Theoretical Machine Learning: Lesson 3

Department of Computer Science
Princeton University

Elad Hazan

COS 511 - Theoretical Machine Learning, Spring 2014-2015

Recap

- Statistical learning model (PAC learning):
  - without restricting hypothesis class - overfitting
  - ERM algorithm learns finite hypothesis classes (examples: apple factory (finite-precision rectangles), conjunctions)
- Agnostic learning in the statistical model
- Agnostic learning as a way for handling noise (HW1)
- Started infinite learnable hypothesis classes: half-lines, rectangles.
- Unlearnability: disjunctive formulas (DNF), restriction to hypothesis class necessary
- Today: unlearnable classes, sufficient and necessary condition for learning (VC-dimension), fundamental theorem of statistical learning.

When can't we learn?

Theorem (No Free Lunch)

Let m be a training set sampled $i.i.d$, and let $\mathcal{X}$ be a domain with at least $2m$ distinct elements. Then, for every learning algorithm A (which creates a hypothesis $A(S) : \mathcal{X} \to \{0,1\}$ from the training data), there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a concept $f : \mathcal{X} \to \{0,1\}$ such that:

1. $\text{err}_{\mathcal{D}}(f) = 0$
2. $\mathbb{E}_{S \sim \mathcal{D}^m} [\text{err}_{\mathcal{D}}(A(S))] \geq \frac{1}{4}$

Thus: "learning without restricting the hypothesis may result in overfitting".

No free lunch thm

Lemma

Let $C \subseteq \mathcal{X}$, $|C| = 2m$ and f be a function $f : C \to \{0,1\} \in \mathcal{F} = \{0,1\}^C$. Let $\mathcal{D}_f$ be the distribution over $C \times \{0,1\}$ defined by:

$$\Pr(x,y) = \begin{cases} \frac{1}{|C|} = \frac{1}{2m} & x \in C, y = f(x) \\ 0 & o/w \end{cases}$$

Then for any learning algorithm $A : (C \times \{0,1\})^m \to F$ there exists a function $f \in F$ such that $\mathbf{E}_{S \sim \mathcal{D}_f^m}\left[\mathrm{err}_{\mathcal{D}_f}(A(S))\right] \geq \frac{1}{4}$

- Lemma implies theorem, since $\underset{\mathcal{D}_f}{\mathrm{err}}(f) = \underset{(x,y)\sim\mathcal{D}_f}{\Pr}[f(x) \neq y] = 0$
- Proof by **the probabilistic method** (showing existence via probably of an event being non-zero). In our case, we show: $Q = \mathbf{E}_{f \in F}[\underset{S \sim U(C)^m}{\mathbf{E}}[\underset{\mathcal{D}_f}{\mathrm{err}}(A(S))]] \geq \frac{1}{4}$, where $U(C)$ is uniform distribution over C.

No free lunch thm

$$Q = \mathop{\mathbf{E}}_{f\in F}[\mathop{\mathbf{E}}_{S\sim U(C)^m}[\mathop{\mathrm{err}}_{\mathcal{D}_f}(A(S))]] = \mathop{\mathbf{E}}_{f\in F}[\mathop{\mathbf{E}}_{S\sim U(C)^m}[\mathop{E}_{c\sim U(C)}[I_{\{A_S(c)\neq f(c)\}}]]]$$

$$= \mathop{\mathbf{E}}_{S\sim U(C)^m}[\mathop{\mathbf{E}}_{c\sim U(C)}[\mathop{\mathbf{E}}_{f\in F}[I_{\{A_S(c)\neq f(c)\}}]]]$$

$$= \mathop{\mathbf{E}}_{S\sim U(C)^m}[\mathop{\mathbf{E}}_{c\sim U(C)}[\mathop{E}_{f\in F}[I_{\{A_S(c)\neq f(c)\}}|c\in S]\cdot \Pr[c\in S]$$

