CSC 411 Lecture 23-24: Learning theory *

Ethan Fetaya, James Lucas and Emad Andrews

University of Toronto

Today

- Generalization bounds
- PAC learning
- VC dimension

Hypothesis set

- Talking about supervised learning and binary classification.
- Data $(\mathbf{x}, t)$ is distributed according to an unknown distribution $\mathcal{D}$
- We want to return a function h that minimizes expected loss (risk) $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}[\ell(h(\mathbf{x}), t)]$
- Cannot minimize the risk as $\mathcal{D}$ is unknown, so we can minimize the empirical risk $L_S^N(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(h(\mathbf{x}^{(i)}), t^{(i)})$
- Minimizing over all functions cannot work so we restrict to a subset hypothesis set
  - Linear classifiers
  - Neural networks (fixed architecture)
  - etc.
- Main goal: Define complexity of $\mathcal{H}$ and use it to say if $\mathcal{H}$ is learnable and how many examples do we need.
  - Learnable means we can find the best function in the hypothesis space

PAC learning

- We are given m samples $\{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$ and a hypothesis space $\mathcal{H}$ and we wish to return $h \in \mathcal{H}$ minimizing $L_\mathcal{D}(h) = \mathbb{E}[\ell(h(x), y)]$.
- Problem 1: Cannot find the exact minimizer after seeing only a sample of the data (or even if we had perfect knowledge). Can only expect an **approximate** solution: $L_\mathcal{D}(h) \leq \min_{h \in \mathcal{H}} L_\mathcal{D}(h) + \epsilon$
- Problem 2: We depend on a random sample. There is always a chance we get a bad sample that doesn't represent $\mathcal{D}$. Our algorithm can only be **probably** correct: there is always some probability δ that we are completely wrong.
- We wish to find a **probably approximately correct (PAC)** hypothesis.

PAC learning

Definition (PAC learnable)

A hypothesis class $\mathcal{H}$ is PAC learnable, if there exists a learning algorithm A, satisfying that for any $\epsilon > 0$ and $\delta \in (0,1)$ there exist $\mathfrak{M}(\epsilon, \delta) = poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ such that for i.i.d samples $S^m = \{(x_i, y_i)\}_{i=1}^m$ drawn from any distribution $\mathcal{D}$ and $m \geq \mathfrak{M}(\epsilon, \delta)$ the algorithm returns a hypothesis $A(S^m) \in \mathcal{H}$ satisfying

$$P_{S^m \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon) < \delta$$

- If you have enough samples you can guarantee a probably approximately correct answer.
  - There is a number of samples needed doesn't depend on $\mathcal{D}$.
- Next will show that if $L_S(h) \approx L_{\mathcal{D}}(h)$ for all h then the empirical risk minimization (ERM) is a PAC learning algorithm.

Uniform Convergence

Definition (Uniform convergence)

A hypothesis class $\mathcal{H}$ has the uniform convergence property, if for any $\epsilon > 0$ and $\delta \in (0,1)$ there exist $\mathfrak{M}(\epsilon, \delta) = poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ such that for any distribution $\mathcal{D}$ and $m \geq \mathfrak{M}(\epsilon, \delta)$ i.i.d samples $S^m = \{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$ with probability at least $1 - \delta$, $|L_S^m(h) - L_{\mathcal{D}}(h)| < \epsilon$ for all $h \in \mathcal{H}$.

- For a single h, law of large numbers says $L_S^m(h) \overset{m \to \infty}{\to} L_{\mathcal{D}}(h)$
- For loss bounded by 1 the Hoeffding inequality states

$$P(|L_S^m(h) - L_{\mathcal{D}}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 m}$$

- The difficulty is to bound all the $h \in \mathcal{H}$ uniformly.

PAC by uniform convergence

Theorem (PAC by uniform convergence)

If $\mathcal{H}$ has the uniform convergence with $\mathfrak{M}(\epsilon, \delta)$ then $\mathcal{H}$ is PAC learnable with the ERM algorithm and $\mathfrak{M}(\frac{\epsilon}{2}, \delta)$ samples.

