1 norm).

Example 5.3 (K-means clustering)**.** Suppose we would like to pick k locations for a chain of stores in KW.
Each h will represent a set of k potential locations $h = (\mu_1, \ldots, \mu_k)$.
Therefore we let $Z =$ location of customers seeking a store and we define our loss to be

$$l((\mu_1, \ldots, \mu_k), z) = \min_{1 \leq i \leq k} |z - \mu_i|$$

i.e. the loss is the L1 distance between a customer z and the closest location μ_i.
Our training data would then be a sample $S = (z_1, \ldots, z_m)$ which is a record of past customers.

We will now explore how learning is achieved under this more general model.
The prior knowledge of the learner is again modeled by a set H of possible predictors (class of hypotheses). Our **input** is a training set $S = (z_1, \ldots, z_m)$ generated iid by some unknown D over Z. The **output** of the learner is a predictor h.
Our goal is the minimize the *true loss* $L_D(h)$.

Remark 5.2. If we know the distribution D over Z then we could solve our problem without learning, but of course all we know is S.

Definition 5.1 (Agnostic PAC learnable)**.** A class H is **agnostic PAC learnable** if $\exists m : (0,1) \times (0,1) \to \mathbb{N}$ and a learner A (mapping S's to h's) such that $\forall \epsilon \forall \delta$, for all distribution D over Z, and for all $m \geq m(\epsilon, \delta)$ we have

$$\Pr_{z \sim D^m} \left[L_D(A(S)) \geq \min_{h \in H} L_D(h) + \epsilon \right] < \delta$$

Remark 5.3. Agnostic PAC learnable is almost identical to PAC learnable except we do not assume a lower bound of 0 on $L_D(A(S))$: instead we lower bound it with $\min_{h\in H} L_D(h)$ plus some small ϵ.
Furthermore we no longer assume a determinisitic f and instead describe a D over Z.

Remark 5.4. More correctly, we require only $\inf_{h\in H} L_D(h)$: we need not require an attainable minimum.

How can we learn in this new model? In many cases ERM_H is still a good strategy.

6 January 24, 2019

6.1 Uniform Convergence Property

Definition 6.1 (ϵ-representative)**.** A sample $S = (z_1, \ldots, z_m)$ is **ϵ-representative of** H with respect to a distribution D if for any $h \in H$ we have $|L_S(h) - L_D(h)| < \epsilon$.

Claim. If S is $\frac{\epsilon}{2}$-representative of H wrt D then for any ERM_H learner A

$$L_D(A(S)) \leq \min_{h\in H} L_D(h) + \epsilon$$

This is the **Uniform Convergence Property**.

Proof. Note that for any $h \in H$

$$\begin{aligned} L_D(A(S)) &\leq L_S(A(S)) + \frac{\epsilon}{2} && \frac{\epsilon}{2} - \text{representative} \\ &\leq L_S(h) + \frac{\epsilon}{2} && A(S) \text{ is } ERM_H \\ &\leq L_D(h) + \epsilon && \frac{\epsilon}{2} - \text{representative} \end{aligned}$$

sine this holds for any h then $L_D(A(S)) \leq \min_{h\in H} L_D(h) + \epsilon$. □

6.2 Finite hypothesis classes are agnostic PAC learnable

Claim. Given a **finite** hypothesis class H, H is agnostic PAC learnable.

Question 6.1. For a given distribution D and **finite** hypothesis class H, how do we determine m large enough such that

$$L_D(A(S)) \leq \min_{h\in H} L_D(h) + \epsilon$$

i.e. such that H is agnostic PAC learnable?

Proof. Note that

$$\begin{aligned} & P_{S\sim D^m}\left[\forall h \in H \text{ s.t. } |L_S(h) - L_D(h)| \leq \frac{\epsilon}{2}\right] > 1 - \delta \\ =& D^m\left[\{S \mid \exists h \in H \text{ s.t. } |L_S(h) - L_D(h)| > \frac{\epsilon}{2}\}\right] < \delta \\ \iff & \bigcup_{h\in H} D^m\left[\{S \mid h \in H \text{ s.t. } |L_S(h) - L_D(h)| > \frac{\epsilon}{2}\}\right] < \delta \end{aligned}$$

where $D^m\big[\{S\}\big]$ is the total probability mass of $\{S\}$ in D^m.

