Learning Theory II

CSci 5525: Machine Learning

Instructor: Nicholas Johnson

November 5, 2020

Announcements

- HW3 due next Tue (Nov 10)
- Project progress report due in $<$ 2 weeks (Nov 17)

Finite Hypothesis Spaces

- Consider event $Z_h = \{(x, y) \in D^n : \mathcal{R}(h) > \hat{\mathcal{R}}_{(x,y)}(h) + \epsilon\}$
- Samples of size n give empirical risk ϵ more than true risk
- By union bound over $\mathcal{H} = \{h_1, \ldots, h_m\}$

$$P\left(\bigcup_h Z_h\right) \leq \sum_{i=1}^{m} P(Z_{h_i}) = m \exp(-2n\epsilon^2)$$

- Given 'tolerance' δ, want $m \exp(-2n\epsilon^2) \leq \delta$
- With probability at least $1 - \delta$

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log 1/\delta}{2n}}$$

- Union bound for general function classes (i.e., infinite-sized function classes)?
- Preview: replace $\log(|\mathcal{H}|)$ with complexity measure of $\mathcal{H}$

VC Dimension

- What if $|\mathcal{H}|$ is infinite?
- Replace $\log(|\mathcal{H}|)$ with complexity measure of $\mathcal{H}$
- Consider a set of points $x_1, \ldots, x_n \in \mathcal{X}$
- The function class $\mathcal{H}$ *shatters* $x_1, \ldots, x_n$ if $\mathcal{H}$ realizes all labelings over the n points
- The VC dimension of $\mathcal{H}$ is the largest number of points $\mathcal{H}$ can shatter

data

$$VCD(\mathcal{H}) = \max\{n \in \mathbb{Z} : \exists (x_1, \ldots, x_n) \in \mathcal{X}^n,$$
$$\forall (y_1, \ldots, y_n) \in \{0,1\}^n,$$
$$\exists h \in \mathcal{H},$$
$$h(x_i) = y_i \forall i\}$$

all labels

classifier

prediction

shatter

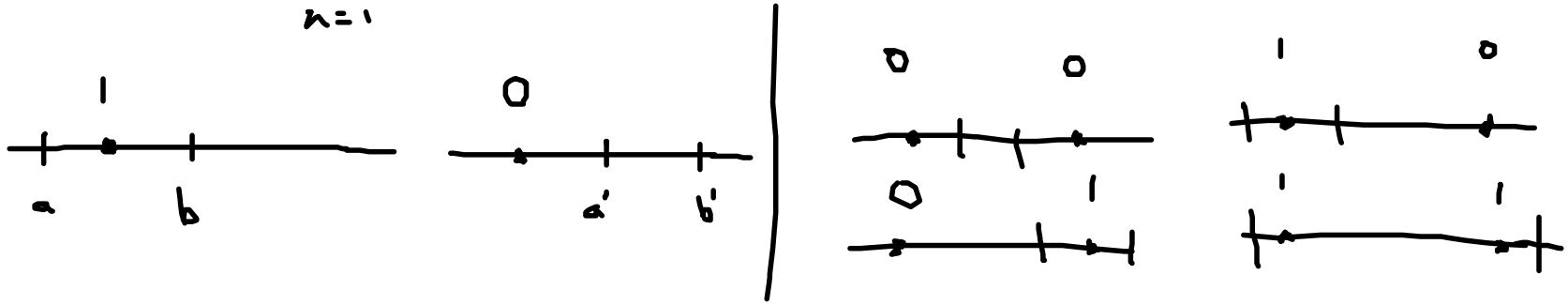
VC Dimension

- Claim: If $\mathcal{H}$ is finite then $VCD(\mathcal{H}) \leq \log(|\mathcal{H}|)$
- To establish VC dimension bound for function class $\mathcal{H}$ prove two things:
  - Upper bound showing no set of $VCD(\mathcal{H}) + 1$ or more points can be shattered by $\mathcal{H}$ and
  - Lower bound given by set of $VCD(\mathcal{H})$ points can be shattered by $\mathcal{H}$

