Foundations of Machine Learning

Learning with Infinite Hypothesis Sets

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Motivation

- With an infinite hypothesis set H, the error bounds of the previous lecture are not informative.
- Is efficient learning from a finite sample possible when H is infinite?
- Our example of axis-aligned rectangles shows that it is possible.
- Can we reduce the infinite case to a finite set? Project over finite samples?
- Are there useful measures of complexity for infinite hypothesis sets?

This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound

Empirical Rademacher Complexity

- Definition:
  - G family of functions mapping from set Z to $[a, b]$.
  - sample $S = (z_1, \ldots, z_m)$.
  - σ_is (Rademacher variables): independent uniform random variables taking values in $\{-1, +1\}$.

$$\widehat{\mathfrak{R}}_S(G) = \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \underbrace{\frac{1}{m} \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \end{bmatrix} \cdot \begin{bmatrix} g(z_1) \\ \vdots \\ g(z_m) \end{bmatrix}}_{\text{correlation with random noise}}\right] = \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right].$$

Rademacher Complexity

- **Definitions**: let G be a family of functions mapping from Z to $[a, b]$.
  - Empirical Rademacher complexity of G:

$$\widehat{\mathfrak{R}}_S(G) = \mathop{\mathrm{E}}_{\sigma}\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right],$$

    where σ_is are independent uniform random variables taking values in $\{-1, +1\}$ and $S = (z_1, \ldots, z_m)$.
  - Rademacher complexity of G:

$$\mathfrak{R}_m(G) = \mathop{\mathrm{E}}_{S \sim D^m}[\widehat{\mathfrak{R}}_S(G)].$$

Rademacher Complexity Bound

(Koltchinskii and Panchenko, 2002)

- Theorem: Let G be a family of functions mapping from Z to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $g \in G$:

$$\mathrm{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

$$\mathrm{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

- Proof: Apply McDiarmid's inequality to

$$\Phi(S) = \sup_{g \in G} \mathrm{E}[g] - \widehat{\mathrm{E}}_S[g].$$

- Changing one point of S changes $\Phi(S)$ by at most $\frac{1}{m}$.

$$
\begin{aligned}
\Phi(S') - \Phi(S) &= \sup_{g \in G} \{\mathrm{E}[g] - \widehat{\mathrm{E}}_{S'}[g]\} - \sup_{g \in G} \{\mathrm{E}[g] - \widehat{\mathrm{E}}_{S}[g]\} \\
&\leq \sup_{g \in G} \{\{\mathrm{E}[g] - \widehat{\mathrm{E}}_{S'}[g]\} - \{\mathrm{E}[g] - \widehat{\mathrm{E}}_{S}[g]\}\} \\
&= \sup_{g \in G} \{\widehat{\mathrm{E}}_{S}[g] - \widehat{\mathrm{E}}_{S'}[g]\} = \sup_{g \in G} \frac{1}{m}(g(z_m) - g(z'_m)) \leq \frac{1}{m}.
\end{aligned}
$$

- Thus, by McDiarmid's inequality, with probability at least $1 - \frac{\delta}{2}$

$$
\Phi(S) \leq \mathop{\mathrm{E}}_{S}[\Phi(S)] + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.
$$

- We are left with bounding the expectation.

- Series of observations:

$$
\begin{aligned}
\mathop{\mathrm{E}}_{S}[\Phi(S)] &= \mathop{\mathrm{E}}_{S}\Big[\sup_{g\in G} \mathrm{E}[g] - \widehat{\mathrm{E}}_S(g)\Big]\\
&= \mathop{\mathrm{E}}_{S}\Big[\sup_{g\in G} \mathop{\mathrm{E}}_{S'}[\widehat{\mathrm{E}}_{S'}(g) - \widehat{\mathrm{E}}_S(g)]\Big]\\
\text{(sub-add. of sup)} &\leq \mathop{\mathrm{E}}_{S,S'}\Big[\sup_{g\in G} \widehat{\mathrm{E}}_{S'}(g) - \widehat{\mathrm{E}}_S(g)\Big]\\
&= \mathop{\mathrm{E}}_{S,S'}\Big[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}(g(z'_i) - g(z_i))\Big]\\
(\text{swap } z_i \text{ and } z'_i) &= \mathop{\mathrm{E}}_{\sigma,S,S'}\Big[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z'_i) - g(z_i))\Big]\\
\text{(sub-additiv. of sup)} &\leq \mathop{\mathrm{E}}_{\sigma,S'}\Big[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z'_i)\Big] + \mathop{\mathrm{E}}_{\sigma,S}\Big[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i)\Big]\\
&= 2\mathop{\mathrm{E}}_{\sigma,S}\Big[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\Big] = 2\mathfrak{R}_m(G).
\end{aligned}
$$