$$+ \mathop{\mathbf{E}}_{f\in F}[I_{\{A_S(c)\neq f(c)\}}|c\notin S)\cdot \Pr[c\notin S]]]$$

$$\geq \mathop{\mathbf{E}}_{S\sim U(C)^m}[\mathop{\mathbf{E}}_{c\sim U(C)}[\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot \mathop{E}_{f\in F}[I_{\{A_S(c)\neq f(c)\}}|c\notin S]]]$$

$$= \mathop{\mathbf{E}}_{S\sim U(C)^m}(\mathop{\mathbf{E}}_{c\sim U(C)}(\frac{1}{2}\cdot\frac{1}{2})) = \frac{1}{4}$$

- Therefore we have $Q \geq \frac{1}{4}$

When is a concept unlearnable?

- Complexity of possible restrictions grows with sample size without limit (or till the limit of inputs, as in DNF formulas).
- Leads to the notion of VC-dimension we have seen last lecture.
- This notion exactly characterizes what is learnable in the statistical (PAC) learning model for both finite and infinite hypothesis classes, and gives rise to tight upper bounds on the sample complexity for learning them.

VC-theory

Definition

A restriction of hypothesis class $\mathcal{H}$ to $C = \{c_1, ..., c_m\} \subseteq \mathcal{X}$ is the set of all binary vectors induced by the hypothesis of $\mathcal{H}$ on items of C.

$$\mathcal{H}_C = \{(h(c_1), ..., h(c_m))|h \in \mathcal{H}\}$$

Definition

A set $C \subseteq \mathcal{X}$ is **shattered** by hypothesis class $\mathcal{H}$ iff $\|\mathcal{H}_C\| = 2^{|C|}$.

Definition

The **VC-dimension** of $\mathcal{H}$ is the largest cardinality of a set $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.

Examples

- positive half-lines
- axis-aligned rectangles
- convex polygons in Euclidean plane.
- finite classes:

Lemma

For any finite hypothesis class $\mathcal{H}$ and arbitrary domain, $\text{VC-dim}(\mathcal{H}) \leq \log |\mathcal{H}|$.

Fundamental theorem of statistical learning

Theorem

A hypothesis class $\mathcal{H}$ is learnable if and only if $\text{VC-dim}(\mathcal{H}) < \infty$*, in which case it is PAC learnable with sample complexity*

$$m_{\mathcal{H}}(\varepsilon, \delta) = O(\frac{\text{VC-dim}(\mathcal{H})}{\varepsilon} \log \frac{1}{\delta\varepsilon})$$

and agnostically learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) = O(\frac{\text{VC-dim}(\mathcal{H})}{\varepsilon^2} \log \frac{1}{\delta})$$

Fundamental theorem of statistical learning - proof outline

Definition

The growth function of $\mathcal{H}$ is defined by $\tau_{\mathcal{H}} : \mathcal{N} \mapsto \mathcal{N}$:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X}\ ,\ |C|=m} |\mathcal{H}_C|$$

Let $d = \text{VC-dim}(\mathcal{H})$. Two main parts:

- Sauer'-Shelah lemma: even though $\mathcal{H}$ can be large/infinite, when restricting it to $C \subseteq \mathcal{X}$ we have $\tau_{\mathcal{H}}(m) = O(m^d)$
- The generalization error as a function of the training set size $m = |S|$ behaves as

$$\frac{\log \tau(m)}{m}$$

Part 1: Sauer-Shelah Lemma

Lemma

The growth function of $\mathcal{H}$ satisfies:

$$\tau_H(m) \leq \left\{ {m \atop d} \right\} \equiv \sum_{i=0}^{d} \binom{m}{i} \overset{m>d}{\leq} \left(\frac{em}{d}\right)^d$$

Proof: by induction on $m + d$. *Base:* $m + d = 0$.
Claim: If $|C| = 0$, then $|\mathcal{H}_C| \leq 1$.

Proof.