Recall $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$ and $L_D(h) = \mathbb{E}_{Z\sim D}\big[l(h,z)\big]$. Let $\theta_i = l(h, z_i)$ and let $L_D(h) = \mu$, thus $\mathbb{E}(\theta_i) = \mu$. Note that on the LHS we have

$$\begin{aligned}
&\bigcup_{h\in H} D^m\left[\{S \mid h \in H \text{ s.t. } |L_S(h) - L_D(h)| > \frac{\epsilon}{2}\}\right] \\
&\leq \sum_{h\in H} D^m\left[\{S \mid h \in H \text{ s.t. } |L_S(h) - L_D(h)| > \frac{\epsilon}{2}\}\right] \\
&\leq \sum_{h\in H} P\left[\left|\frac{1}{m}\sum \theta_i - \mu\right| > \frac{\epsilon}{2}\right] \\
&\Rightarrow 2|H| \exp\left(-2m\left(\frac{\epsilon}{2}\right)^2\right) < \delta
\end{aligned}$$

where the second last inequality follows from **Hoeffding's inequality** (assuming $l \in [0,1]$):

Theorem 6.1 (Hoeffding's inequality)**.** Let $\theta_1, \ldots, \theta_n$ be random variables where $\mathbb{E}(\theta_i) = \mu$ and $a \leq \theta_i \leq b$. Then

$$P\left[\left|\frac{1}{m}\sum \theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

Solving for m we get

$$m \geq \frac{2\log\left(\frac{2|H|}{\delta}\right)}{\epsilon^2}$$

□

Since there exists such a function $m(\epsilon, \delta)$ a finite hypothesis class H is agnostic PAC learnable.

7 January 31, 2019

7.1 Minimal sample size for the class of all functions

Theorem 7.1. If X has size $2m$ then we need $\geq d$ examples to learn the hypothesis class of *all functions* over $X \times \{0,1\}$ to an accuracy of $\frac{1}{4}$ with $\delta \leq \frac{1}{4}$.

Proof. Let X be our domain and D a *uniform distribution* over X.
Choose an $f : X \to \{0,1\}$ to label our points. Our learner's input is $\{(x_1, f(x_1)), \ldots, (x_m, f(x_m))\}$.
Suppose $m < \frac{|X|}{2}$ i.e. our sample is less than half the size of our domain.
We then have a region $X \setminus S$ where any new point x has probability $\geq \frac{1}{2}$ of being in $X \setminus S$ which our learner is impartial to labelling as either 0 or 1 (since we have no information from our sample).
Furthermore since we have an arbitrary labelling function f, then for every $x \notin S$ we have

$$\Pr\big[A(S) \neq f(x)\big] = \frac{1}{2}$$

or $\Pr_{(D,f)}\big[A(s) \neq f(x)\big] \geq \frac{1}{4}$. □

7.2 No Free Lunch Theorem

Theorem 7.2 (No Free Lunch Theorem)**.** Let A be any learning algorithm for the task of binary classification ($0-1$ loss) over a domain X. Let the sample size m be any number smaller than $\frac{|X|}{2}$. Then there exists a distribution D over $X \times \{0,1\}$ such that:

1. There exists a function $f : X \to \{0,1\}$ with $L_D(f) = 0$
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ we have $L_D(A(S)) \geq \frac{1}{8}$

i.e. this this theorem states that there exists a task for any learner A that it fails on, but which there is another learner that can successfully learn it. A trivial successful ERM learner would be one with $H = \{f\}$ or more generally an ERM with finite hypothesis class whose size satisfies $m \geq 8 \log\left(\frac{7|H|}{6}\right)$.

Proof. Let $C \subseteq X$ be of size $2m$.
The intuition is that any learner that has observed only half of the instances of C has no information regarding the labels in the rest of C. Therefore there exists some "reality" i.e. some target function f that would always contradict labels assigned by $A(S)$ on unobserved instances.
Note that there are $T = 2^{2m}$ possible functions/labellings from C to $\{0,1\}$. Denote these functions as $f_1, \ldots, f_T$.
Let D_i be distributions over $C \times \{0,1\}$ where

$$D_i((x,y)) = \begin{cases} \frac{1}{|C|} & \text{if } y = f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

That is $L_{D_i}(f_i) = 0$ (wrong labels have probability 0).
We show that for any algorithm A with sample size m it holds that