- Now, changing one point of S makes $\widehat{\mathfrak{R}}_S(G)$ vary by at most $\frac{1}{m}$. Thus, again by McDiarmid's inequality, with probability at least $1-\frac{\delta}{2}$,

$$\mathfrak{R}_m(G) \leq \widehat{\mathfrak{R}}_S(G) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

- Thus, by the union bound, with probability at least $1-\delta$,

$$\Phi(S) \leq 2\widehat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Loss Functions - Hypothesis Set

- Proposition: Let H be a family of functions taking values in $\{-1, +1\}$, G the family of zero-one loss functions of H: $G = \{(x, y) \mapsto 1_{h(x) \neq y} : h \in H\}$. Then,

$$\mathfrak{R}_m(G) = \frac{1}{2} \mathfrak{R}_m(H).$$

- Proof: $$\begin{aligned}\mathfrak{R}_m(G) &= \underset{S,\sigma}{\mathrm{E}}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i 1_{h(x_i) \neq y_i}\right] \\ &= \underset{S,\sigma}{\mathrm{E}}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i \frac{1}{2}(1 - y_i h(x_i))\right] \\ &= \underbrace{\frac{1}{2} \underset{S,\sigma}{\mathrm{E}}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i\right]}_{=0} + \frac{1}{2} \underset{S,\sigma}{\mathrm{E}}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m -\sigma_i y_i h(x_i)\right] \\ &= \frac{1}{2} \underset{S,\sigma}{\mathrm{E}}\left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i)\right].\end{aligned}$$

Generalization Bounds - Rademacher

- **Corollary**: Let H be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \mathfrak{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

$$R(h) \leq \widehat{R}(h) + \widehat{\mathfrak{R}}_S(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Remarks

- First bound distribution-dependent, second data-dependent bound, which makes them attractive.
- But, how do we compute the empirical Rademacher complexity?
- Computing $\mathrm{E}_\sigma[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i)]$ requires solving ERM problems, typically computationally hard.
- Relation with combinatorial measures easier to compute?

This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound

Growth Function

- Definition: the growth function $\Pi_H\colon \mathbb{N} \to \mathbb{N}$ for a hypothesis set H is defined by

$$\forall m \in \mathbb{N},\ \Pi_H(m) = \max_{\{x_1,\ldots,x_m\} \subseteq X} \Big| \Big\{ \big(h(x_1), \ldots, h(x_m) \big) \colon h \in H \Big\} \Big|.$$

- Thus, $\Pi_H(m)$ is the maximum number of ways m points can be classified using H.

Massart's Lemma

(Massart, 2000)

- **Theorem**: Let $A \subseteq \mathbb{R}^m$ be a finite set, with $R = \max_{x \in A} \|x\|_2$, then, the following holds:

$$\mathop{\mathrm{E}}_{\sigma}\left[\frac{1}{m} \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i\right] \leq \frac{R\sqrt{2\log|A|}}{m}.$$

- **Proof**: $\exp\left(t \mathop{\mathrm{E}}_{\sigma}\left[\sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i\right]\right) \leq \mathop{\mathrm{E}}_{\sigma}\left(\exp\left[t \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i\right]\right)$ (Jensen's ineq.)