Suppose we have more than one hypothesis in $\mathcal{H}_C$. Then, there exists x in X s.t. $h_1(x) \neq h_2(x)$. Therefore, if we chose $C = \{x\}$ then $\mathcal{H}$ shatters C, and thus $VCdim(\mathcal{H}) > 1$. Contradicting the assumption that $d = 0$. □

Part 1: Sauer-Shelah Lemma - cont.

Inductive Step: assume correctness for $m + d \leq k$ Define:

1. $C' := C/\{x\}$ thus $|C'| = m - 1$
2. $\mathcal{H}_1 := \mathcal{H}_{C'}$
3. $\mathcal{H}_2 := \{h \in \mathcal{H}_C | h(x) = 1 \wedge \exists h' \in \mathcal{H}_C$ s.t. $h'(x) = 0 \wedge h(y) = h'(y)\ \forall y \in C'\}$

We're going to apply induction to $\mathcal{H}_1, \mathcal{H}_2$. But first, **Claim:**

$$|\mathcal{H}_c| = |\mathcal{H}_1| + |\mathcal{H}_2|$$

proof of Claim.

Let $h \in \mathcal{H}_{C'}$ and let $\tilde{h} \in \mathcal{H}_C$ s.t. $\tilde{h}(y) = h(y)$ for all y in C'.
Q: How many $\tilde{h}$ does exist? □

proof of Claim cont.

A: either one or two:
Case 1: There exist one $\tilde{h}$. Thus, $\tilde{h}(x) = 0$ or $\tilde{h}(x) = 1$ thus $\tilde{h}$ not in $\mathcal{H}_2$. So we count this $\tilde{h}$ only in $\mathcal{H}_1$.
Case 2: There exist two $\tilde{h}$; $\tilde{h}_1$ and $\tilde{h}_2$. *w.l.o.g*, $\tilde{h}_1(x) = 0$ and $\tilde{h}_2(x) = 1$. Thus, by definition: $\tilde{h}_1$ projection over C' is in $\mathcal{H}_1$ and $\tilde{h}_2$ in $\mathcal{H}_2$. So again, we count both of them.
This proves that $|\mathcal{H}_c| = |\mathcal{H}_1| + |\mathcal{H}_2|$. □

Observation 1: Applying induction assumption, since $(m-1)+d < m-d$,

$$|\mathcal{H}_1| = |\mathcal{H}_{C'}| \leq \begin{Bmatrix} m-1 \\ d \end{Bmatrix}$$

Claim.

$$VCdim(\mathcal{H}_2) \leq VCdim(\mathcal{H}_c) - 1$$

□

Proof of Claim.

Suppose $\mathcal{H}_2$ shatters some $S \subseteq C'$.
Then by definition, $\mathcal{H}_C$ shatters $S \cup \{x\}$ thus $VCdim(\mathcal{H}_c) \geq VCdim(\mathcal{H}_2) + 1$. □

Observation 2: Applying induction assumption again,

$$|\mathcal{H}_2| \leq \begin{Bmatrix} m-1 \\ d-1 \end{Bmatrix}$$

Combining both observations and the Claim we get:

$$|\mathcal{H}_c| = |\mathcal{H}_1| + |\mathcal{H}_2| \leq \begin{Bmatrix} m-1 \\ d \end{Bmatrix} + \begin{Bmatrix} m-1 \\ d-1 \end{Bmatrix} = \begin{Bmatrix} m \\ d \end{Bmatrix}$$

(last identity is a combinatorial identity in homework) This completes Sauer-Shelah lemma.

