Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Classification: Rademacher Complexity

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 6

Slides adapted from Rob Schapire

Setup

Nothing new ...

- Samples $S = ((x_1, y_1), \ldots, (x_m, y_m))$
- Labels $y_i = \{-1, +1\}$
- Hypothesis $h : X \rightarrow \{-1, +1\}$
- Training error: $\hat{R}(h) = \frac{1}{m} \sum_i^m \mathbb{1}\left[h(x_i) \neq y_i\right]$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m} \sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

(2)

(3)

(4)

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == \left(1, -1\right) \text{ or } \left(-1, 1\right) \\ 0 & \left(h(x_i, y_i\right) == \left(1, 1\right) \text{ or } \left(-1, -1\right) \end{cases} \tag{2}$$

$$\tag{3}$$

$$\tag{4}$$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == \left(1, -1\right) \text{ or } \left(-1, 1\right) \\ 0 & \left(h(x_i, y_i\right) == \left(1, 1\right) \text{ or } \left(-1, -1\right) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$\tag{4}$$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == \left(1, -1\right) \text{ or } \left(-1, 1\right) \\ 0 & \left(h(x_i, y_i\right) == \left(1, 1\right) \text{ or } \left(-1, -1\right) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \tag{4}$$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == (1,-1) \text{ or } (-1,1) \\ 0 & \left(h(x_i, y_i\right) == (1,1) \text{ or } (-1,-1) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \tag{4}$$

Correlation between predictions and labels

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == (1,-1) \text{ or } (-1,1) \\ 0 & \left(h(x_i, y_i\right) == (1,1) \text{ or } (-1,-1) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \tag{4}$$

Minimizing training error is thus equivalent to maximizing correlation

$$\arg\max_{h} \frac{1}{m}\sum_{i}^{m} y_i h(x_i) \tag{5}$$

Playing with Correlation

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

Playing with Correlation

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Playing with Correlation

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Notation: $\mathbb{E}_p[f] \equiv \sum_x p(x) f(x)$

Playing with Correlation

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Note: Empirical Rademacher complexity is with respect to a sample.

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$\mathbb{E}_{\sigma}\left[\max_{h \in H} \frac{1}{m} \sum_{i}^{m} \sigma_i h(x_i)\right]$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$h(x_i)\mathbb{E}_\sigma \left[\frac{1}{m}\sum_i^m \sigma_i\right]$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma \left[\frac{1}{m} \sum_i^m \sigma_i \right] = 0$$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right] = 0$$

$|H| = 2^m$

$$\mathbb{E}_\sigma\left[\max_{h \in H}\frac{1}{m}\sum_i^m \sigma_i h(x_i)\right]$$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right] = 0$$

$|H| = 2^m$

$$\frac{m}{m} = 1$$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma \left[\frac{1}{m} \sum_i^m \sigma_i \right] = 0$$

$|H| = 2^m$

$$\frac{m}{m} = 1$$

- Rademacher correlation is larger for more complicated hypothesis space.
- What if you're right for stupid reasons?

Generalizing Rademacher Complexity

We can generalize Rademacher complexity to consider all sets of a particular size.

$$\mathscr{R}_m(H) = \mathbb{E}_{S \sim D^m}\left[\hat{\mathscr{R}}_S(H)\right] \tag{8}$$

Generalizing Rademacher Complexity

Theorem

Convergence Bounds *Let F be a family of functions mapping from Z to $[0,1]$, and let sample $S = (z_1, \dots, z_m)$ were $z_i \sim D$ for some distribution D over Z. Define $\mathbb{E}[f] \equiv \mathbb{E}_{z \sim D}[f(z)]$ and $\hat{\mathbb{E}}_S[f] \equiv \frac{1}{m}\sum_{i=1}^{m} f(z_i)$. With probability greater than $1-\delta$ for all $f \in F$:*

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_s[f] + 2\mathscr{R}_m(F) + \mathscr{O}\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right) \tag{8}$$

Generalizing Rademacher Complexity

Theorem

Convergence Bounds *Let F be a family of functions mapping from Z to $[0,1]$, and let sample $S=(z_1,\ldots,z_m)$ were $z_i \sim D$ for some distribution D over Z. Define $\mathbb{E}[f] \equiv \mathbb{E}_{z\sim D}[f(z)]$ and $\hat{\mathbb{E}}_S[f] \equiv \frac{1}{m}\sum_{i=1}^m f(z_i)$. With probability greater than $1-\delta$ for all $f \in F$:*

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_s[f] + 2\mathscr{R}_m(F) + \mathscr{O}\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{8}$$

f is a surrogate for the accuracy of a hypothesis (mathematically convenient)

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1,\ldots,x_i,\ldots x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1,\ldots,x_m) \geq \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1,\ldots,x_i,\ldots x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1,\ldots,x_m) \geq \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Proofs online and in Mohri (requires Martingale, constructing $V_k = \mathbb{E}\left[V \,|\, x_1 \ldots x_k\right] - \mathbb{E}\left[V \,|\, x_1 \ldots x_{k-1}\right]$).