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] \geq \frac{1}{4}$$

that is there exists some $f : X \to \{0,1\}$ and distribution D where $L_D(f) = 0$ and

$$\mathbb{E}_{S \sim D^m}\left[L_D(A(S))\right] \geq \frac{1}{4}$$

if the above holds then our claim holds.
Note that there are $k = (2m)^m$ possible sequences of m examples from C. Denote these sequences as $S_1, \ldots, S_k$.
Also if $S_j = (x_1, \ldots, x_m)$ then we denote $S_j^i = \{(x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))\}$ (sample labelled by f_i).
If the distribution is D_i then A can receive training sets $S_1^i, \ldots, S_k^i$. Note that all these samples have equal probability of being sampled (because of D_i's definition) thus

$$\mathbb{E}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

Using the fact that maximums $\geq$ averages $\geq$ minimums we have

$$\begin{aligned}\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \\ &\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i))\end{aligned}$$

Let us fix some $j \in [k]$ (fix some sample). Let $S_j = (x_1, \ldots, x_m)$ and let $v_1, \ldots, v_p$ be the examples in $C \setminus S_j$. Note that $p \geq m$ (since we only have half of C). Therefore for every $h : C \to \{0, 1\}$ and every i

$$\begin{aligned}L_{D_i}(h) &= \frac{1}{2m}\sum_{x\in C} \mathbb{1}_{[h(x)\neq f_i(x)]} \\ &\geq \frac{1}{2m}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \\ &\geq \frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]}\end{aligned}$$

Therefore we have

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \\ &\geq \frac{1}{2}\min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\end{aligned}$$

For any $r \in [p]$ (any point not in our sample) we can partition $f_1, \ldots, f_T$ into $T/2$ disjoint pairs $(f_i, f_{i'})$ such that for $c \in C$ $f_i(c) \neq f_{i'}(c)$ if and only if $c = v_r$ (they only differ labelling on one point v_r). Since for this pair we must have $S_j^i = S_j^{i'}$ (points in sample must all be the same) then

$$\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^{i'})(v_r)\neq f_{i'}(v_r)]} = 1$$

(at most one is labelled incorrectly by A), thus we have

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} = \frac{1}{2}$$

thus combining everything our claim holds. □

8 February 5, 2019

8.1 Summary of PAC learnability

We have proved PAC learnabiliy for a number of hypothesis classes, namely:

- Every finite H
- The set of intervals on $\mathbb{R}$

We also note that every ERM_H is a *good PAC learner.*
Some classes we've seen that are PAC *unlearnable*:

- If X has size $\geq 2d$ then we need $\geq d$ examples to learn the class of *all functions* to accuracy $\frac{1}{4}$ with $\delta \leq \frac{1}{4}$

8.2 Infinite domain on class of all functions

A corollary to our theorem before regarding sample sizes $< \frac{|X|}{2}$:

Corollary 8.1. If X is infinite then the class of all functions (from X to $\{0,1\}$) is *not PAC learnable.*
That is, without inductive bias (prior knowledge) we cannot learn from an infinite domain.

Proof. Assume for contradiction that H_{All} is PAC learnable.
Namely $\exists$ learner A and $m(\epsilon, \delta)$ such that for all D over $X \times \{0,1\}$ and $\forall \epsilon, \delta > 0$ on sample size $\geq m(\epsilon, \delta)$ we have

$$\Pr_{(S \sim D^m, f)} \left[L_{(D,f)}\big(A(S)\big) > \epsilon \right] < \delta$$

Consider the number $m(0.1, 0.1)$. Pick a $W \subseteq X$ of size $> 2m(0.1, 0.1)$.
Pick D to be uniform over W where for all $x \in X$ we have

$$D(x) = \begin{cases} \frac{1}{|W|} & \text{if } x \in W \\ 0 & \text{if } x \notin W \end{cases}$$

Since H contains every function over W from our theorem above we require $> \frac{|W|}{2}$ for $\epsilon \leq \frac{1}{8}, \delta \leq \frac{1}{7}$, but we promised that $m(0.1, 0.1)$ should suffice, thus we have a contradiction. □

8.3 Shattering

So when exactly does ERM_H succeed? We saw that $H^{\mathbb{R}}_{intervals}$ has a good ERM_H learner but we also saw that $H^{\mathbb{R}}_{finite}$ where

$$H^{\mathbb{R}}_{finite} = \{f : \mathbb{R} \to \{0,1\} \mid f^{-1}(1) \text{ is finite}\}$$

would not succeed.
In 1970 Vapnik-Chervonenkis and in 1989 EBHW both measured teh compleity of class H that fully determines whether H is learnable.
We begin with a few definitions:

Definition 8.1 (Shattering)**.** A class of functions H (from X to $\{0,1\}$) **shatters** $W \subseteq X$ if **for every** $f : W \to \{0,1\}$ there is some $h \in H$ such that for every $x \in W$: $h(x) = f(x)$.
That is: for any possible labelling of X (which we have $2^{|X|}$) there exists some $h \in H$ that can produce the same labelling.