$$= \mathop{\mathrm{E}}_{\sigma}\left(\sup_{x \in A} \exp\left[t \sum_{i=1}^{m} \sigma_i x_i\right]\right)$$

$$\leq \sum_{x \in A} \mathop{\mathrm{E}}_{\sigma}\left(\exp\left[t \sum_{i=1}^{m} \sigma_i x_i\right]\right) = \sum_{x \in A} \prod_{i=1}^{m} \mathop{\mathrm{E}}_{\sigma}\left(\exp\left[t \sigma_i x_i\right]\right)$$

(Hoeffding's ineq.) $\leq \sum_{x \in A}\left(\exp\left[\frac{\sum_{i=1}^{m} t^2 (2|x_i|)^2}{8}\right]\right) \leq |A| e^{\frac{t^2 R^2}{2}}.$

- Taking the log yields:

$$\mathop{\mathrm{E}}_{\sigma}\left[\sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i\right] \leq \frac{\log |A|}{t} + \frac{tR^2}{2}.$$

- Minimizing the bound by choosing $t = \frac{\sqrt{2 \log |A|}}{R}$ gives

$$\mathop{\mathrm{E}}_{\sigma}\left[\sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i\right] \leq R\sqrt{2 \log |A|}.$$

Growth Function Bound on Rad. Complexity

- **Corollary**: Let G be a family of functions taking values in $\{-1, +1\}$, then the following holds:

$$\mathfrak{R}_m(G) \leq \sqrt{\frac{2 \log \Pi_G(m)}{m}}.$$

- **Proof**:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(G) &= \mathop{\mathrm{E}}_{\sigma}\left[\sup_{g \in G} \frac{1}{m} \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \end{bmatrix} \cdot \begin{bmatrix} g(z_1) \\ \vdots \\ g(z_m) \end{bmatrix}\right] \\
&\leq \frac{\sqrt{m}\sqrt{2 \log |\{(g(z_1), \ldots, g(z_m)) \colon g \in G\}|}}{m} \qquad \text{(Massart's Lemma)} \\
&\leq \frac{\sqrt{m}\sqrt{2 \log \Pi_G(m)}}{m} = \sqrt{\frac{2 \log \Pi_G(m)}{m}}.
\end{aligned}$$

Generalization Bound - Growth Function

- Corollary: Let H be a family of functions taking values in $\{-1,+1\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2\log \Pi_H(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- But, how do we compute the growth function? Relationship with the VC-dimension (Vapnik-Chervonenkis dimension).

This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound

VC Dimension

(Vapnik & Chervonenkis, 1968-1971; Vapnik, 1982, 1995, 1998)

- Definition: the VC-dimension of a hypothesis set H is defined by

$$\mathrm{VCdim}(H) = \max\{m \colon \Pi_H(m) = 2^m\}.$$

- Thus, the VC-dimension is the size of the largest set that can be fully shattered by H.

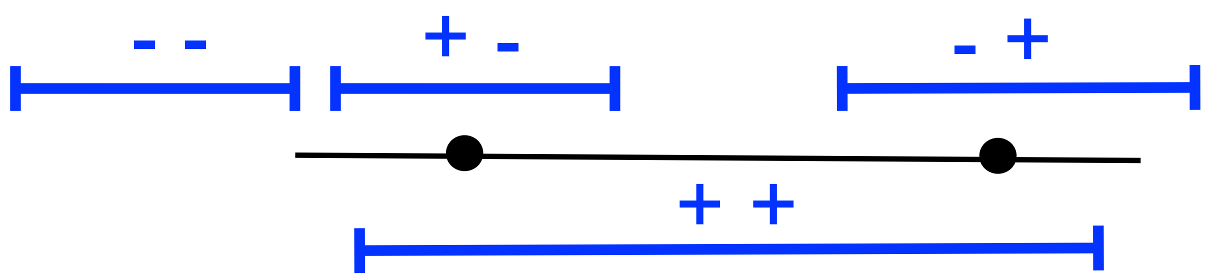
- Purely combinatorial notion.

