Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 3:

PAC Learning and VC Theory II

- What is learnable?
- With how many samples?

Statistical No Free Lunch

- **Theorem**: For any domain $\mathcal{X}$ of size $|\mathcal{X}|$ and any learning rule A, there exists a source distribution $\mathcal{D}$ with $\mathbb{P}_{x,y\sim\mathcal{D}}[f(x)=y]=1$ for some $f:\mathcal{X}\to\{\pm 1\}$, such that for $m<\frac{|\mathcal{X}|}{2}$,

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

(and so w.p.$\geq 1/7$, $L_{\mathcal{D}}\left(A(S)\right)\geq 1/8$)

- **Conclusion**: For an infinite domain $\mathcal{X}$, for any learning rule A and *any* sample size m, there exists a source distribution and f as above such that

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

Statistical No Free Lunch—Stronger Statement

- For a finite domain $\mathcal{X}$, $\mathcal{Y} = \{\pm 1\}$, and $f: \mathcal{X} \to \mathcal{Y}$, denote $\mathcal{U}_f$ the source distribution s.t.:
  - x is uniform over $\mathcal{X}$
  - $y = f(x)$ with probability one

- Consider a uniform distribution over $f: \mathcal{X} \to \mathcal{Y}$
  (i.e. for each x set $f(x) = \pm 1$ w.p. $1/2$, independent of all other values)

- **Theorem**: For any learning rule A and any sample size m,

$$\frac{1}{2} - \frac{m}{2|\mathcal{X}|} \leq \mathbb{E}_f \mathbb{E}_{S \sim \mathcal{U}_f^m} \left[L_{\mathcal{U}_f}(A(S)) \right] \leq \frac{1}{2} + \frac{m}{2|\mathcal{X}|}$$

Statistical No Free Lunch: Proof

- Define:

  "S is consistent with f" if $\forall_{(x_i,y_i)\in S} f(x_i) = y_i$

  "S is self-consistent" if it is consistent with some f (i.e. if $x_i = x_j$ then $y_i = y_j$)

- For any learning rule A, and any self-consistent sample S:

$$\mathbb{E}_f\left[L_{\mathcal{U}_f}(A(S)) \mid S \text{ cons with } f\right] = \mathbb{E}_f\left[\frac{1}{|\mathcal{X}|}\sum_x [[A(S)(x) \neq f(x)]] \mid S \text{ cons with } f\right]$$

$$= \frac{1}{|\mathcal{X}|}\left(\sum_{x \text{ appears in } S} [[A(S)(x) \neq f(x)]] + \sum_{x \notin S} \underbrace{\mathbb{P}_f[A(S)(x) \neq f(x)]}_{1/2}\right) = \frac{1}{2} \pm \frac{m'}{2|\mathcal{X}|}$$

$|\{x \mid x \text{ appears in } S\}|$

- And so:

$$\mathbb{E}_f \mathbb{E}_{S\sim\mathcal{U}_f^m}\left[L_{\mathcal{U}_f}(A(S))\right] = \mathbb{E}_f \mathbb{E}_{\text{self}-\text{cons } S}\left[L_{\mathcal{U}_f}(A(S)) \mid f \text{ cons with } S\right]$$

$$= \mathbb{E}_{\text{self}-\text{cons } S} \mathbb{E}_f\left[L_{\mathcal{U}_f}(A(S)) \mid f \text{ cons with } S\right] = \frac{1}{2} \pm \frac{m}{2|\mathcal{X}|}$$

QED

Learning

- No Free Lunch:
  - Without assuming anything on f, can't do any better than memorization
  - For a random f, all learning rules essentially the same

- If we assume $f \in \mathcal{H}$, **with $\mathcal{H}$ known**, or just want to compete with $h \in \mathcal{H}$, we can learn with $O\left(\mathrm{VCdim}(\mathcal{H})\right)$ samples

VC Learning Guarantees

- Theorem: For any hypothesis class $\mathcal{H}$:

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \qquad L_{\mathcal{D}}\left(\hat{h}\right) \leq L_S\left(\hat{h}\right) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) + \log {}^{1}/_{\delta}}{m}}\right)$$

- Conclusion: If $\mathrm{VCdim}(\mathcal{H}) < \infty$ then $\mathcal{H}$ is **agnostically PAC learnable** using $ERM_{\mathcal{H}}$ with sample complexity

$$m(\epsilon, \delta) \leq O\left(\frac{\mathrm{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right)$$