Example 8.1. Let $X = \mathbb{R}$ and $W = \{a, b, c\}$ where $a < b < c$.
Does $H_{intervals}$ shatter W? **No**. Consider

$$f(x) = \begin{cases} 1 & \text{if } x \in \{a, b\} \\ 0 & \text{if } x = c \end{cases}$$

Does H_{finite} shatter W? **Yes**, since for every labelling function over $\{a, b, c\}$ (we have 2^3 such labelling functions) there exists $h \in H_{finite}$ that corresponds to every such labelling function.

Example 8.2. Let $X = [0, 1]^2$ (unit square). Let $W = \{(0.2, 0.2), (0.3, 0.3), (0.4, 0.1)\}$.
Does $H_{rectangles}$ shatter W? Yes (we can clearly see we can draw rectangles around any subset of the points).
However if W are three collinear points e.g. $\{(0.1, 0.1), (0.2, 0.2), (0.3, 0.3)\}$ then the corrsponding labelling $1 - 0 - 1$ would not be shatter-able by $H_{rectangles}$.

9 February 7, 2019

9.1 V-C dimension

Definition 9.1 (V-C dimension)**.** Given a class H the **V-C dimension of** H is the size of the **largest set** that H shatters, that is

$$VC(H) = \max_{|A|}\{H \text{ shatters } A\}$$

It is ∞ if H shatters aribtrarily large W's.

Remark 9.1. We can represent functions $f : X \to \{0, 1\}$ as subsets of X where for a given function f we have

$$S_f = \{x \in X \mid f(x) = 1\}$$

Similarly the converse holds: given $B \subseteq X$ consider

$$f_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

Thus we have a 1-1 correspondence.

Remark 9.2. H shatters A if

$$\{B \mid B \subseteq A\} = \{h \cap A \mid h \in H\}$$

that is: H shatters A if we can produce label every subset of A with 1 (and all else as 0).

Remark 9.3. For the hypothesis class of all functions H_{all} note that $|H_{all}^X| = 2^{|X|}$. Furthermore note that the collection of all subsets of $B \subseteq X$ is $|\{B \mid B \subseteq X\}| = 2^{|X|}$.
Since both sets have the same cardinality and the collection of $B \subseteq X$ is maximal in XS then $VC(H_{all}) = \infty$.

For example, the V-C dimensions of the classes:

- $VC(H_{intervals}^{\mathbb{R}}) = 2$
- $VC(H_{finite}^{\mathbb{R}}) = \infty$
- $VC(H_{rect}^{\mathbb{R}^2}) \geq 4$: note that we can easily draw 4 points which we can shatter with rectangles.

Claim. H_{rect} cannot shatter *any* set of 5 points ($VC(H^{\mathbb{R}^2}_{rect}) \leq 4$).

Proof. Let A be any set of > 4 points. Pick the 4 points in A that are farthest left, right, up and down i.e. let $B = \{top_A, bot_A, left_A, right_A\}$.

Note that we can never have an $h \in H_{rect}$ that *exactly labels* only B since h would pick all of A if h contains B. □

Therefore $VC(H^{R^2}_{rect}) = 4$.