Examples

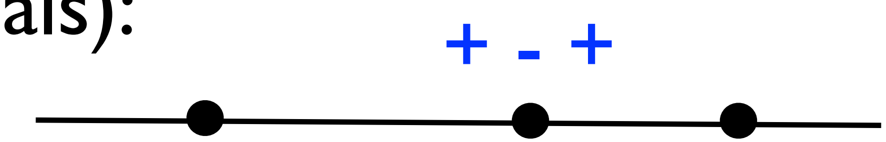
- In the following, we determine the VC dimension for several hypothesis sets.
- To give a lower bound d for $\mathrm{VCdim}(H)$, it suffices to show that a set S of cardinality d can be shattered by H.
- To give an upper bound, we need to prove that no set S of cardinality $d+1$ can be shattered by H, which is typically more difficult.

Intervals of The Real Line

- **Observations:**
  - Any set of two points can be shattered by four intervals



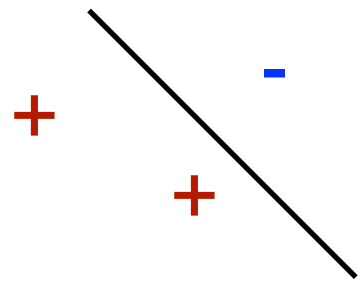
  - No set of three points can be shattered since the following dichotomy "+ - +" is not realizable (by definition of intervals):



  - Thus, $\text{VCdim}(\text{intervals in } \mathbb{R}) = 2$.

Hyperplanes

- Observations:
  - Any three non-collinear points can be shattered:

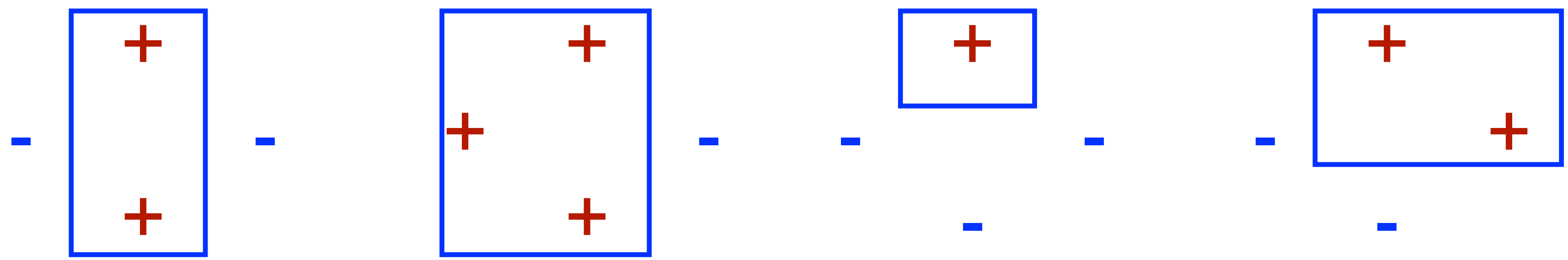


  - Unrealizable dichotomies for four points:

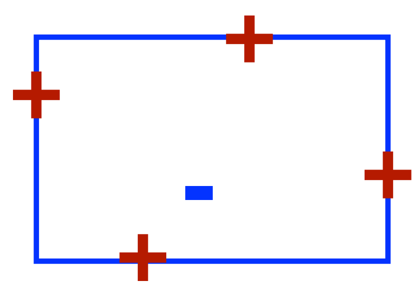
  - Thus, $\mathrm{VCdim}(\text{hyperplanes in } \mathbb{R}^d) = d+1$.

Axis-Aligned Rectangles in the Plane

- **Observations:**
  - The following four points can be shattered:



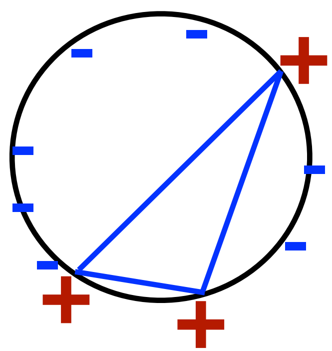
  - No set of five points can be shattered: label negatively the point that is not near the sides.