Fundamental theorem of statistical learning - recall proof outline

Two main parts:

- Sauer'-Shelah lemma: even though $\mathcal{H}$ can be large/infinite, when restricting it to $C \subseteq \mathcal{X}$ we have $\tau_{\mathcal{H}}(m) = O(m^d)$
- The generalization error as a function of the training set set $m = |S|$ behaves as

$$\frac{\log \tau(m)}{m}$$

Part II (main)

Theorem

Let $\mathcal{D}$ be an unknown distribution function and $\mathcal{H}$ a realizable hypothesis class. Given a set S consisted of m iid samples from $\mathcal{D}$, so that

$$m = O(\frac{1}{\varepsilon} \log \frac{\tau_{\mathcal{H}}(2m)}{\varepsilon\delta})$$

then $\forall \varepsilon, \delta > 0$ and $\forall h \in \mathcal{H}$ such that $err_S(h) = 0$:

$$\Pr[\underset{\mathcal{D}}{\text{err}}(h) < \varepsilon] > 1 - \delta$$

Remark: this can be strengthened a bit in terms of ε, δ, or altenatively realizeability can be removed at a cost of additional ε.

Part II - main idea

The proof reduces the concentration argument to a finite sample.

1. First, we reduce the question of learnability to that of learning a finite sample.
2. Prove concentration on the finite sample.

Part 1:
Let S, S' be two i.i.d. samples of size m.

- Let A be the event where exists $h \in \mathcal{H}$ such that $\mathrm{err}_S(h) = 0$, but $\mathrm{err}_{\mathcal{D}}(h) > \varepsilon$.
- Let B be the event where $\exists h \in \mathcal{H}$ so that $\mathrm{err}_S(h) = 0$ and $\mathrm{err}_{S'}(h) > \varepsilon/2$.

Claim

$\Pr[A] \leq 2\Pr[B]$

By law of complete probability:

$$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr\left(B \mid \bar{A}\right)\Pr\left(\bar{A}\right) \geq \Pr(B \mid A)\Pr(A)$$

Therefore it is enough to show that $\Pr_{S'}(B \mid A) \geq \frac{1}{2}$ Define the random variable $Y = \frac{1}{m}\sum_{i=1}^{m} Z_i$, where Z_i is:

$$Z_i = \begin{cases} 1, & \Pr_{(x_i,y_i)\sim\mathcal{D}}\left(h(x_i) \neq y_i\right) \\ 0, & \text{otherwise} \end{cases}.$$

Notice: $0 \leq Y \leq 1$, $Y = \mathrm{err}_{S'}(h)$ and also $\mathbf{E}[Y] = \mathrm{err}_{\mathcal{D}}(h) \geq \varepsilon$, thus,

$$\Pr(Y < \varepsilon/2) \leq \Pr(|Y - \mathbf{E}[Y]| > \varepsilon/2) \leq 2\exp\left(-\frac{m\varepsilon}{2}\right)$$

where for the last transition we used Chernoff's inequality (stronger version) Therefore, by our choice of m

$$\Pr(B \mid A) = \Pr(Y \geq \varepsilon/2) \geq 1 - 2e^{-\frac{m\varepsilon}{2}} \geq \frac{1}{2}$$

part 2

Recall: S, S' are two i.i.d. samples of size m.

- Let A be the event where exists $h \in \mathcal{H}$ such that $\mathrm{err}_S(h) = 0$, but $\mathrm{err}_{\mathcal{D}}(h) > \varepsilon$.
- Let B be the event where $\exists h \in \mathcal{H}$ so that $\mathrm{err}_S(h) = 0$ and $\mathrm{err}_{S'}(h) > \varepsilon/2$.

We've shown $\Pr[A] \leq 2\Pr[B]$. We proceed to bound $\Pr[B]$.

heart of the proof

Claim

$$\Pr[B] = \Pr[\exists h \in \mathcal{H}, \underset{S}{\text{err}}(h) = 0, \underset{S'}{\text{err}}(h) > \frac{\varepsilon}{2}] \leq \tau_{\mathcal{H}}(2m)\, 2^{-\frac{m\varepsilon}{2}}$$