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1,\ldots,x_i,\ldots x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1,\ldots,x_m) \geq \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Proofs online and in Mohri (requires Martingale, constructing $V_k = \mathbb{E}\left[V \,|\, x_1 \ldots x_k\right] - \mathbb{E}\left[V \,|\, x_1 \ldots x_{k-1}\right]$).

What function do we care about for Rademacher complexity? Let's define

$$\Phi(S) = \sup_f \left(\mathbb{E}\left[f\right] - \hat{\mathbb{E}}_S\left[f\right]\right) = \sup_f \left(\mathbb{E}\left[f\right] - \frac{1}{m}\sum_i f(z_i)\right) \tag{11}$$

Step 1: Bounding divergence from true Expectation

Lemma

Moving to Expectation *With probability at least* $1-\delta$*,*
$\Phi(S) \leq \mathbb{E}_s[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$

Since $f(z_1) \in [0,1]$, changing any z_i to z_i' in the training set will change $\frac{1}{m}\sum_i f(z_i)$ by at most $\frac{1}{m}$, so we can apply McDiarmid's inequality with $\epsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$ and $c_i = \frac{1}{m}$.

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f (\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$\tag{13}$$

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \ldots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f \in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f]\right] - \hat{\mathbb{E}}_S[f])\right] \tag{13}$$

$$\tag{14}$$

The expectation is equal to the expectation of the empirical expectation of all sets S'

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f\in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f]\right] - \hat{\mathbb{E}}_S[f])\right] \tag{13}$$

$$= \mathbb{E}_S\left[\sup_{f\in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]\right])\right] \tag{14}$$

$$\tag{15}$$

S and S' are distinct random variables, so we can move inside the expectation

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f (\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f\in F} (\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]\right])\right] \tag{13}$$

$$\leq \mathbb{E}_{S,S'}\left[\sup_f (\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{14}$$

The expectation of a max over some function is at least the max of that expectation over that function

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_{S}[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] \tag{15}$$

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_{S}[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] \tag{15}$$

Let's introduce σ_i:

$$\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] = \frac{1}{m} \begin{cases} f(z_i) - f(z_i') & \text{with prob .5} \\ f(z_i') - f(z_i) & \text{with prob .5} \end{cases} \tag{16}$$

$$= \frac{1}{m} \sum_i \sigma_i (f(z_i') - f(z_i)) \tag{17}$$

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f] \tag{15}$$

Let's introduce σ_i:

$$\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f] = \frac{1}{m} \begin{cases} f(z_i) - f(z_i') \text{ with prob .5} \\ f(z_i') - f(z_i) \text{ with prob .5} \end{cases} \tag{16}$$

$$= \frac{1}{m} \sum_i \sigma_i (f(z_i') - f(z_i)) \tag{17}$$

Thus:
$\mathbb{E}_{S,S'}\left[\sup_{f \in F}\left(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]\right)\right] = \mathbb{E}_{S,S',\sigma}\left[\sup_{f \in F}\left(\sum_i \sigma_i (f(z_i') - f(z_i))\right)\right]$.

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i\left(f(z'_i)-f(z_i)\right)\right]$

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i(f(z'_i) - f(z_i))\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z'_i) + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

$$\tag{19}$$

Taking the sup jointly must be less than or equal the individual sup.

