The Fundamental Theorem

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

PAC Learnability

We have seen that $\mathcal{H}$ is

- PAC learnable if $\mathcal{H}$ is finite
- *not* PAC learnable if $VC(\mathcal{H}) = \infty$

Today we will characterize exactly what it takes to be PAC learnable:

> *$\mathcal{H}$ is PAC learnable if and only if $VC(\mathcal{H})$ is finite*

This is known as the *fundamental theorem*.

Moreover, we will provide bounds

- on sample complexity
- and error

for hypothesis classes of finite VC complexity

- also known as classes of *small effective size*.

By Bad Samples

We already have seen a few of such proofs

- proving that finite hypothesis sets are PAC learnable

They all have the same main idea

- prove that the probability of getting a ‘bad’ sample is small

Not surprisingly, that is what we’ll do again

But first we’ll discuss (and prove) a technical detail which we’ll need in our proof

- Jensen’s inequality

Convex Functions

Jensen's inequality – in as far as we need it – is about expectations and convex functions. So we first recall what a convex function is.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* iff

- for all $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in [0, 1]$
- we have that

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

When $n = 1$, i.e., $f : \mathbb{R} \to \mathbb{R}$, this means that if we draw the graph of f and choose two points on that graph, the line that connects these two points is always above the graph of f.

Convex Examples

With the intuition given it is easy to see that, e.g.,

- $x \to |x|$,
- $x \to x^2$ and
- $x \to e^x$

are convex functions; with a little high school math, you can, of course, also prove this

If you draw the graph of $x \to \sqrt{x}$ or $x \to \log x$,

- you'll see that if you connect two points by a line, this line is always under the graph

Functions for which

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

are known as *concave* functions

Larger Sums

If we have $\lambda_1, \ldots, \lambda_m \in [0,1] : \sum_{i=1}^{m} \lambda_i = 1$, natural induction proves that for $x_1, \ldots, x_m$ we have

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$$

At least one of the $\lambda_i > 0$, say, λ_1. then we have

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_1 x_1 + \sum_{i=2}^{n+1} \lambda_i x_i\right) \\ &= f\left(\lambda_1 x_1 + (1-\lambda_1)\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} x_i\right) \\ &\leq \lambda_1 f(x_1) + (1-\lambda_1) f\left(\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} x_i\right) \\ &\leq \lambda_1 f(x_1) + (1-\lambda_1)\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} f(x_i) = \sum_{i=1}^{n+1} \lambda_i f(x_i) \end{aligned}$$

Jensen's Inequality

A special case of the previous result is when all the $\lambda_i = \frac{1}{m}$ then we have:

$$f\left(\sum_{i=1}^{m} \frac{x_i}{m}\right) \leq \sum_{i=1}^{m} \frac{f(x_i)}{m}$$

That is, the value of f at the average of the x_i is smaller than the average of the $f(x_i)$.

The average is an example of an expectation. Jensen's inequality tells us that the above inequality holds for the expectation in general, i.e., for a convex f we have

$$f\left(\mathbb{E}(X)\right) \leq \mathbb{E}(f(X))$$

We already saw that $x \rightarrow |x|$ is a convex function.

- the same is true for taking the supremum

This follows from the fact that taking the supremum is a monotone function:

$$A \subset B \rightarrow \sup(A) \leq \sup(B)$$

Proof by Uniform Convergence

To prove the fundamental theorem, we prove that classes of small effective size have the uniform convergence property.

- which is sufficient as we have seen that classes with the uniform convergence property are agnostically PAC learnable

Recall:

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* wrt domain Z and loss function l if

- there exists a function $m_{\mathcal{H}}^{UC} : (0, 1)^2 \to \mathbb{N}$
- such that for all $(\epsilon, \delta) \in (0, 1)^2$
- and for any probability distribution $\mathcal{D}$ on Z

If D is an i.i.d. sample according to $\mathcal{D}$ over Z of size $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$. Then D is ϵ-representative with probability of at least $1 - \delta$.