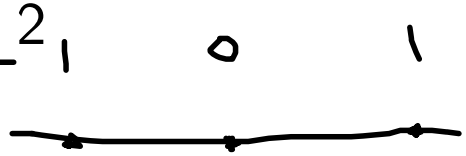
$VCD \leq VCD < VCD+1$

$VCD-1 < VCD < VCD+1$

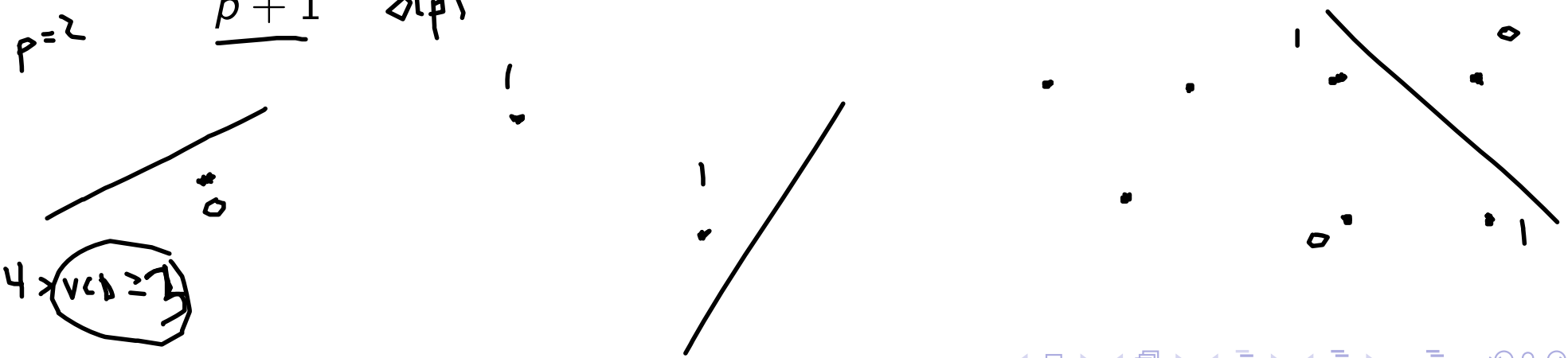
VC Dimension Examples



- **Example 1:** The class of all intervals on the real line $\mathcal{H} = \{\mathbb{1}[x \in [a, b]] | a, b \in \mathbb{R}\}$ has VC dimension 2



- **Example 2:** The class of all affine functions $\mathcal{H} = \{\mathbb{1}[\langle a, x \rangle + b \geq 0] | a, x \in \mathbb{R}^p, b \in \mathbb{R}\}$ has VC dimension $p + 1$



Uniform Convergence with VC Dimension

- Using VC dimension we can obtain uniform convergence for infinite function classes
- For a function class $\mathcal{H}$ with bounded VC dimension, with probability $1-\delta$, for all $h \in \mathcal{H}$ we have

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \tilde{O}\left(\sqrt{\frac{VCD(\mathcal{H}) + \log 1/\delta}{n}}\right)$$

where $\tilde{O}(\cdot)$ hides some log terms

Uniform Convergence with VC Dimension

- Bound with finite function class $\mathcal{H}$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log 1/\delta}{2n}}$$

- Bound with possibly infinite function class $\mathcal{H}$ with bounded VC dimension:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \tilde{O}\left(\sqrt{\frac{VCD(\mathcal{H}) + \log 1/\delta}{n}}\right)$$

VC Dimension Neural Net Example

- **Example 3:** Neural network classifier prediction:

$$f(x, \theta) = \text{sign}[h_L(W_L(\dots W_2 h_1(W_1 x + b_1) + b_2 \dots) + b_L)]$$

  - ρ number of parameters (weights and biases)
  - L number of layers

If we use the same activation for all h_i, we obtain:

- Binary activation $h(z) = \mathbb{1}[z \geq 0]$, $VCD = O(\rho \log \rho)$
  - Thm 4 in VC Dimension of Neural Networks by E. Sontag
- ReLU activation $h(z) = max(0, z)$, $VCD = O(\rho L \log(\rho L))$
  - Thm 6 in Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks by P. Bartlett et al.

VC Dimension Neural Net Example

- The VC dimension of neural networks scale with number of parameters (roughly)
- In practice number of parameters exceeds number of training examples so generalization bound using VC dimension is not very useful

Rademacher Complexity

- VC dimension designed for binary classification
- What about multi-class classification or regression?
- Rademacher complexity is more general complexity measure
- Given set of examples $\{z_1 \ldots, z_n\}$ where $z_i = (x_i, y_i)$ and function class $\mathcal{H}$ the Rademacher complexity is

$$Rad(\mathcal{H}) = E_\epsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(z_i) \right]$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables: $P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$

Rademacher Complexity

- Why does Rademacher complexity capture complexity of function class?
- Intuition: captures ability of $\mathcal{H}$ to fit random signs given by Rademacher random variables

Rademacher Complexity

- For any loss $\ell : \mathcal{Y} \times \mathcal{Y}$ and predictor $h \in \mathcal{H}$ let $\ell \circ h$ be a function such that for any example $z = (x, y)$ we have $\ell \circ h(z) = \ell(y, h(x))$
- Let corresponding function class $\ell \circ \mathcal{H} = \{\ell \circ h | h \in \mathcal{H}\}$
- Assume we have $|\ell(y, h(x))| \leq c \; \forall x, y, h$ and let $(x_1, y_1), \ldots (x_n, y_n)$ be i.i.d. draws from distribution P. With probability at least $1 - \delta$, for all $h \in \mathcal{H}$ we have

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2Rad(\ell \circ \mathcal{H}) + 4c\sqrt{\frac{2\log(4/\delta)}{n}}$$

- If ℓ is γ-Lipschitz then $Rad(\ell \circ \mathcal{H}) \leq \gamma Rad(\mathcal{H})$ and so

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2\gamma Rad(\mathcal{H}) + 4c\sqrt{\frac{2\log(4/\delta)}{n}}$$

Rademacher Complexity

- Rademacher complexity depends on underlying data distribution
- For simple function classes we can obtain complexity bounds only by assuming boundedness in the data

Rademacher Complexity Example

- **Example 1:** Consider two classes of linear functions

$$\mathcal{H}_1 = \{x \to w^\top x : w \in \mathbb{R}^p, \|w\|_1 \leq W_1\}$$

$$\mathcal{H}_2 = \{x \to w^\top x : w \in \mathbb{R}^p, \|w\|_2 \leq W_2\}$$

For $x_1, \ldots, x_n \in \mathbb{R}^p$:

$$Rad(\mathcal{H}_1) \leq (\max_i \|x_i\|_\infty) W_1 \sqrt{\frac{2\log(2p)}{n}}$$

$$Rad(\mathcal{H}_2) \leq (\max_i \|x_i\|_2) W_2 \sqrt{\frac{1}{n}}$$

- For linear functions, Rademacher complexity picks up explicit depends on norm bounds of weight vectors
- VC dimension for affine functions is just $p + 1$