I.e., for any $\mathcal{D}$, w.p. $\geq 1-\delta$ over $S \sim \mathcal{D}^{m(\epsilon,\delta)}$, $L_{\mathcal{D}}\left(ERM_{\mathcal{H}}(S)\right) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

VC Dimension

- $C = \{x_1, \ldots, x_m\}$ is **shattered** by $\mathcal{H}$ if we can get all 2^m behaviors:

$$\forall_{y_1,\ldots,y_m \in \pm 1}, \exists_{h \in H} \text{ s.t. } \forall_i h(x_i) = y_i$$

- The VC-dimension of $\mathcal{H}$ is the largest number of points that can be shattered by $\mathcal{H}$

VC-Dimension: Examples

- Circles in $\mathbb{R}^2$: $\mathcal{H} = \left\{ h_{c,r}(x) = \left[\left[\|x - c\| \leq r\right]\right] \mid c \in \mathbb{R}^2, r \in \mathbb{R} \right\}$
  - Can shatter 3 points
- Circles and their complement
  - Can shatter 4 points
- Circles around origin: $\mathcal{H} = \left\{ h_{c,r}(x) = \left[\left[\|x\| \leq r\right]\right] \mid r \in \mathbb{R} \right\}$
  - Can shatter only 1 point
- Axis aligned ellipses:

$$\mathcal{H} = \left\{ h_{c,r[1],r[2]}(x) = \left[\left[\frac{(x[1] - c[1])^2}{r[1]^2} + \frac{(x[2] - c[2])^2}{r[2]^2} \leq 1\right]\right] \mid c \in \mathbb{R}^2, r[1], r[2] \in \mathbb{R} \right\}$$

  - Can shatter 4 points
- General ellipses
  - Can shatter 5 points

- Upper bounds?

VC dim of Homogenous Half Spaces

$$\mathcal{H}_\phi = \{ [[\langle w, \phi(x)\rangle \geq 0]] \mid w \in \mathbb{R}^d \}, \qquad \phi: \mathcal{X} \to \mathbb{R}^d$$

- Can shatter the d points: $e_1, \ldots, e_d$. Use $w = (y_1, y_2, \ldots, y_d)$.

- Claim: can't shatter any set of d+1 points
  - For any d+1 points $x_1, \ldots, x_{d+1}$, there must be some linear dependency:
$$\sum_i a_i x_i = 0$$
  - Let $I = \{i | a_i > 0\}$, $J = \{j | a_j < 0\}$
  - At least one coefficient is non-zero. By negating all coefficients if necessary, can assume without loss of generality that J is non-empty.
  - Consider labeling where $y_i = +1$ for $i \in I$ and $y_j = -1$ for $j \in J$ (and arbitrary label for points not in either one).
  - The linear predictor w that attains this labeling satisfies:
$$0 \leq \sum_{i \in I} a_i \langle w, x_i \rangle = \langle w, \sum_{i \in I} a_i x_i \rangle = -\langle w, \sum_{j \in J} a_j x_j \rangle = -\sum_{j \in I} a_j \langle w, x_j \rangle < 0$$

- Conclusion: $\text{VCdim}(\mathcal{H}_\phi) = d$

Half Space Representations

- **Theorem**: for a hypothesis class $\mathcal{H}$, if there exists $\phi: \mathcal{X} \to \mathbb{R}^D$ s.t.
$$\mathcal{H} \subseteq \mathcal{H}_\phi,$$
i.e. s.t. every hypothesis $h \in \mathcal{H}$ can we written as $h(x) = \text{sign}(\langle w_h, \phi(x) \rangle)$ for some $w_h \in \mathbb{R}^D$, then $\text{VCdim}(\mathcal{H}) \leq D$.

- **Example**: non-homogenous half-spaces over $\mathbb{R}^d$, use $D = d + 1$ with $\tilde{\phi}(x) = [\phi(x), 1]$.