To Prove Uniform Convergence

Now recall that D is ϵ-representative wrt Z, $\mathcal{H}$, l and $\mathcal{D}$ if

$$\forall h \in \mathcal{H} : |L_D(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

Hence, we have devise a bound on $|L_D(h) - L_{\mathcal{D}}(h)|$ that is for almost all $D \sim \mathcal{D}^m$ small.
Markov's inequality (lecture 2) tells us that

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

So, one way to prove uniform convergence is by considering

$$\mathbb{E}_{D \sim \mathcal{D}^m} |L_D(h) - L_{\mathcal{D}}(h)|$$

Or, more precisely since it should be small for all $h \in \mathcal{H}$:

$$\mathbb{E}_{D \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} |L_D(h) - L_{\mathcal{D}}(h)| \right)$$

The supremum as $\mathcal{H}$ may be infinite and a maximimum doesn't have to exist

The First Step

The first step to derive a bound on

$$\mathbb{E}_{D\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|L_D(h)-L_{\mathcal{D}}(h)|\right)$$

is to recall that $L_{\mathcal{D}}(h)$ is itself defined as the expectation of the loss on a sample, i.e.,

$$L_{\mathcal{D}}(h)=\mathbb{E}_{D'\sim\mathcal{D}^m}\left(L_{D'}(h)\right)$$

So, we want to derive a bound on

$$\mathbb{E}_{D\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|\mathbb{E}_{D'\sim\mathcal{D}^m}(L_D(h)-L_{D'})|\right)$$

We can manipulate this expression further using Jensen's inequality

By Jensen

By Jensen's inequality we firstly have:

$$|\mathbb{E}_{D'\sim\mathcal{D}^m}(L_D(h) - L_{D'}(h))| \leq \mathbb{E}_{D'\sim\mathcal{D}^m}|L_D(h) - L_{D'}(h)|$$

And secondly we have:

$$\sup_{h\in\mathcal{H}} (\mathbb{E}_{D'\sim\mathcal{D}^m}|L_D(h) - L_{D'}(h)|) \leq \mathbb{E}_{D'\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}} |L_D(h) - L_{D'}(h)|\right)$$

Plugging in then gives us:

$$\sup_{h\in\mathcal{H}} (|\mathbb{E}_{D'\sim\mathcal{D}^m}(L_D(h) - L_{D'}(h))|) \leq \mathbb{E}_{D'\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}} |L_D(h) - L_{D'}(h)|\right)$$

Using this in the result of the first step gives us the second step

Second Step

Combining the result of the first step with the result on the previous page, we have:

$$\mathbb{E}_{D\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|L_D(h)-L_{\mathcal{D}}(h)|\right)\leq\mathbb{E}_{D,D'\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|L_D(h)-L_{D'}(h)|\right)$$

By definition, the right hand side of this inequality can be rewritten to:

$$\mathbb{E}_{D,D'\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\left|\sum_{i=1}^{m}(l(h,z_i)-l(h,z_i'))\right|\right)\right)$$

with $z_i \in D$ and $z_i' \in D'$ and both D and D' are i.i.d samples of size m sampled according to the distribution $\mathcal{D}$

An Observation

Both D and D' are i.i.d samples of size m

- it could be that the D and D' we draw today
- are the D' and D we drew yesterday

that is

- a z_i of today was a z_i' yesterday
- an a z_i' of today was a z_i yesterday

If we have this – admittedly highly improbable – coincidence

- a term $(l(h, z_i) - l(h, z_i'))$ of today
- was $-(l(h, z_i) - l(h, z_i'))$ yesterday because of the switch
- and the expectation doesn't change!