Proof.

By uniform convergence: With probability at least $1 - \delta$ for all $h \in \mathcal{H}$, $|L_S(h) - L_\mathcal{D}(h)| \leq \frac{\epsilon}{2}$.

Define $h_{ERM} = \arg\min_{h \in \mathcal{H}} L_S(h)$ and $h^* = \arg\min_{h \in \mathcal{H}} L_\mathcal{D}(h)$.

$L_\mathcal{D}(h_{ERM}) \leq L_S(h_{ERM}) + \frac{\epsilon}{2} \leq L_S(h^*) + \frac{\epsilon}{2} \leq L_\mathcal{D}(h^*) + \epsilon$ □

- This shows it is sufficient to show uniform convergence.
- For binary classification it is necessary as well, not true more generally.

Finite hypothesis space

A first simple example of PAC learnable spaces - finite hypothesis spaces.

Theorem (uniform convergence for finite $\mathcal{H}$)

Let $\mathcal{H}$ be a finite hypothesis space and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a bounded loss function, then $\mathcal{H}$ has the uniform convergence property with $\mathfrak{M}(\epsilon, \delta) = \frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2}$ and is therefore PAC learnable by the ERM algorithm.

Proof .

For any $h \in \mathcal{H}$, $\ell(h(x_1), y_1), ..., \ell(h(x_m), y_m)$ are i.i.d random variables with expected value $L_{\mathcal{D}}(h)$.

According to the Hoeffding inequality,

$$P(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 m} \leq 2e^{-2\epsilon^2 \mathfrak{M}(\epsilon,\delta)} = \frac{\delta}{|\mathcal{H}|} \tag{1}$$

□

Proof (Cont.)

We can now use the union bound: For all events $A_1, ..., A_n$

$$P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \tag{2}$$

For all $h \in \mathcal{H}$ define A_h as the event that $|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon$. By equation 1 we know that $P(A_h) \leq \frac{\delta}{|\mathcal{H}|}$. With equation 2 we can conclude

$$P(\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = P(\cup_{h \in \mathcal{H}} A_h) \leq \sum_{h \in \mathcal{H}} P(A_h)$$

$$\leq \sum_{h \in \mathcal{H}} \frac{\delta}{|\mathcal{H}|} = \delta$$

□

Infinite hypothesis sets

- We have seen that finite hypothesis class can be learned, but what about infinite ones like linear predictors?
- We can discretize (after all we are working on a finite precision machines), but this is not a great solution.
  - If we move from float to double does generalization changes?
  - The union bound is very suboptimal as similar hypothesis will fail on similar samples
- The solution is the check how many effective hypothesis there are on a sample of size m
- We will restrict ourselves to binary classification with $0-1$ loss.

Growth function

Definition

Let $\mathcal{H}$ be a set of function from $\mathcal{X}$ to $\{\pm 1\}$ and let $C \subset \mathcal{X}$ be a subset of the input space. We denote by $\mathcal{H}|_C$ all the function that can be derived by restricting functions in $\mathcal{H}$ to C.

$$\mathcal{H}|_C = \{h|_C : C \to \{\pm 1\} : \ h \in \mathcal{H}\}$$

Definition (Growth function)

The growth function of $\mathcal{H}$, $\Pi_{\mathcal{H}}(m)$ is the size of the largest restriction of $\mathcal{H}$ to a set of size m.

$$\Pi_{\mathcal{H}}(m) = \max\{|\mathcal{H}|_C| : \ C \subset \mathcal{X}, \, |C| = m\}$$

$\Pi_{\mathcal{H}}$ measures the maximum number of ways to label your inputs

Growth function - Examples

Notice that $\Pi_{\mathcal{H}}(m) \leq 2^m$.