9.2 Hyper-rectangle class

We extend our rectangle examples in $\mathbb{R}^1$ and $\mathbb{R}^2$ to any $\mathbb{R}^d$ for $d \in \mathbb{N}$:

Definition 9.2 (Hyper-rectangle class)**.** The **class of hyper-rectangles** is defined as

$$H^{\mathbb{R}^d}_{rect} = \{h_{\{(a_1,b_1),\ldots,(a_d,b_d)\}} \mid a_1, b_1, a_2, b_2, \ldots, a_d, b_d \in \mathbb{R}\}$$

For any set of intervals $\{(a_1, b_1), (a_2, b_2), \ldots, (a_d, b_d)\}$ (min and max bounds for every dimension d) let

$$h_{\{(a_1,b_1),\ldots,(a_d,b_d)\}} = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$$

define a **hyper-rectangle** in $\mathbb{R}^d$.
Then for all $X = (x_1, \ldots, x_d)$ we have

$$h_{\{(a_1,b_1),\ldots,(a_d,b_d)\}}(x_1, \ldots, x_d) = \begin{cases} 1 & \text{if for all } i \leq d, a_i \leq x_i \leq d_i \\ 0 & \text{otherwise} \end{cases}$$

Claim. We claim $VC(H^{\mathbb{R}^d}_{rect}) \geq 2d$.

Proof. Let

$$A = \{(1, 0, \ldots, 0), (-1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), (0, -1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1), (0, \ldots, 0, -1)\} \subseteq \mathbb{R}^d$$

where A is the set of one-hot vectors in $\mathbb{R}^d$ and their negations.
More compactly if $l^d_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ where only the ith dimension of l_i is 1 then

$$A = \{l_i \mid 1 \leq i \leq d\} \cup \{-l_i \mid 1 \leq i \leq d\}$$

Given any $B \subseteq A$ let h_B be the hyper-rectangle $[a_1, b_1] \times \ldots \times [a_d, b_d]$ such that for every $1 \leq i \leq d$:

if both l_i and $-l_i \in B$	$a_i = -2$	$b_i = 2$
if $l_i \in B, -l_i \notin B$	$a_i = 0$	$b_i = 2$
if $l_i \notin B, -l_i \in B$	$a_i = -2$	$b_i = 0$
otherwise	$a_i = 0$	$b_i = 0$

Note that the boundaries for dimension i must always include 0 since all points $l_j, -l_j$ where $j \neq i$ have their ith coordinate as 0. □

Claim. We claim $VC(H^{\mathbb{R}^d}_{rect}) \leq 2d$.

Proof. The proof follows similarly from the $\mathbb{R}^2$ case.
Given any $A \subseteq \mathbb{R}^d$ where $|A| > 2d$ we pick subset $B \subseteq A$ where for every $1 \leq i \leq d$, we pick *a point* t_i and s_i with maximum and minimum value, respectively, in the ith coordinate. Note that $|B| = 2d$ so $B \neq A$.
For every rectangle h that includes all members of B we have $A \subseteq h$ (h bounds all of A), so h cannot cut B from A thus h cannot shatter A. □

9.3 Bounding V-C dimension of a class

We notice that for $H^{\mathbb{R}^2}_{rect}$ and $H^{\mathbb{R}^d}_{rect}$ we proved minimum and upper bounds of V-C dimension in the following way:

Minimum bound $\exists A \forall B \subseteq A$ such that $\exists h \in H$ that can cut B.

Then $VC(\cdot) \geq |A|$.

Maximum bound $\forall A \exists B \subseteq A$ such that $\nexists h \in H$ that can cut B.

Then $VC(\cdot) < |A|$.

We provide another example:

Example 9.1. Let $X = \mathbb{N}$ and $H^{\mathbb{N}}_5 = \{A \subseteq N \mid |A| = 5\}$ (all functions that mark exactly 5 points as 1).

Claim. $VC(H^{\mathbb{N}}_5) \geq 5$.

Proof. Pick $A = \{6, 7, 8, 9, 10\}$. For any $B \subseteq A$, let

$$h_B = B \cup (5 - |B|) \text{ points above } 10$$

where $|h_B| = 5$. Clearly $h_B \cap A = B$ so h_B cuts B and thus shatters A. □

Claim. $VC(H^{\mathbb{N}}_5) \leq 5$.

Proof. For every A of size > 5 let $B = A$. No members of H include all points in B so A is not shattered. □

It follows $VC(H^{\mathbb{N}}_5) = 5$.

9.4 Size of hypothesis class from V-C dimension

Claim. For every H and every d if $VC(H) \geq d$ then $|H| \geq 2^d$.

Proof. There is some A of size d shattered by H so for every $B \subseteq A$ there is a corresponding $h_B \in H$.
Since $|A| \geq d$, then A has $\geq 2^d$ subsets thus $|H| \geq 2^d$. □

Note that the converse does not hold. Consider the following:

Example 9.2. For example, the hypothesis class of intervals on $\mathbb{R}$ obviously has size ∞ ($|H_{int}| = \infty$) but $VC(H_{int}) = 2$.