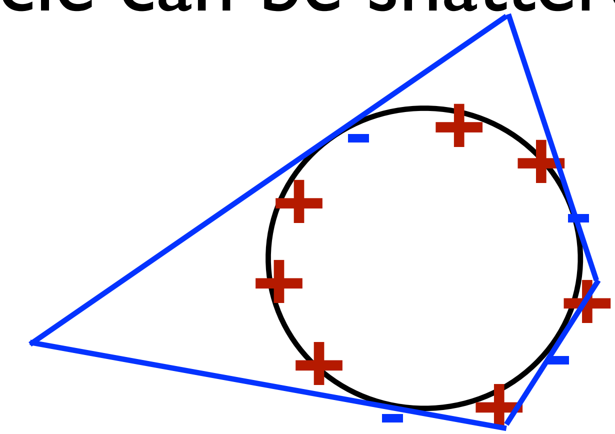
  - Thus, $\mathrm{VCdim}(\text{axis-aligned rectangles}) = 4$.

Convex Polygons in the Plane

- Observations:
  - $2d+1$ points on a circle can be shattered by a *d*-gon:



|positive points| < |negative points|

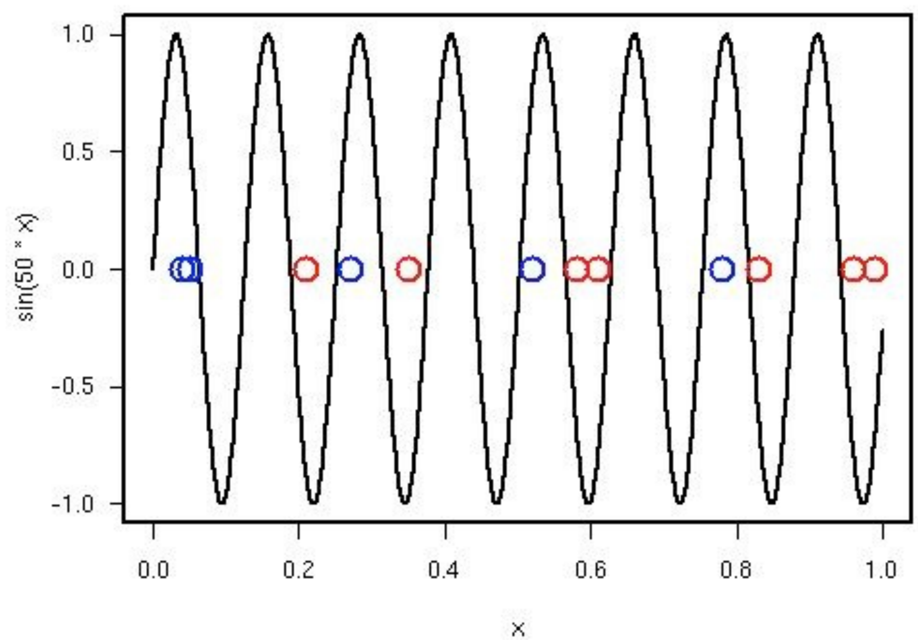


|positive points| > |negative points|

  - It can be shown that choosing the points on the circle maximizes the number of possible dichotomies. Thus, $\mathrm{VCdim}(\text{convex } d\text{-gons}) = 2d+1$. Also, $\mathrm{VCdim}(\text{convex polygons}) = +\infty$.

Sine Functions

- Observations:
  - Any finite set of points on the real line can be shattered by $\{t \mapsto \sin(\omega t) : \omega \in \mathbb{R}\}$.
  - Thus, $\text{VCdim}(\text{sine functions}) = +\infty$.



Sauer's Lemma

(Vapnik & Chervonenkis, 1968-1971; Sauer, 1972)

- Theorem: let H be a hypothesis set with $\mathrm{VCdim}(H) = d$ then, for all $m \in \mathbb{N}$,
$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$
- Proof: the proof is by induction on $m+d$. The statement clearly holds for $m=1$ and $d=0$ or $d=1$. Assume that it holds for $(m-1, d-1)$ and $(m-1, d)$.
    - Fix a set $S = \{x_1, \ldots, x_m\}$ with $\Pi_H(m)$ dichotomies and let $G = H_{|S}$ be the set of concepts H induces by restriction to S.