1. $\mathcal{H} = 2^{\mathcal{X}}$ for infinite $\mathcal{X}$, $\Pi_{\mathcal{H}}(m) = 2^m$.
2. For finite $\mathcal{H}$, $\Pi_{\mathcal{H}}(m) \leq |\mathcal{H}|$
3. For $\mathcal{H} = \{h_a(x) = sign(x - a), a \in \mathbb{R}\}$, $\Pi_{\mathcal{H}}(m) = m + 1$.
4. For $\mathcal{H} = \{h_a^{\pm}(x) = sign(\pm x - a), a \in \mathbb{R}\}$, $\Pi_{\mathcal{H}}(m) = 2m$.

As we can see, even for an infinite hypothesis set it is possible that $\Pi_{\mathcal{H}}(m) \ll 2^m$.

Uniform Convergence Bound

We can now state the main theorem that shows the importance of the growth function.

Theorem (Uniform convergence bound)

Let $\mathcal{H}$ be a hypothesis set of $\{\pm 1\}$ valued functions and ℓ be the $0-1$ loss, then for any distribution $\mathcal{D}$ on $\mathcal{X} \times \{\pm 1\}$, any $\epsilon > 0$ and positive integer m, we have

$$P_{S\sim\mathcal{D}^m}\left(\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\right) \leq 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)$$

Immediate corollary - if $\Pi_{\mathcal{H}}(m)$ grows sub-exponentially then $\mathcal{H}$ is PAC learnable.

Proof sketch

- This is not a simple proof, we will just go over the main idea.
- We will discretize and use the union bound + Hoeffding inequality.
- The trick is to have the "right" discretization.
- We want to look only at $\Pi_{\mathcal{H}}(m)$ classifiers but it isn't that straightforward
  - The test loss depends on infinite number of data-points
  - $\mathcal{H}|_C$ on the training input $\mathbf{x}_1, ..., \mathbf{x}_N$.
- We start by showing the we can replace $L_{\mathcal{D}}$ by $L_{\tilde{S}}$ - the error on another m independent "test" samples.
- The next step to show you can fix the samples, and look at the probability of permuting between the train and test sets.
- Then we can use the union bound and Hoeffding on this reduced case.

VC dimension

- In order to prove uniform convergence, and therefore PAC learnability, it is enough to show that the growth function is sub-exponential.
- As we will see, the behavior $\Pi_{\mathcal{H}}(m)$ is greatly controlled by a single parameter - the VC dimension.

Definition (Shattering)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$. We say that $\mathcal{H}$ shatters $C \subset \mathcal{X}$ if $\mathcal{H}|_C = 2^C$.

Definition (VC-dimension)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$. The VC-dimension of $\mathcal{H}$ is the size of the largest finite set that $\mathcal{H}$ shatters (or ∞ if there is no maximum).

So $VC(\mathcal{H}) = d \Leftrightarrow \Pi_{\mathcal{H}}(d) = 2^d \wedge \Pi_{\mathcal{H}}(d+1) < 2^{d+1}$

VC dimension examples

1. $\mathcal{H} = 2^{\mathcal{X}}$ for infinite $\mathcal{X}$, $\Pi_{\mathcal{H}}(m) = 2^m$ Therefore $VC(\mathcal{H}) = \infty$.
2. For finite $\mathcal{H}$, $\Pi_{\mathcal{H}}(m) \leq |\mathcal{H}| \Rightarrow VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$
3. For $\mathcal{H} = \{h_a(x) = sign(x - a),\ a \in \mathbb{R}\}$, $\Pi_{\mathcal{H}}(m) = m + 1 \Rightarrow VC(\mathcal{H}) = 1$.
4. For $\mathcal{H} = \{h_a^{\pm}(x) = sign(\pm x - a),\ a \in \mathbb{R}\}$, $\Pi_{\mathcal{H}}(m) = 2m \Rightarrow VC(\mathcal{H}) = 2$.
5. The class of axis aligned rectangles $h_{(x_1,x_2,y_1,y_2)}(x, y) := 1 \Leftrightarrow x_1 < x < x_2 \wedge y_1 < y < y_2$ for $x_1 < x_2,\ y_1 < y_2$.
   - It is easy to find a 4 element set that $\mathcal{H}$ shatters. To show $VC(\mathcal{H}) = 4$, we need to show it cannot shatter any set of five elements.
   - This can be done by observing that one point is always in the convex hall of the other points that cannot get zero if all others are one.
6. The class of convex sets in the plane $h_C(x, y) := 1 \Leftrightarrow (x, y) \in C$ for a convex set $C \subset \mathbb{R}^2$. We can see that $VC(\mathcal{H}) = \infty$ by arranging points on the circle.