10 February 14, 2019

10.1 Fundamental theorem of statistical learning

Theorem 10.1 (Fundamental theorem of statistical learning)**.** The follow statements are equivalent for every class H:

1. H has the uniform convergence property
2. Every ERM_H learner is a successful PAC agnostic PAC learner
3. H is agnostic PAC learnable
4. Every ERM_H is successful PAC learnable for H
5. H is PAC learnable
6. $VC(H)$ is finite

We have shown that $1 \Rightarrow 2$, $2 \Rightarrow 3$, $1 \Rightarrow 4$, $4 imply 5$.
To show $5 \Rightarrow 6$ is equivalent to showing:

Claim. If $VC(H) = \infty$ then H is not PAC learnable.

Proof. Proof is basically applying the No Free Lunch Theorem. Recall that the NFL theorem states that if there is a domain subset $W \subseteq X$ of size d such that H *contains all functions from* W *to* $\{0,1\}$ then H shatters W. To PAC learn H to $\delta = \frac{1}{8}, \epsilon = \frac{1}{8}$ we need $\geq \frac{d}{2}$ sample size.
In other words: if H shatters a set of size d then $m_H^{PAC}(\frac{1}{8}, \frac{1}{8}) \geq \frac{d}{2}$.

Corollary 10.1. If $VC(H) = \infty$ then $m_H^{PAC}(\frac{1}{8}, \frac{1}{8})$ is not any finite number (since H shatters an arbitrarily large W).
Therefore H is not PAC learnable.

□

10.2 Shatter function and Sauer's Lemma

We now show $6 \Rightarrow 1$ in order to prove the fundamental theorem of statistical learning holds.

Definition 10.1 (Shatter function)**.** The **shatter function** of a class H is a function $\pi_H : \mathbb{N} \to \mathbb{N}$ defined by

$$\pi_H(m) = \max_{|A|=m} |H_A|$$

where H_A are the functions in H restricted to $A \subseteq X$ i.e. $H_A = \{h_{|A} \mid h \in H\}$ where $h_{|A}$ is the function from A to Y such that for every $x \in A$ we have $h_{|A}(x) = h(x)$.

Some observations about the shatter function π_H:

1. For every m (and every H) note that $\pi_H(m) \leq 2^m$.
2. If $VC(H) \geq m$ then $\pi_H(m) = 2^m$.
3. If $\pi_H(m) < 2^m$ then $VC(H) < m$.

   Note if $\pi_H(m) < 2^m$ then there are no A of size m does H gets all of its behaviours. That is H shatters no set of size m, which implies H shatters no set of size $\geq m$, which implies $VC(H) < m$.

We also require the following lemma and corollary:

Lemma 10.1 (Sauer-Shelah-Perles-Vapnik-Chervonenkis lemma)**.** (AKA **Sauer's or Sauer-Shelah lemma**). For every H and every m

$$\pi_H(m) \leq \sum_{i=0}^{VC(H)} \binom{m}{i} = |\{B \subseteq A \mid |B| \leq d\}|$$

Note that $\binom{m}{i} \leq m^i$.
The RHS is exactly the number of subsets of A of at most size d.

Corollary 10.2. If $VC(H) = d$ then for all m we have $\Pi_H(m) \leq m^d$.

Remark 10.1. We have $\Pi_H(m)$ where $VC(H) = d$ is bounded by both the functions 2^m and m^d.
For a fixed d we eventually have $m^d << 2^m$.
That is: $\Pi_H(m)$ grows exponentially until it reaches $m = VC(H)$, then $\Pi_H(m)$ becomes bounded by a polynomial m^d.

Corollary 10.3. The number of linearly separable subsets of m points is at most m^3.

Proof. Consider Hs^2 the set of linear partitions of $\mathbb{R}^2$.
We claim $VC(Hs^2) = 3$. Clearly it is easy to show there exists an arrangemnet of 3 points we can shatter with linear halfspaces.
We omit the proof that Hs^2 cannot shatter any 4 points.
By Sauer's lemma it follows $\Pi_{Hs^2}(m) \leq m^3$. □

Let's take $m = 1000$ for example. The number of functions on m points in $\mathbb{R}^2$ to $\{0, 1\}$ is 2^{1000}. By the above corollary we know there can be at most $1000^3 \approx 2^{30}$ linear halfspace functions, a very small fraction of possible functions.