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i\left(f(z_i') - f(z_i)\right)\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i') + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i')\right] + \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{19}$$

$$\tag{20}$$

Linearity

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i(f(z_i') - f(z_i))\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i') + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i')\right] + \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{19}$$

$$= \mathscr{R}_m(F) + \mathscr{R}_m(F) \tag{20}$$

Definition

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Step 1

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\sup_f \left(\mathbb{E}\left[f\right] - \hat{\mathbb{E}}_S\left[h\right] \right) \leq \mathbb{E}_S\left[\Phi(S)\right] + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Definition of Φ

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[h] \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Drop the sup, still true

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[h] \leq \mathbb{E}_{S,S'}\left[\sup_f (\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])\right] + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Step 2

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[h] \leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f \in F}\left(\sum_i \sigma_i (f(z'_i) - f(z_i))\right)\right] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \quad (21)$$

Step 3

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[h] \leq 2\mathscr{R}_m(F) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Step 4

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[h] \leq 2\mathscr{R}_m(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Recall that $\hat{\mathscr{R}}_S(F) \equiv \mathbb{E}_\sigma\left[\sup_f \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$, so we apply McDiarmid's inequality again (because $f \in [0,1]$):

$$\hat{\mathscr{R}}_S(F) \leq \mathscr{R}_m(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{22}$$

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}\left[f\right]-\hat{\mathbb{E}}_S\left[h\right] \leq 2\mathscr{R}_m(F)+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Recall that $\hat{\mathscr{R}}_S(F) \equiv \mathbb{E}_\sigma\left[\sup_f \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$, so we apply McDiarmid's inequality again (because $f \in [0,1]$):

$$\hat{\mathscr{R}}_S(F) \leq \mathscr{R}_m(F)+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{22}$$

Putting the two together:

$$\mathbb{E}\left[f\right] \leq \hat{\mathbb{E}}_s\left[f\right]+2\mathscr{R}_m(F)+\mathscr{O}\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{23}$$

What about hypothesis classes?

Define:

$$
\begin{aligned}
Z \equiv & X \times \{-1, +1\} && (24) \\
f_h(x, y) \equiv & \mathbb{1}\left[h(x) \neq y\right] && (25) \\
F_H \equiv & \{f_h : h \in H\} && (26)
\end{aligned}
$$

What about hypothesis classes?

Define:

$$Z \equiv X \times \{-1, +1\} \tag{24}$$

$$f_h(x, y) \equiv \mathbb{1}\left[h(x) \neq y\right] \tag{25}$$

$$F_H \equiv \{f_h : h \in H\} \tag{26}$$

We can use this to create expressions for generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\left[h(x) \neq y\right]\right] = \mathbb{E}\left[f_h\right] \tag{27}$$

$$\hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}\left[h(x_i) \neq y\right] = \hat{\mathbb{E}}_S\left[f_h\right] \tag{28}$$

What about hypothesis classes?

Define:

$$Z \equiv X \times \{-1, +1\} \tag{24}$$

$$f_h(x, y) \equiv \mathbb{1}\left[h(x) \neq y\right] \tag{25}$$

$$F_H \equiv \{f_h : h \in H\} \tag{26}$$

We can use this to create expressions for generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\left[h(x) \neq y\right]\right] = \mathbb{E}\left[f_h\right] \tag{27}$$

$$\hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}\left[h(x_i) \neq y\right] = \hat{\mathbb{E}}_S\left[f_h\right] \tag{28}$$

We can plug this into our theorem!

Generalization bounds

- We started with expectations

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_S[f] + 2\hat{\mathscr{R}}_S(F) + \mathscr{O}\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{29}$$

- We also had our definition of the generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\left[h(x) \neq y\right]\right] = \mathbb{E}\left[f_h\right] \quad \hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}\left[h(x_i) \neq y\right] = \hat{\mathbb{E}}_S\left[f_h\right]$$

- Combined with the previous result:

$$\hat{\mathscr{R}}_S(F_H) = \frac{1}{2}\hat{\mathscr{R}}_S(H) \tag{30}$$

- All together:

$$R(h) \leq \hat{R}(h) + \mathscr{R}_m(H) + \mathscr{O}\left(\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right) \tag{31}$$

Wrapup

- Interaction of data, complexity, and accuracy
- Still very theoretical
- Next up: How to evaluate generalizability of specific hypothesis classes

Recap

- Rademacher complexity provides nice guarantees

$$R(h) \leq \hat{R}(h) + \mathscr{R}_m(H) + \mathcal{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right) \tag{32}$$

- But in practice hard to compute for real hypothesis classes
- Is there a relationship with simpler combinatorial measures?

Growth Function

Define the **growth function** $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set H as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1,\ldots,x_m\} \in X} \left| \{ (h(x_1), \ldots, h(x_m) : h \in H \} \right| \tag{33}$$

Growth Function

Define the **growth function** $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set H as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1,\ldots,x_m\} \in X} \left| \{ (h(x_1), \ldots, h(x_m) : h \in H\} \right| \tag{33}$$

i.e., the number of ways m points can be classified using H.

