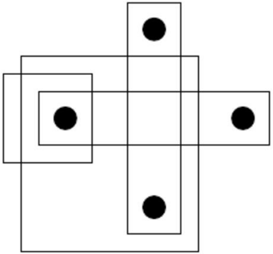
Lecture 19: Shatter functions

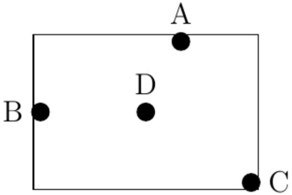
Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

- Set system: $(X, \mathcal{H})$
  - X — instance space
  - $\mathcal{H}$, subsets of X — possible classifiers / hypotheses
- $A \subseteq X$ is shattered by $\mathcal{H}$ if every subset of A is given by $A \cap h$ for some $h \in \mathcal{H}$
- VC-Dimension of $\mathcal{H}$ — size of the largest subset of X shattered by $\mathcal{H}$
  - VC-dimension of axis-parallel rectangles is 4
  - VC-dimension may be finite or infinite
- How can we relate VC-dimension to bounds on training error and test error?





Shatter function

- Shatter function: $\pi_{\mathcal{H}}(n)$ — maximum number of subsets of any set A of size n that can be expressed as $A \cap h$ for some $h \in \mathcal{H}$
  - Let $d = \text{VC-dim}(\mathcal{H})$
  - For $n \leq d$, $\pi_{\mathcal{H}}(n) = 2^n$
  - If d is infinite, $\pi_{\mathcal{H}}(n) = 2^n$ for all n
  - What if d is finite?

Sauer's Lemma

For any set system $(X, \mathcal{H})$ of VC-dimension at most d, $\pi_{\mathcal{H}}(n) \leq \sum_{i=1}^{d} \binom{n}{i}$ for all n

- $\sum_{i=1}^{d} \binom{n}{i} = O(n^d)$, polynomial with respect to VC-dimension

Shatter functions for combinations of hypothesis

- Let $(X, \mathcal{H}_1)$ and $(X, \mathcal{H}_2)$ be two set systems
- Intersection system — $(X, \mathcal{H}_1 \cap \mathcal{H}_2)$
  - Combine $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ using AND

Claim $\pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \leq \pi_{\mathcal{H}_1}(n) \cdot \pi_{\mathcal{H}_2}(n)$

- For $A \subseteq X$, $|A| = n$, interested in size of $\mathcal{S} = \{A \cap h \mid h \in \mathcal{H}_1 \cap \mathcal{H}_2\}$
- $\mathcal{S} = \{A \cap (h_1 \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$
- $\mathcal{S} = \{(A \cap h_1) \cap (A \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$
- $|\mathcal{S}| \leq |\{(A \cap h_1) \mid h_1 \in \mathcal{H}_1\}| \times |\{(A \cap h_2) \mid h_2 \in \mathcal{H}_2\}|$

Generalizes to boolean combination $f(h_1, \ldots, h_k)$, where each $h_i \in \mathcal{H}$

- $\pi_{f(\mathcal{H})}(n) \leq \pi_{\mathcal{H}}(n)^k$

VC-dimension and machine learning

- Let $(X, \mathcal{H})$ be a set system with probability distribution D over X
- Given target concept c^* and $h \in \mathcal{H}$, error region is symmetric difference $h \Delta c^*$
- For a training sample S, want h with $Pr(h\Delta c^*) \geq \epsilon$ to have $|S \cap (h \Delta c^*)| > 0$
- Set of error regions is $\mathcal{H}' = \{h\Delta c^* \mid h \in \mathcal{H}\}$
- **Claim:** $\mathcal{H}$ and $\mathcal{H}'$ have same VC-dimension and shatter function
  - If $\text{VC-dim}(\mathcal{H}) = d$, there is $A \subseteq X$, $|A| = d$, shattered by $\mathcal{H}$
  - Three cases: $A \cap c^* = \emptyset$, $A \subseteq c^* = \emptyset$, $A \cap c^* \neq \emptyset$
  - In all cases, can shatter A with $\mathcal{H}'$ as well

VC-dimension and machine learning

- For a training sample S, want h with $Pr(h\Delta c^*) \geq \epsilon$ to have $|S \cap (h\Delta c^*)| > 0$
- Set of error regions is $\mathcal{H}' = \{h\Delta c^* \mid h \in \mathcal{H}\}$
- Apply the following result to $\mathcal{H}'$

Key Theorem

Let $(X, \mathcal{H})$ be a set system, D a probability distribution over X, and let n be an integer satisfying $n \geq \frac{8}{\epsilon}$ and $n \geq \frac{2}{\epsilon}\left[\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \frac{1}{\delta}\right]$.

Let S consists of n points drawn from D. With probability $\geq 1-\delta$, every $h \in \mathcal{H}$ with $Pr(h) > \epsilon$ intersects S.

- Proof omitted. Strange constants like $\frac{8}{\epsilon}$ arise from use of Chebyshev's inequality in the proof.

VC-dimension and machine learning

- This gives us the following analogue of the PAC learning guarantee

Sample bound

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn using D of size

$$n > \frac{2}{\epsilon}\left[\log(2\pi_{\mathcal{H}}(2n)) + \log\frac{1}{\delta}\right]$$

then with probability $\geq 1-\delta$,

- every $h \in \mathcal{H}$ with true error $\mathrm{err}_D(h) \geq \epsilon$ has training error $\mathrm{err}_S(h) > 0$,
- i.e., every $h \in \mathcal{H}$ with training error $\mathrm{err}_S(h) = 0.$ has true error $\mathrm{err}_D(h) < \epsilon$

- There is a corresponding version of uniform convergence.

Shatter function uniform convergence

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn using D of size

$$n > \frac{8}{\epsilon^2}\left[\ln(2\pi_{\mathcal{H}}(2n)) + \ln\frac{1}{\delta}\right]$$

then with probability $\geq 1-\delta$, every $h \in \mathcal{H}$ will have $|\mathrm{err}_S(h) - \mathrm{err}_D(h)| \leq \epsilon$.

VC-dimension and machine learning

- Using Sauer's Lemma we can rewrite the sample bound directly using VC-dimension.

Sample bound using VC-dimension

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn using D of size

$$O\left(\frac{1}{\epsilon}\left[\text{VC-dim}(\mathcal{H}) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right]\right)$$

then with probability $\geq 1 - \delta$,

- every $h \in \mathcal{H}$ with true error $\text{err}_D(h) \geq \epsilon$ has training error $\text{err}_S(h) > 0$,
- i.e., every $h \in \mathcal{H}$ with training error $\text{err}_S(h) = 0$. has true error $\text{err}_D(h) < \epsilon$

- We can similarly rewrite the uniform convergence criterion

Summary

- VC-dimension gives rise to the shatter function
- For finite VC-dimension, shatter function grows as a polynomial in VC-dimension (Sauer's Lemma)
- We can prove analogues of PAC learning guarantee and uniform convergence in terms of shatter function
- Note that these theoretical bounds are hard to use in practice
- Difficult, if not impossible, to compute VC-dimension and shatter function for complex models