This is true whether we switch 1, 2, or all elements of D and D'.
That is, for every $\sigma \in \{-1, 1\}^m$:

$$\mathbb{E}_{D,D' \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \left| \sum_{i=1}^{m} (l(h, z_i) - l(h, z_i')) \right| \right) \right)$$

$$= \mathbb{E}_{D,D' \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (l(h, z_i) - l(h, z_i')) \right| \right) \right)$$

Observing Further

Since this equality holds for any $\sigma \in \{-1, 1\}^m$, it also holds if we sample a vector from $\{-1, 1\}^m$. So, also if we sample each $-1/+1$ entry in the vector at random under the uniform distribution, denoted by $U_{\pm}$. That is,

$$\mathbb{E}_{D,D' \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \left| \sum_{i=1}^{m} (l(h, z_i) - l(h, z_i')) \right| \right) \right)$$

$$= \mathbb{E}_{\sigma \sim U^m_{pm}} \mathbb{E}_{D,D' \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (l(h, z_i) - l(h, z_i')) \right| \right) \right)$$

And since $\mathbb{E}$ is a linear operation, this equals

$$\mathbb{E}_{D,D' \sim \mathcal{D}^m} \mathbb{E}_{\sigma \sim U^m_{\pm}} \left(\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \left| \sum_{i=1}^{m} \sigma_l (l(h, z_i) - l(h, z_i')) \right| \right) \right)$$

From Infinite to Finite

In computing the inner expectation of

$$\mathbb{E}_{D,D'\sim\mathcal{D}^m}\mathbb{E}_{\sigma\sim U_\pm^m}\left(\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\left|\sum_{i=1}^m \sigma_i(l(h,z_i)-l(h,z_i'))\right|\right)\right)$$

both D and D' are fixed, they vary for the outer expectation computation

- just like nested loops

So, if we denote $C = D \cup D'$, then we do not range over the (possibly) infinite set $\mathcal{H}$, but just over the finite set $\mathcal{H}_C$. That is

$$\mathbb{E}_{\sigma\sim U_\pm^m}\left(\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\left|\sum_{i=1}^m \sigma_i(l(h,z_i)-l(h,z_i'))\right|\right)\right)$$
$$= \mathbb{E}_{\sigma\sim U_\pm^m}\left(\max_{h\in\mathcal{H}_C}\left(\frac{1}{m}\left|\sum_{i=1}^m \sigma_i(l(h,z_i)-l(h,z_i'))\right|\right)\right)$$

Step 3

For $h \in \mathcal{H}_C$ define the random variable θ_h by

$$\theta_h = \frac{1}{m}\sum_{i=1}^{m} \sigma_i(l(h, z_i) - l(h, z_i'))$$

Now note that

- $\mathbb{E}(\theta_h) = 0$
- θ_h is the average of independent variables, taking values in $[-1, 1]$

Hence, we can apply Hoeffding's inequality. Hence, $\forall \rho > 0$

$$\mathbb{P}(|\theta_h| > \rho) \leq 2e^{-2m\rho^2}$$

Applying the union bound we have:

$$\mathbb{P}(\bigvee_{h\in\mathcal{H}_C} |\theta_h| > \rho) \leq 2|\mathcal{H}_C|e^{-2m\rho^2}$$

Which is equivalent to:

$$\mathbb{P}(\max_{h\in\mathcal{H}_C} |\theta_h| > \rho) \leq 2|\mathcal{H}_C|e^{-2m\rho^2}$$

A Useful Lemma

We now have a bound on $\mathbb{P}(\max_{h \in \mathcal{H}_C} |\theta_h| > \rho)$

- but we need a bound on $\mathbb{E}(\max_{h \in \mathcal{H}_C} |\theta_h|)$

To make this step, there is a useful lemma.

Let X be a random variable and $x \in \mathbb{R}$ If

- there exists an $a > 0$ and $b > e$ such that
- $\forall t \geq 0 : \mathbb{P}(|X - x| > t) \leq 2be^{-\frac{t^2}{a^2}}$

then

$$\mathbb{E}(|X - x|) \leq a(4 + \sqrt{\log(b)})$$

Which can be proven by straightforward calculus (see Lemma A4 in the book).