Half Space Representation: Circles

$$h_{a,r}(x) = \left[\left[(x[1]-a[1])^2 + (x[2]-a[2])^2 \le r^2\right]\right]$$

$$= \left[\left[x[1]^2 - 2a[1]x[1] + a[1]^2 + x[2] - a[2]x[2] + a[2] \le r^2\right]\right]$$

$$= \text{sign}\left((-x[1]^2 - x[2]^2) + 2a[1]x[1] + 2a[2]x[2] + (r^2 - a[1]^2 - a[2]^2)\right)$$

$$= \text{sign}\left(\langle w_{a,r}, \phi(x)\rangle\right)$$

$$\phi(x) = (-x[1]^2 - x[2]^2, x[1], x[2], 1)$$

$$w_{a,r} = (1, 2a[1], 2a[2], r^2 - a[1]^2 - a[2]^2)$$

- Conclusion: VCdim ≤ 4
- Why not tight?
- If we allow $w[1] < 0$, we get circles and their complement, and a tight bound on its VCdim

Half Space Representation

- Axis-aligned ellipses (and their complement):
$$\phi(x) = (\, x[1]^2, \quad x[2]^2, \quad x[1], \quad x[2], \quad 1)$$

- Conic cuts (including all ellipses):
$$\phi(x) = (\, x[1]^2, \quad x[2]^2, \quad x[1]x[2], \quad x[1], \quad x[2], \quad 1)$$

- Degree-k polynomials over $\mathbb{R}^2$:
$$\phi(x) = (x[1]^k, x[1]^{k-1}x[2]^1, x[1]^{k-2}x[2]^2, \dots, x[1]^1x[2]^{k-1}, x[2]^k,$$
$$x[1]^{k-1}, x[1]^{(k-2)x[2]^1}, \dots, x[2]^{k-1},$$
$$x[1]^{k-2}, \dots$$
$$x[1]^2, x[1]x[2], x[2]^2,$$
$$x[1], x[2], 1) \in \mathbb{R}^{(k+1)k/2}$$

- Degree-k polynomials over $\mathbb{R}^d$: $\phi(x) \in \mathbb{R}^{\binom{d+k-1}{k}}$ ➔ $D = O(d^k)$

VCdim always = #params?

$$\mathcal{X} = \mathbb{R} \qquad \mathcal{H} = \left\{ h_{\theta,\nu}(x) = \text{sign}\left(\sin(\nu x + \theta)\right) \mid \nu, \theta \in \mathbb{R} \right\}$$

- Claim: $\text{VCdim}(\mathcal{H}) = \infty$
- Proof: consider the infinite set of points $\left\{ x_i = 10^{-i} \right\}_{i=1,2,\ldots}$. Any labeling $y_1, y_2, \ldots$ is attained by $\theta = 0$ and:

$$\nu = \pi \left(1 - \sum_{i=1}^{\infty} \frac{y_i}{2x_i} \right)$$

Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0, \exists m(\epsilon, \delta), \forall \mathcal{D}, \forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

- Sample complexity of a learning rule:
$m_{A,\mathcal{H}}(\epsilon, \delta) = \min m \; s.t. \; \forall \mathcal{D}, \forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}, L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$
- Sample complexity for learning a hypothesis class:
$$m_{\mathcal{H}}(\epsilon, \delta) = \min_{A} m_{A,\mathcal{H}}(\epsilon, \delta)$$

- What hypothesis classes are learnable?
- What controls the sample complexity?

Probably Approximately Correct (PAC)

- Definition: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0, \exists m(\epsilon, \delta), \forall \mathcal{D}, \forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

- Sample complexity of a learning rule:
$$m_{A,\mathcal{H}}(\epsilon, \delta) = \min m \; s.t. \; \forall \mathcal{D}, \forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}, L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$
- Sample complexity for learning a hypothesis class:
$$m_{\mathcal{H}}(\epsilon, \delta) = \min_{A} m_{A,\mathcal{H}}(\epsilon, \delta)$$

- Finite classes are PAC-learnable, with $m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log|\mathcal{H}| + \log^{1}/_{\delta}}{\epsilon^2}\right)$
- VC classes are PAC-learnable, with $m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\mathrm{VCdim}(\mathcal{H}) + \log^{1}/_{\delta}}{\epsilon^2}\right)$
- **Can a class with infinite VC-dimension be learnable?**

VC Dimension: Converse

- Suppose $\mathrm{VCdim}(\mathcal{H}) = D$. Might it be possible to learn with $\omega(D)$ samples?
- There exists D points $\mathcal{X}' = \{x_1, x_2, \dots, x_D\}$ that are shattered by $\mathcal{H}$
- Restricting attention only of $\mathcal{X}'$, $\mathcal{H}$ does not constrain us at all, and we can apply the No Free Lunch Theorem on $\mathcal{X}'$.