Given $S = \{x_1, ..., x_m\}, S' = \{x'_1, ..., x'_m\}$, we define T, T' as follows: $\forall 1 \leq i \leq m$ we assign x_j to T and x'_j to T' with probability $\frac{1}{2}$, and vice versa w.p. $\frac{1}{2}$. Since S, S' were sampled iid, the sets T, T' are also iid. Define B_T as the event where $\exists h \in \mathcal{H}$ so that $\text{err}_T(h) = 0$, and $\text{err}_{T'}(h) \geq \varepsilon/2$. Therefore:

$$\Pr_{S,S'}(B) = \Pr_{S,S',T}(B_T) = \mathop{\mathbf{E}}_{S,S'} \Pr_{T}[B_T | S, S']$$

heart of the proof

The amount of error configurations, given S, S', h is limited by $\tau_{\mathcal{H}}(2m)$, so:

$$\begin{aligned}
\Pr_T[B_T|S,S'] &= \Pr_T\left[\cup_{h\in\mathcal{H}} \operatorname*{err}_T(h)=0 \wedge \operatorname*{err}_{T'}(h) > \frac{\varepsilon}{2}|S,S'\right] \\
&\leq |\mathcal{H}_{S\cup S'}| \max_{h,S,S'} \Pr_T[B_T|S,S',h] \\
&\leq \tau_{\mathcal{H}}(2m) \max_{h,S,S'} \Pr_T[B_T|S,S',h]
\end{aligned}$$

Now suppose there are k mistakes for h in the samples S, S'. If $\#\text{mistakes} = k < \frac{m\varepsilon}{2}$, then event B_T cannot hold, and its probability is zero. Else, mistakes must occupy disjoint places, and

$$\max_{h,S,S'} \Pr_T[B_T|S,S',h] \leq 2^{-k} \leq 2^{-\frac{m\varepsilon}{2}}$$

and hence

$$\Pr_T[B_T|S,S'] \leq \tau_{\mathcal{H}}(2m) \cdot 2^{-\frac{m\varepsilon}{2}}$$

conclusion

Thus,

$$\begin{aligned}\Pr[B] &= \textstyle\sum_{S,S'} \Pr_T[B_T|S,S'] \Pr[S,S'] \\ &\leq \tau_{\mathcal{H}}(2m)\, 2^{-\frac{m\varepsilon}{2}} \textstyle\sum_{S,S'} \Pr[S,S'] \\ &\leq \tau_{\mathcal{H}}(2m)\, 2^{-\frac{m\varepsilon}{2}}\end{aligned}$$

And we conclude,

$$\Pr[A] \leq 2\Pr[B] \leq 2\tau_{\mathcal{H}}(2m)\, 2^{-\frac{m\varepsilon}{2}} \leq \delta$$

The last inequality is by our choice of

$$m = O(\frac{1}{\varepsilon} \log \frac{\tau_{\mathcal{H}}(2m)}{\delta})$$

conclusion

We have seen:

- For $d = \text{VC-dim}(\mathcal{H})$, that $\tau_{\mathcal{H}}(m) \leq (\frac{em}{d})^d$
- For sample size $m = O(\frac{1}{\varepsilon} \log \frac{\tau_{\mathcal{H}}(2m)}{\delta})$ we have $\forall h \in \mathcal{H}$ such that $err_S(h) = 0$ that

$$\Pr[\underset{\mathcal{D}}{\text{err}}(h) < \varepsilon] > 1 - \delta$$

This shows sample complexity bound of (exercise...)

$$m_{\mathcal{H}}(\varepsilon, \delta) = O(\frac{d}{\varepsilon} \log \frac{d}{\delta\varepsilon})$$

Concluding the positive part of the fundamental theorem of statistical learning.

conclusions

Claim

If $\text{VC-dim}(\mathcal{H}) = \infty$, *then* $\mathcal{H}$ *is not learnable*

Sketch: for every sample size m, there exists a subset $C \subseteq \mathcal{X}$ for which the "no free lunch" theorem applies.

Recap + what's next

- Fundamental theorem of statistical learning exactly characterizes finite and infinite hypothesis classes that are learnable.
- No efficient algorithms (only ERM till now...).
- Restricted to statistic model (training set available).

Next:

- Efficient algorithms for learning.
- Online model.