Substituting ρ for t, $1/\sqrt{2m}$ for a, and $|\mathcal{H}_C|$ for b, we get a bound on the expectation

Step 4

The lemma on the previous page gives us that

$$\mathbb{P}(\max_{h \in \mathcal{H}_C} |\theta_h| > \rho) \leq 2|\mathcal{H}_C| e^{-2m\rho^2}$$

implies that

$$\mathbb{E}(\max_{h \in \mathcal{H}_C} |\theta_h|) \leq \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{2m}}$$

Now C has maximal $2m$ distinct elements

- and $\tau_{\mathcal{H}}(k)$ is the maximal size of $|\mathcal{H}_C|$ for a set C with k elements

we have:

$$\mathbb{E}(\max_{h \in \mathcal{H}_C} |\theta_h|) \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$$

Working our way back through this (long) computation we have:

$$\mathbb{E}_{D \sim \mathcal{D}^m} \left(\sup_{h \in \mathcal{H}} |L_D(h) - L_{\mathcal{D}}(h)| \right) \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$$

Step 5

Since $\sup_{h \in \mathcal{H}} |L_D(h) - L_{\mathcal{D}}(h)|$ is obviously a non-negative random variable, we can now apply Markov's inequality to get:

Let $\mathcal{H}$ be a hypothesis class. Then for any distribution $\mathcal{D}$ and for every $\delta \in (0, 1)$ with a probability of at least $1 - \delta$ over the choice of $D \sim \mathcal{D}^m$ we have for all $h \in \mathcal{H}$:

$$|L_D(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

To prove uniform convergence, we now have to show

- that there exists an m depending on ϵ and δ
- such that the right hand side is less than ϵ

Uniform Convergence

If $m > d = VC(\mathcal{H})$ we have by Sauer: $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$.
Hence,

$$|L_D(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d \log(2em/d)}}{\delta\sqrt{2m}}$$

For large enough m, $\sqrt{d \log(2em/d)} \geq 4$, so

$$|L_D(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d \log(2em/d)}{m}}$$

Some tedious algebra shows that this implies that

$$|L_D(h) - L_{\mathcal{D}}(h)| \leq \epsilon \text{ if}$$
$$m \geq 4\frac{2d}{(\delta\epsilon)^2} \log\left(\frac{2d}{(\delta\epsilon)^2}\right) + \frac{4d \log(2e/d)}{(\delta\epsilon)^2}$$

That is, for $\mathcal{H}$ with finite VC dimension we have uniform convergence.

The Fundamental Theorem

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0, 1\}$ with $0/1$ loss. Then the following statements are equivalent

1. $\mathcal{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathcal{H}$
6. $\mathcal{H}$ has a finite VC dimension

Our calculation leading up to this theorem – its proof, actually – gives us a bound on the sample complexity. This bound is not as good as possible. I'll give you better bounds, without proof (it depends on yet another interesting concept: ϵ-nets).

The Fundamental Theorem: the Bounds

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0, 1\}$ with $0/1$ loss. Then

1. $\mathcal{H}$ has the uniform convergence property with sample complexity
$$m_{\mathcal{H}}^{UC} = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$$
2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity
$$m_{\mathcal{H}} = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$$
3. $\mathcal{H}$ is PAC learnable with sample complexity
$$m_{\mathcal{H}} = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$$

Polynomial Sample Complexity

When Valiant introduced PAC learning he required that

- the sample complexity should be polynomial in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$.

The bounds on the sample complexity we just discussed show that this requirement is not necessary

- PAC learnability implies a polynomial sample complexity (under the conditions of the theorem)

Hence there is no reason to stipulate this requirement

Valiant's other requirement

- the existence of a polynomial learning algorithm

of course still makes perfect sense. Non-polynomial algorithms on polynomially sized samples are still not practical.

Bounds in Terms of Growth

Analogously to the proof of the Fundamental Theorem, one can prove:

For any hypothesis space $\mathcal{H}$ (finite or infinite), for any D of size m and for any $\epsilon > 0$

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) > L_D(h) + \epsilon\right) \leq 8\tau_{\mathcal{H}}(m)e^{-m\epsilon^2/32}$$

So, with probability at least $1 - \delta$

$$\forall h \in \mathcal{H} : L_{\mathcal{D}}(h) \leq L_D(h) + \sqrt{\frac{32(\ln(\tau_{\mathcal{H}}(m) + \ln(8/\delta))}{m}}$$

For Consistent Hypotheses Only

If we restrict ourselves to hypothesis that are consistent with D only

- they make 0 errors on D
- that is $L_D(h) = 0$

we get slightly tighter bounds.