VC of linear classifiers

- What is the VC dimension of linear classifiers: $\mathcal{H}_d = \{h_w(x) = sign(\langle w, x \rangle) : \ w \in \mathbb{R}^d\}$?
- To show that $VC(\mathcal{H}_d) = d$ it is enough to prove the following lemma:

Lemma

The vectors $x_1, ..., x_k \in \mathbb{R}^d$ are shattered by $\mathcal{H}_d$ if and only if they are linearly independent.

Proof.

$\Rightarrow$ assume by contradiction that they are linearly dependent, so there exist some j such that $x_j = \sum_{i=1}^{j-1} \alpha_i x_i$. Any labeling y_i such that $\alpha_i y_i \geq 0$ has to have $y_j = 1$, therefore the set is not shattered - a contradiction.

$\Leftarrow$ Let X be the matrix with rows x_i^T, then the vector of labels given by any w is just $sign(X \cdot w)$. Our assumptions means that X has rank k and is therefore an onto mapping, and the set is shattered. □

- This shows that the VC dimension and the algebraic dimension are the same.

Bounding the growth function

The next step is bounding $\Pi_{\mathcal{H}}(m)$ using the VC dimension.

Theorem (Sauer Shelah)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ with VC-dimension $d < \infty$, then $\Pi_{\mathcal{H}}(m) \leq \sum_{k=1}^{d} \binom{m}{k}$

Notice that $S(m, d) = \sum_{k=0}^{d} \binom{m}{k}$ is the number of subset of size smaller or equal to d of a set of size m.

Proof.

We will show a stronger claim $|\mathcal{H}|_C| \leq |\{B \subset C : \mathcal{H} \textit{ shatters } B\}| \leq \sum_{k=0}^{d} \binom{m}{k}$.

This is done by induction on m. For $m = 1$ the claim is trivial. □

Sauer-Shelah Proof

Proof (Cont.)

Let $C = \{x_1, ..., x_{m+1}\}$ and $\tilde{C} = \{x_1, ..., x_m\}$. Each function of $\mathcal{H}|_{\tilde{C}}$ corresponds to either one function in $\mathcal{H}|_C$ if it has a unique extension, or to two function if both extensions are possible.

Define $\mathcal{F} \subset \mathcal{H}|_{\tilde{C}}$ as all the function that correspond to two functions in $\mathcal{H}|_C$, then $|\mathcal{H}|_C| = |\mathcal{H}|_{\tilde{C}}| + |\mathcal{F}|$.

From our induction hypothesis
$|\mathcal{H}|_{\tilde{C}}| \leq |\{B \subset \tilde{C} : \mathcal{H}\, shatters\, B\}| = |\{B \subset C : \mathcal{H}\, shatters\, B \wedge x_{m+1} \notin B\}|$.

For $\mathcal{F}$: $|\mathcal{F}| \leq |\{B \subset \tilde{C} : \mathcal{F}\, shatters\, B\}|$. For each such B shattered by $\mathcal{F}$, $B \cup \{x_{m+1}\}$ is shattered by $\mathcal{H}$, so
$|\mathcal{F}| \leq |\{B \subset C : \mathcal{H}\, shatters\, B \wedge x_{m+1} \in B\}|$. □

Can simplify the bound $\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d = \mathcal{O}(m^d)$.

PAC Learnability

- We can combine all our results to show that if $\mathcal{H}$ has VC-dimension $d < \infty$ then it is PAC learnable

Theorem (PAC learnability of finite VC-dimension)

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ with VC-dimension $d < \infty$, then $\mathcal{H}$ has the uniform convergence property with $\mathfrak{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{d \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$ and is therefore PAC learnable with the ERM algorithm.