  I.e., we consider distributions $\mathcal{D}_h$ where x is uniform on $\mathcal{X}'$, and $y = h(x)$ w.p. 1, for $h \in \mathcal{H}$ (recall this allows any labeling on $\mathcal{X}'$)
- **Conclusion**: for any learning rule A, there exists a distribution $\mathcal{D}_h$ and $h \in \mathcal{H}$ with $L_{\mathcal{D}_h}(h) = 0$, s.t. with $m < D/2$ samples, w.p. $\geq 1/7$, $L_{\mathcal{D}_h}\big(A(S)\big) \geq 1/8$.

$$m_{\mathcal{H}}\left(1/8, 6/7\right) \geq \mathrm{VCdim}(\mathcal{H})/2$$

Fundamental Theorem of Statistical Learning Theory

- If $\text{VCdim}(\mathcal{H}) < \infty$ then $\mathcal{H}$ is agnostic-PAC learnable with sample complexity

$$\Omega\left(\frac{\text{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H},ERM}(\epsilon, \delta) \leq O\left(\frac{\text{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right)$$

- If $\mathcal{H}$ is PAC-learnable using any learning rule, even in the realizable case, i.e. even if only when $\exists_{h \in \mathcal{H}} L(h) = 0$, then it must have finite VC-dimension.

- In the realizable case, the sample complexity is

$$\Omega\left(\frac{\text{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{\text{VCdim}(\mathcal{H}) \log 1/\epsilon + \log(1/\delta)}{\epsilon}\right)$$

Note: in homework, you will only show $m_{\mathcal{H},ERM}(\epsilon, \delta) \leq O\left(\frac{\text{VCdim}(\mathcal{H}) \log 1/\epsilon + \log(1/\delta)}{\epsilon^2}\right)$

Implications of Fundamental Theorem

- Exact characterization of what is learnable
- Tight understanding of sample complexity
  - #samples $\propto$ VC-dimension $\approx$ #parameters
  - Once we can't explain everything (fit every possible behavior), we start learning
- One learning rule to rule them all: ERM

Question: Harvard, 1984

Answer: Moscow, 1971

Implications of Fundamental Theorem

- Exact characterization of what is learnable
- Tight understanding of sample complexity
  - #samples $\propto$ VC-dimension $\approx$ #parameters
  - Once we can't explain everything (fit every possible behavior), we start learning
- One learning rule to rule them all: ERM

But:

- What about computation? Can we implement ERM?
  - Valiant's actual question: what is ***efficiently*** PAC learnable?
- Other forms of prior knowledge beyond "captured by $\mathcal{H}$"

Non-Uniform Bias

- Up until now: "flat" prior on $\mathcal{H}$—every $h \in \mathcal{H}$ equally likely
- Instead: "Prior" $p(h)$ encodes bias

$$p: \mathcal{H} \to [0,1] \text{ (or } p: \mathcal{Y}^{\mathcal{X}} \to [0,1]), \quad \sum_h p(h) \leq 1$$

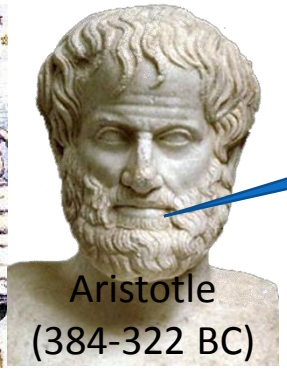
- Expert says higher $p(h)$ more likely (e.g. relying on "better" features)
- Bias toward simpler predictors; $p(h)$ encodes "simplicity"
- Bias toward "shorter" explanations; $p(h)$ encodes "description length"

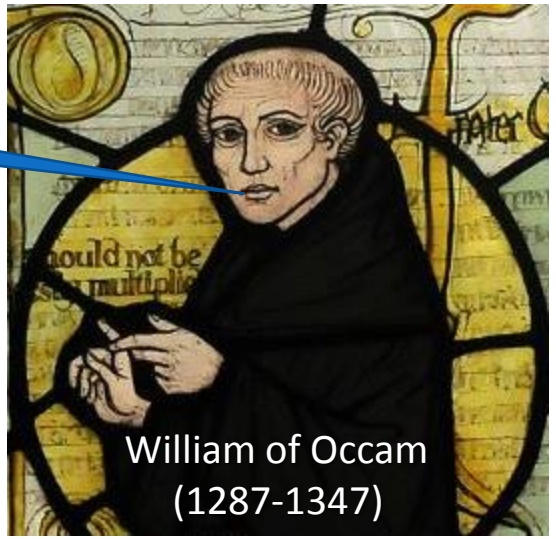
Occam's Razor: "A short explanation is preferred over a longer one"

We said it first!