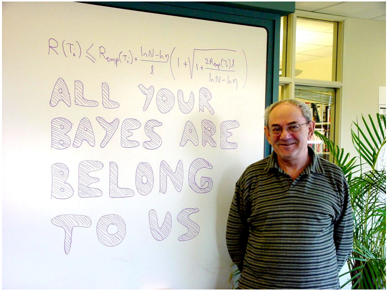
Rademacher Complexity vs. Growth Function

If G is a function taking values in $\{-1, +1\}$, then

$$\mathscr{R}_m(G) \leq \sqrt{\frac{2 \ln \Pi_G(m)}{m}} \tag{34}$$

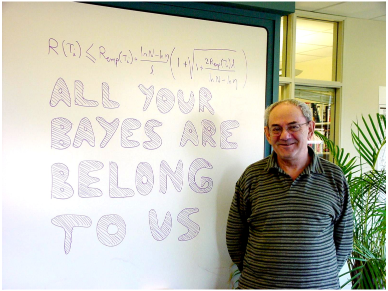
Uses Masart's lemma

Vapnik-Chervonenkis Dimension



$$\mathrm{VC}(H) \equiv \max\{m : \Pi_H(m) = 2^m\} \tag{35}$$

Vapnik-Chervonenkis Dimension



$$\mathrm{VC}(H) \equiv \max\{m : \Pi_H(m) = 2^m\} \tag{35}$$

The size of the largest set that can be fully shattered by H.

VC Dimension for Hypotheses

- Need upper and lower bounds
- Lower bound: example
- Upper bound: Prove that no set of $d+1$ points can be shattered by H (harder)

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

Intervals

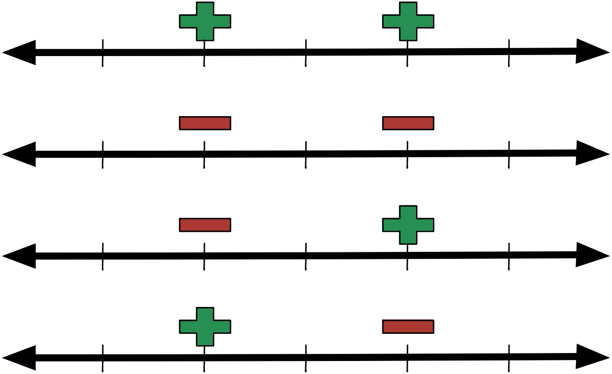
What is the VC dimension of $[a, b]$ intervals on the real line.

- What about two points?

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

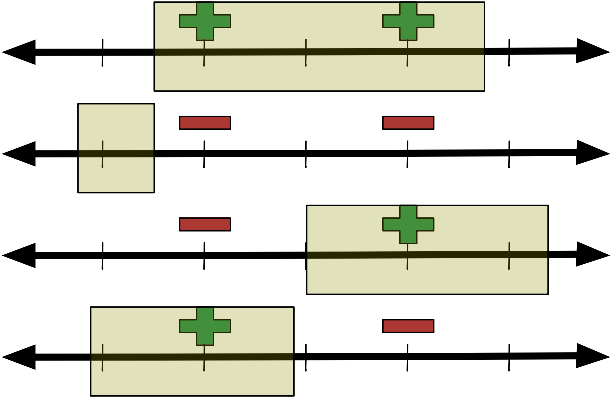
- What about two points?



Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- What about two points?



What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2

Intervals

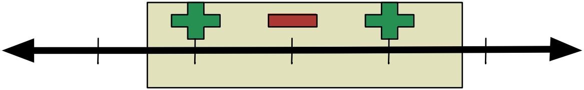
What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?



Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?
- **No set** of three points can be shattered

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?
- **No set** of three points can be shattered
- Thus, VC dimension of intervals is 2

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

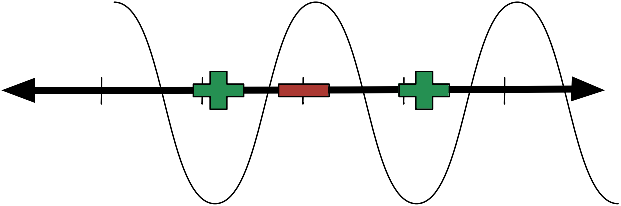
- Can you shatter three points?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Can you shatter three points?



Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \to \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Can you shatter four points?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

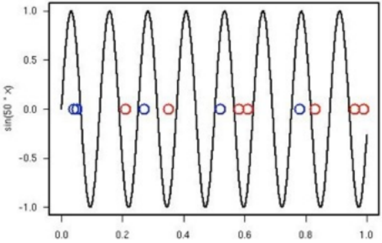
- How many points can you shatter?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Thus, VC dim of sine on line is ∞



Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Theorem

Sauer's Lemma *Let H be a hypothesis set with VC dimension d. Then $\forall m \in \mathbb{N}$*

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \equiv \Phi_d(m) \tag{37}$$

Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Theorem

Sauer's Lemma *Let H be a hypothesis set with VC dimension d. Then $\forall m \in \mathbb{N}$*

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \equiv \Phi_d(m) \tag{37}$$

This is good because the sum when multiplied out becomes $\binom{m}{i} = \frac{m \cdot (m-1) \dots}{i!} = \mathcal{O}\left(m^d\right)$. When we plug this into the learning error limits: $\log(\Pi_H(2m)) = \log(\mathcal{O}\left(m^d\right)) = \mathcal{O}\left(d \log m\right)$.

Proof of Sauer's Lemma

Prelim:

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \quad \text{This comes from Pascal's Triangle}$$

$$\binom{m}{k} = 0 \quad \text{if} \begin{cases} k < 0 \\ k > m \end{cases} \quad \text{This convention is consistent with Pascal's Triangle}$$

Proof of Sauer's Lemma

Prelim:

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \quad \text{This comes from Pascal's Triangle}$$
$$\binom{m}{k} = 0 \quad \text{if} \begin{cases} k < 0 \\ k > m \end{cases} \quad \text{This convention is consistent with Pascal's Triangle}$$

We'll proceed by induction. Our two base cases are:

- If $m = 0$, $\Pi_H(m) = 1$. You have no data, so there's only one (degenerate) labeling
- If $d = 0$, $\Pi_H(m) = 1$. If you can't even shatter a single point, then it's a fixed function

Induction Step

Assume that it holds for all m', d' for which $m' + d' < m + d$. We are given H, $|S| = m$, $S = \langle x_1, \dots, x_m \rangle$, and d is the VC dimension of H.

Induction Step

Assume that it holds for all m', d' for which $m' + d' < m + d$. We are given H, $|S| = m$, $S = \langle x_1, \ldots, x_m \rangle$, and d is the VC dimension of H.

Build two new hypothesis spaces

$\mathcal{H}$							$\mathcal{H}_1$						$\mathcal{H}_2$				
	$\mathbf{x}_1, \ldots, \mathbf{x}_m$							$\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}$						$\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}$			
h1	0	1	1	0	0	→	h1	0	1	1	0	→	h1	0	1	1	0
h2	0	1	1	0	1	↗											
h3	0	1	1	1	0	→	h3	0	1	1	1						
h4	1	0	0	1	0	→	h4	1	0	0	1	→	h4	1	0	0	1
h5	1	0	0	1	1	↗											
h6	1	1	0	0	1	→	h6	1	1	0	0						

Encodes where the extended set has differences on the first m points.

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (38) \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (39) \\
& && (40)
\end{aligned}
$$

Bounding Growth Function

$$|\Pi_H(S)| = |H_1| + |H_2| \quad (38)$$

$$\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \quad (39)$$

$$(40)$$

We can rewrite this as $\sum_{i=0}^{d} \binom{m-1}{i-1}$ because $\binom{x}{-1} = 0$.

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (38)\\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (39)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (40)\\
& && (41)
\end{aligned}
$$

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (38)\\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (39)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (40)\\
&= \sum_{i=0}^{d} \binom{m}{i} && (41)\\
& && (42)
\end{aligned}
$$

Pascal's Triangle

Bounding Growth Function

$$
\begin{aligned}
|\Pi_H(S)| &= |H_1| + |H_2| && (38)\\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} && (39)\\
&= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] && (40)\\
&= \sum_{i=0}^{d} \binom{m}{i} && (41)\\
&= \Phi_d(m) && (42)
\end{aligned}
$$

Wait a minute . . .

Is this combinatorial expression really $\mathcal{O}\left(m^d\right)$?

$$\begin{aligned}
\sum_{i=0}^{d}\binom{m}{i} &\leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m} \leq \left(\frac{m}{d}\right)^{d} e^{d}.
\end{aligned}$$

Generalization Bounds

Combining our previous generalization results with Sauer's lemma, we have that for a hypothesis class H with VC dimension d, for any $\delta > 0$ with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{43}$$

Whew!

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension

Whew!

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice ...

Whew!

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . . Support Vector Machines

Whew!

- We now have some theory down
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . . Support Vector Machines
- In class: more VC dimension examples