11 February 26, 2019

11.1 Extended Sauer's Lemma

Lemma 11.1 (Extended (Sauer's) Lemma)**.** For every set $A \subseteq X$

$$|H_A| \leq |\{B \subseteq A \mid H \text{ shatters } B\}|$$

Remark 11.1. This Extended Lemma implies the Sauer Lemma since if H shatters B and $VC(H) = d$ then this implies $|B| \leq d$.

Remark 11.2. Note the inequality says that the number of subsets that H cuts is *fewer than* the number of subsets H shatters: this is counterintuitive since shattering a set seems harder than cutting a set.

Example 11.1. Let us look at $H_{intervals}$ and the inequality. Suppose $A = \{1, 2, 3, 4, 5\}$.
Note that

$$H_{intervals,A} = \{h_{|A} \mid h \in H\} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}, \{1, 2, 3, 4\}, \{2, 3, 4, 5\}, \varnothing, \{1, 2, 3, 4, 5\}\} \quad (11.1)$$

where $|H_A| = 16$.
Note that $H_{intervals}$ can shatter any sets of size $0, 1, 2$, thus $|\{B \subseteq A \mid H \text{ shatters } B\}| = \binom{5}{2} + \binom{5}{1} + \binom{5}{0} = 16$.
Therefore the Extended Lemma does indeed hold for $H_{intervals}$.

11.2 VC bound on union of classes

How might we use Sauer's Lemma to characterize the behaviour of classes H?

Question 11.1. Let H_1, H_2 be two classes over X and assume $VC(H_1) = VC(H_2) = d$. Can we bound $VC(H_1 \cup H_2$? Let m be the size of a set shattered by $H_1 \cup H_2$. Let A be a subset of size m that $H_1 \cup H_2$ shatters. Then

$$|\{h_{|A} \mid h \in H_1\} \cup \{h_{|A} \mid h \in H_2\}| = |\{h_{|A} \mid h \in H_1 \cup H_2\}| = 2^m$$

where the last equality holds since h shatters A.
Note that

$$\begin{aligned} |\{h_{|A} \mid h \in H_1\} \cup \{h_{|A} \mid h \in H_2\}| &\leq |\{h_{|A} \mid h \in H_1\}| + |\{h_{|A} \mid h \in H_2\}| \\ &\overset{Sauer's lemma}{\leq} m^d + m^d \end{aligned}$$

therefore $2^m \leq 2m^d$ or $m \leq 1 + d \log m$ (we can actually show that $m \leq d \log d$).
This implies that m cannot be too large.

11.3 Finite VC implies uniform convergence property

We now use Sauer's Lemma to prove $6 \Rightarrow 1$ of the fundamental theorem of statistical learning.
Recall: H has the **uniform convergence property** if there exists a function $m_H(\epsilon, \delta)$ such that for every distribution D over $X \times \{0, 1\}$ and $\epsilon, \delta > 0$, if $m \geq m_H(\epsilon, \delta)$ then

$$\Pr_{S \sim D^m} \left[S \text{ is } \epsilon - \text{representative of } H \text{ wrt } D \right] > 1 - \delta$$

A sample S is ϵ-representative of H wrt D if $\forall h \in H$

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

Proof. Idea: Let D be any distribution over $X \times \{0, 1\}$ and S a D-sample of size m such that $m >> VC(H)$.
For each $h \in H$, we wish to show that $|L_S(h) - L_D(h)| < \epsilon$.
If we fix h, then Hoeffding's Lemma guarantees that

$$\Pr \left[|L_S(h) - L_D(h)| > \epsilon \right] \leq 2e^{-2m\epsilon^2}$$

Thus the probability this will hold for all $h \in H$ is $\leq |H| \cdot 2e^{-2m\epsilon^2}$.
We only care about h's behaviour on S thus we may replace $|H|$ with $|\{h_{|S} \mid h \in H\}| \leq m^d$ by Sauer's Lemma thus we have

$$\Pr \left[|L_S(h) - L_D(h)| > \epsilon \right] \leq 2m^d e^{-2m\epsilon^2}$$

We can simply choose m large enough such that $2m^d e^{-2m\epsilon^2} < \delta$. □