In terms of growth, with probability at least $1 - \delta$

$$L_{\mathcal{D}}(h) \leq \frac{2\log(\tau_{\mathcal{H}}(2m)) + 2\log(2/\delta)}{m}$$

In terms of the VC dimension d, with $m \geq d \geq 1$ with probability at least $1 - \delta$

$$L_{\mathcal{D}}(h) \leq \frac{2\log(2em/d) + 2\log(2/\delta)}{m}$$

Starting From Big Data

Our journey towards this Fundamental Theorem started with the analysis of Big Data. Next to serious problems such as

- the curse of dimensionality
- and the fact that Big Data makes every difference statistically significant
  - however small and pragmatically insignificant it may be

we identified the, perhaps largest, problem as

> *Big Data is too big to process*

Superlinear algorithms

- are quite soon infeasible on very large data sets

Hence, the quest we set out for

- can we sample D to make (superlinear) learning feasible?

Frequent Itemsets

To make the Big Data problem more concrete we introduced a typical data mining problem

Frequent Itemset Mining

and we noted that the A Priori algorithm

- which can be used to mine all frequent itemsets efficiently

actually applies to a far larger class of problems

Frequent Pattern Mining

Given that frequent itemset mining requires multiple scans over the database

- which can be very expensive for very large databases

the natural question was

- can we sample for frequent itemset mining?

Sampling for Frequent Itemset Mining

We discussed a paper by Toivonen, in which he showed that with a sample of size

- $n \geq \frac{1}{\epsilon^2}\left(|\mathcal{I}| + \ln\left(\frac{2}{\delta}\right)\right)$
- our estimate of the frequency of an itemset is with probability of at least $1 - \delta$ off by at most ϵ

The problem with this approach is that we

- may have false negatives: itemsets that are frequent on the database but not on the sample

We can mitigate that problem by

- lowering the threshold by $\sqrt{\frac{1}{2n} \ln \frac{1}{\mu}}$
- checking whether or not the border of our (estimated) set of frequent itemsets contains such false negatives

This gives us indirect control over the probability of false negatives

- can we get direct control?

From Itemsets to Classification

We saw that an itemset Z, or better its associated indicator function, acts as a *classifier* on D:

$$1_Z(t) = \begin{cases} 1 & \text{if } Z \subseteq t \\ 0 & \text{otherwise} \end{cases}$$

This observation allows us to go from

- unsupervised learning – what itemset mining is
- to supervised learning – what classification is

The advantage that supervised learning problems have over unsupervised ones

- is that they have objective quality measures,
- e.g., higher accuracy = better model

Exploiting such measures might give us a better grip on sampling

From Classification

We started this quest with the analysis of a simple classification problem (finite hypothesis class and the realizability assumption). From this analysis, we proved:

Let $\mathcal{H}$ be a finite hypothesis space. Let $\delta \in (0, 1)$, let $\epsilon > 0$ and let $m \in \mathbb{N}$ such that

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

Then, for any labelling function f and distribution $\mathcal{D}$ for which the realizability assumption holds, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample D of size m we have that for every ERM hypothesis h_D:

$$L_{\mathcal{D},f}(h_D) \leq \epsilon$$

To PAC learning

Then we turned this result upside down and made it into the definition of

- Probably Approximately Correct learning

Learning problems that give almost always reasonably good results

- with (polynomial) sized data sets

And that last point is very important in the Big Data context

- as was discussed in the first two lectures

At first we limited ourselves to the realizable case

- colloquially: the hypothesis set contains the true hypothesis

and an immediate consequence of our previous theorem was

- finite hypothesis classes are PAC learnable

In Full Generality

Then we loosened the requirements

- firstly the realizability assumption
- secondly allowing for arbitrary loss functions

To arrive at the general definition of PAC Learning:

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $l : Z \times \mathcal{H} \to \mathbb{R}_+$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

- for every $\epsilon, \delta \in (0,1)$
- for every distribution $\mathcal{D}$ over Z
- when running A on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. samples generated by $\mathcal{D}$
- A returns a hypothesis $h \in \mathcal{H}$ such that with probability at least $1 - \delta$

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

Desirable, but Attainable?