Maimonides (1138-1204)


Ptolemy (90-168)


Aristotle (384-322 BC)

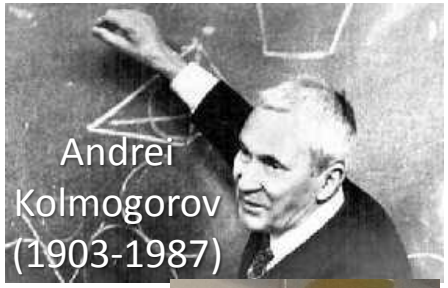

William of Occam (1287-1347)

Bias to Shorter Description

$$p: \mathcal{Y}^{\mathcal{X}} \to [0,1] \qquad \sum_h p(h) \leq 1$$

- Based on length of (prefix-ambiguous) description $d(h)$
  - $d: \mathcal{H} \to \{0,1\}^*$, $d(h)$ is never a prefix of $d(h')$ for any h, h'
  - $p(h) = 2^{-|d(h)|}$
  - Kraft Inequality: $\sum \frac{1}{2^{|d(h)|}} = \sum p(h) \leq 1$
- Based on c: $U \to \mathcal{Y}^{\mathcal{X}}$ (e.g. python code↦function it implements)
  - Set of prefix-ambiguous "legal programs" $U \subset \{0,1\}^*$
  - $p(h) = 2^{-\min_{c(\sigma)=h} |\sigma|}$ (can think of: $d(h) = \arg\min_{c(\sigma)=h} |\sigma|$)

Kolmogorov Complexity



- Minimum Description Length learning rule:

$$MDL_p(S) = \arg\max_{L_S(h)=0} p(h) = \arg\min_{L_S(h)=0} |d(h)|$$

MDL and Non-Uniform Concentration

- Recall: for any h, $P_S\left(|L_S(h) - L(h)| \geq \sqrt{\frac{\log 2/\delta}{2m}}\right) \leq \delta$
- Set $\delta_h = p(h) \cdot \delta$:

$$P\left(\exists_h |L_S(h) - L(h)| \geq \sqrt{\frac{\log 2/\delta_h}{2m}}\right) \leq \sum_h \delta_h = \sum_h p(h)\delta \leq \delta$$

$$\sqrt{\frac{\log 2/(p(h)\delta)}{2m}} = \sqrt{\frac{\log 1/p(h) + \log 2/\delta}{2m}}$$

- Conclusion: w.p. $\geq 1 - \delta$, for all h concurrently,

$$L(h) \leq L_S(h) + \sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}}$$

Minimized by MDL

- If $L(h^*) = 0$ for some h^*, we necessarily also have $L_S(h^*) = 0$ and so:

$$L_S\left(MDL_p(S)\right) = 0, \qquad p\left(MDL_p(S)\right) \geq p(h^*)$$

- Conclusion:

$$L\left(MDL_p(S)\right) \leq \sqrt{\frac{-\log p(h^*) + \log 2/\delta}{2m}}$$

MDL and Universal Learning

- **Theorem:** For any prior $p(h)$, $\sum_h p(h) \leq 1$ (e.g. $p(h) = 2^{-|d(h)|}$ for a prefix-ambiguous $d(h)$), and any source distribution $\mathcal{D}$, if there exists h^* with $L(h^*) = 0$, then w.p. $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$:

$$L\left(MDL_p(S)\right) \leq \sqrt{\frac{-\log p(h^*) + \log 2/\delta}{2m}} = \sqrt{\frac{|d(h^*)| + \log 2/\delta}{2m}}$$

- Sample complexity: $m = O\left(\frac{|d(h^*)|}{\epsilon^2}\right)$ (more careful analysis: $O\left(\frac{|d(h^*)|}{\epsilon}\right)$)
- Can learn any countable class!
  - Class of all computable functions, with $p(h) = 2^{-\min_{c(\sigma)=h} |\sigma|}$.
  - Class enumerable with $n: \mathcal{H} \to \mathbb{N}$ with $p(h) = 2^{-n(h)}$
- But VCdim(all computable functions)=∞ !
- Why no contradiction to Fundamental Theorem?