12 February 28, 2019

12.1 Fundamental theorem of PAC learning

Theorem 12.1. H is learnable $\iff$ $VC(H) < \infty$.
Alternatively we have a quantitative version: Let H be any class of finite VC dimension. Let $m_H(\epsilon, \delta)$ be the sample size needed to learn H to accuracy $< \epsilon$ with probability $\geq 1 - \delta$. Then for some constants c_1, c_2 we have for the

realizable PAC setting

$$c_1 \frac{VC(H) + \log\left(\frac{1}{\delta}\right)}{\epsilon} \leq m_H(\epsilon, \delta) \leq c_2 \frac{VC(H) + \log\left(\frac{1}{\delta}\right)}{\epsilon}$$

and for the **agnostic PAC setting**

$$c_1 \frac{VC(H) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2} \leq m_H^A(\epsilon, \delta) \leq c_2 \frac{VC(H) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

Remark 12.1. As we require a better accuracy ($\epsilon \to 0$) or higher probability ($\delta \to 0$) then we require more samples.

Remark 12.2. As we increase our class complexity/size e.g. adding more variates to our model in H then $VC(H)$ increases requiring a larger sample size.

As we increase the complexity of our class H we have a lower $\min_{h \in H} L_D(h)$, but we require more examples. This is the **bias complexity tradeoff**.
Recall in the NFL theorem we showed (in the realizable case) that $m_H(\epsilon = \frac{1}{8}, \delta = \frac{1}{7}) \geq \frac{d}{2}$ with a *uniform distribution*. How does this bound depend on ϵ, δ in general? How does this bound change for the agnostic setup?

Dependence of ϵ We show the inverse relation with ϵ.

Let $VC(H) = d$ and let $W \subseteq X$ be a set of size d shattered by H.

Define a probability distribution over W as follows: pick some $x_0 \in W$ for every $x \in X$. We let

$$D_\epsilon(x) = \begin{cases} 1 - \epsilon & \text{if } x = x_0 \\ \frac{\epsilon}{d-1} & \text{if } x \in W \setminus \{x_0\} \\ 0 & \text{if } x \notin W \end{cases}$$

By the NFL argument, a sample S that misses $\geq \frac{1}{2}$ of the points in $W \setminus \{x_0\}$ then for every learner and some $h \in H$ the learner will have an expected error $\frac{1}{2}$ on every point in $W \setminus S$ i.e. an expected total error of

$$L_{(D,h)}(A(S)) \geq \frac{1}{4}\epsilon$$

since the total weight on our points in $W \setminus \{x_0\}$ is ϵ.

To get an expected error $< \frac{1}{4}\epsilon$ we require an S that hits $W \setminus \{x_0\}$ at least $\frac{d-1}{2}$ times. A sample of size m is expected to hit $W \setminus \{x_0\}$ $m \cdot \epsilon$ times. Therefore

$$m(\frac{\epsilon}{4}, \cdot) \geq \frac{d-1}{2\epsilon}$$

Agnostic dependence on ϵ We sketch why in the agnostic case we have a $\frac{1}{\epsilon^2}$ relationship.

We make an analogy with flipping coins. Suppose our task to predict heads or tails.

For an unbiased coin $\min_{h \in H} L_D(h) = 0.5$.

However suppose we have a biased coin where either $P(\text{heads}) = \frac{1}{2} + \epsilon, P(\text{tails}) = \frac{1}{2} - \epsilon$ OR $P(\text{heads}) = \frac{1}{2} - \epsilon, P(\text{tails}) = \frac{1}{2} + \epsilon$. In this case $\min_{h \in H} L_D(h) = \frac{1}{2} - \epsilon$.

The best learner is still ERM, and we can show that in order to estimate within ϵ we require more tosses where $m_H^A(\epsilon, \cdot) \propto \frac{1}{\epsilon^2}$.

12.2 Issues with Agnostic PAC

While the agnostic PAC model is well understood it is not practically satisfactory.
Consider a class H of half spaces. We note from agnostic PAC we can guarantee $L_D(A(S)) \leq \min_{\text{half spaces } h} L_D(h)+\epsilon$ with a probability $\geq 1-\delta$.
However, we note that $\min_{\text{half spaces } h} L_D(h)$ may be very high relative to the true error (class with all functions): that is we may have **very high bias** in our class.
How do we overcome this high bias in practice? We might use a class with very large VC dimension i.e. $VC(H) = \infty$ e.g. deep neural networks.