- One can get (with some extra effort) a better bound without the $\ln(\frac{1}{\epsilon})$ factor.
- Can show the bound (without the $\ln(\frac{1}{\epsilon})$ factor) is tight.
- Number of samples needed scales linearly with the dimension.

Proof

Proof (sketch).

To prove uniform convergence we need to show that

$$P_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H}:\ |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq\delta\qquad\forall m\geq\mathfrak{M}(\epsilon,\delta)$$

We already showed that

$$P_{S\sim\mathcal{D}^m}\left(\exists h\in\mathcal{H}:\ |L_S(h)-L_{\mathcal{D}}(h)|\geq\epsilon\right)\leq 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)$$

Using the inequality $\Pi_{\mathcal{H}}(m)\leq\left(\frac{em}{d}\right)^d$ and the inequality $\forall\alpha,x>0:\ \ln(x)\leq\alpha x-\ln(\alpha)$, we can show (with some algebra) that $4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{\epsilon^2 m}{8}\right)\leq\delta$. □

No Free Lunch

We have seen when ML can do, what can't it do?

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

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

Main concept - Generalizing is extrapolating to new data. To do that you need to make assumptions and for problems where the assumptions don't hold you will be suboptimal. (skipping proof)

Infinite VC dimension

We will use the No-Free-Lunch theorem to show that any $\mathcal{H}$ with infinite VC dimension is not PAC learnable.

Theorem

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0,1\}$ with $VC(\mathcal{H}) = \infty$ and let the loss function be the $0-1$ loss. The hypothesis class $\mathcal{H}$ is not PAC learnable.

Proof.

Assume by contradiction that $\mathcal{H}$ is PAC learnable. Then there exists some learning algorithm A (not necessarily ERM) such that for all $\epsilon, \delta > 0$ there exists $\mathcal{M}(\epsilon, \delta)$ such that if $m > \mathcal{M}(\epsilon, \delta)$ then for all distributions $\mathcal{D}$, $P_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) < \delta$ where $h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ □

Proof.

Assume by contradiction that such algorithm exists. Pick some $\epsilon < 1/8$, $\delta < 1/7$ and $m > \mathcal{M}(\epsilon, \delta)$. Since $VC(\mathcal{H}) = \infty$ there exists some $x_1, ..., x_{2m} \in \mathcal{X}$ that $\mathcal{H}$ shatters.

From the No-Free-Lunch theorem there is a distribution $\mathcal{D}$ on $x_1, ..., x_{2m}$ (and labels) such that: There exists some $f : \mathcal{X} \to \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$ and $P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) > 1/7$.

Since the distribution is supported only by $\{x_1, ..., x_{2m}\}$ and this set is shattered by $\mathcal{H}$, this means that $L_{\mathcal{D}}(h^*) = 0$.

This finishes the proof as $P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) \geq P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) > 1/7 > \delta$. □

Fundamental Theorem of Statistical Learning

We can combine everything we did so far and get the fundamental theorem of statistical learning (binary classification):

Theorem (Fundamental Theorem of Statistical Learning)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0, 1\}$ and let the loss function be the $0 - 1$ loss. The following are equivalent:

1. *$\mathcal{H}$ has uniform convergence.*
2. *The ERM is a PAC learning algorithm for $\mathcal{H}$.*
3. *$\mathcal{H}$ is PAC learnable.*
4. *$\mathcal{H}$ has finite VC dimension.*

Proof.

1 ⇒ 2 We have seen uniform convergence implies that ERM is PAC learnable in lecture 2.

2 ⇒ 3 Obvious.

3 ⇒ 4 We just proved that PAC learnability implies finite VC dimension.

4 ⇒ 1 We proved that finite VC dimension implies uniform convergence. □

Recap

- We define the PAC learning theoretical framework.
- We have seen that the VC dimension fully determines learnability for binary classification
- The VC dimension doesn't just determine learnability, it also gives a bound on the sample complexity (which can be shown to be tight).
- Can extend to regression/multiclass classification but theory isn't as simple.
- Recent advances in neural network pose a serious challenge to ML theory as the performence is much better then the theoretical bounds predict.