Clearly, PAC learnability is a desirable property

- you have the guarantee that you almost always get results that are almost as good as it gets.

But, then the question is

- are there hypothesis sets that have this property?

We first showed that hypothesis sets that have the uniform convergence property

- on almost all (large enough) data sets your estimate of the loss of a hypothesis is close to the true loss

are PAC learnable (in the general sense). And, with that result we proved that

- finite hypothesis sets are PAC learnable

Finite can be very large

- and you can always approximate your favourite infinite classes with a finite one

But, then your choice of a finite class has a direct influence on the result you achieve.

Infinite Classes

So, it would be nice if we could PAC learn infinitely large hypothesis classes. But then came our first negative result

- the No Free Lunch theorem says: there are infinitely large hypotheses classes you can not PAC learn
- you would need infinite data samples
  - even larger than Big Data!

We then first showed that

- the infinite set of thresholds functions can be PAC learned in the general sense
  - we had already seen that this class could not be learned in the more restricted realizable case
  - so, that in itself is already a relief

We then compared the proof of the No Free Lunch theorem

- with the threshold classifiers

And, from that comparison we came up with

- with the VC dimension

VC Dimension

The VC dimension of a hypothesis class $\mathcal{H}$ is the size of the largest (finite) set of data points that $\mathcal{H}$ shatters, that is, it is the size of the largest $C \subset X$ such that

$$|\mathcal{H}_C| = 2^{|C|}$$

The proof of the No Free Lunch theorem showed that if the size of our sample D is such that

$$m \leq 2VC(\mathcal{H})$$

then it is may be hard to find a good $h \in \mathcal{H}$

In other words, a finite VC dimension tells us

- that we can distinguish between the different hypotheses relatively quickly
  - from a modestly sized sample

Growth

This ability of the VC dimension is further illustrated by the growth function, defined by

$$\tau_{\mathcal{H}}(m) = \max_{C \subset X : |C| = m} |\{f(c_1), \ldots, f(c_m)\}_{f \in \mathcal{H}}|$$

For $m \leq d = VC(\mathcal{H})$, we have $\tau_{\mathcal{H}}(m) = 2^m$.

More in general, we have by Sauer's Lemma that if $d = VC(\mathcal{H}) \leq \infty$:

- $\forall m : \tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$
- if $m \geq d : \tau_{\mathcal{H}}(m) < (em/d)^d$

The growth function starts of as an exponential function, but from d on forwards it is a polynomial function.

Hence, the expectation

- perhaps I should say hope

that infinite hypothesis classes with a finite VD dimension will be PAC learnable

The Fundamental Theorem

The Fundamental Theorem tells us that our expectation was correct

- Hypothesis classes are PAC learnable iff they have a finite VC dimension
- moreover the sample size you need is polynomial in the parameters that matter
  - in d, $1/\delta$ (in fact $\log(1/\delta)$) and $1/\epsilon$

In other words, we appear to have ended our quest

> *as long as we use hypothesis classes with a finite VC dimension we can conquer the problem of Big Data by sampling*

So the question is now:

- can we use PAC learning to derive sample bounds for frequent itemset mining?

We'll study that next, but it is not the end of the story

There is More

The concept of PAC learning requires

- a sample size that holds for all $h \in \mathcal{H}$ at the same time
- and that we can get arbitrarily close to the truth

What if we relax those requirements

- would that allow us to battle Big Data with a larger class of hypotheses sets?

The answer,

- somewhat surprisingly

is: not really.

This does not have direct ramifications for our frequent itemset mining problem

- but it tells us that PAC learning is a reasonable way to battle the problem of induction