- Consider the following families over $S' = \{x_1, \ldots, x_{m-1}\}$:

$$G_1 = G_{|S'} \quad G_2 = \{g' \subseteq S' \colon (g' \in G) \wedge (g' \cup \{x_m\} \in G)\}.$$

x_1	x_2	$\cdots$	x_{m-1}	x_m
1	1	0	1	0
1	1	0	1	1
0	1	1	1	1
1	0	0	1	0
1	0	0	0	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

- Observe that $|G_1| + |G_2| = |G|$.

- Since $\mathrm{VCdim}(G_1) \leq d$, by the induction hypothesis,

$$|G_1| \leq \Pi_{G_1}(m-1) \leq \sum_{i=0}^{d} \binom{m-1}{i}.$$

- By definition of G_2, if a set $Z \subseteq S'$ is shattered by G_2, then the set $Z \cup \{x_m\}$ is shattered by G. Thus,

$$\mathrm{VCdim}(G_2) \leq \mathrm{VCdim}(G) - 1 = d - 1$$

and by the induction hypothesis,

$$|G_2| \leq \Pi_{G_2}(m-1) \leq \sum_{i=0}^{d-1} \binom{m-1}{i}.$$

- Thus, $|G| \leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i}$
$= \sum_{i=0}^{d} \binom{m-1}{i} + \binom{m-1}{i-1} = \sum_{i=0}^{d} \binom{m}{i}.$

Sauer's Lemma - Consequence

- **Corollary**: let H be a hypothesis set with $\mathrm{VCdim}(H)=d$ then, for all $m \geq d$,

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

- **Proof**:

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

Remarks

- Remarkable property of growth function:
  - either $\mathrm{VCdim}(H) = d < +\infty$ and $\Pi_H(m) = O(m^d)$
  - or $\mathrm{VCdim}(H) = +\infty$ and $\Pi_H(m) = 2^m$.

Generalization Bound - VC Dimension

- Corollary: Let H be a family of functions taking values in $\{-1,+1\}$ with VC dimension d. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- Proof: Corollary combined with Sauer's lemma.
- Note: The general form of the result is

$$R(h) \leq \widehat{R}(h) + O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right).$$

Comparison - Standard VC Bound

(Vapnik & Chervonenkis, 1971; Vapnik, 1982)

- **Theorem**: Let H be a family of functions taking values in $\{-1, +1\}$ with VC dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}{m}}.$$

- **Proof**: Derived from growth function bound

$$\Pr\left[\left|R(h) - \widehat{R}(h)\right| > \epsilon\right] \leq 4\Pi_H(2m) \exp\left(-\frac{m\epsilon^2}{8}\right).$$

This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound

VCDim Lower Bound - Realizable Case

(Ehrenfeucht et al., 1988)

- Theorem: let H be a hypothesis set with VC-dimension $d>1$. Then, for any learning algorithm L,

$$\exists D, \exists f \in H, \ \Pr_{S \sim D^m}\left[R_D(h_S, f) > \frac{d-1}{32m}\right] \geq 1/100.$$

- Proof: choose D such that L can do no better than tossing a coin for some points.
  - Let $X = \{x_0, x_1, \ldots, x_{d-1}\}$ be a set fully shattered. For any $\epsilon > 0$, define D with support X by

$$\Pr_D[x_0] = 1 - 8\epsilon \quad \text{and} \quad \forall i \in [1, d-1], \Pr_D[x_i] = \frac{8\epsilon}{d-1}.$$

- We can assume without loss of generality that L makes no error on x_0.
- For a sample S, let $\overline{S}$ denote the set of its elements falling in $X_1 = \{x_1, \ldots, x_{d-1}\}$ and let $\mathcal{S}$ be the set of samples of size m with at most $(d-1)/2$ points in X_1.
- Fix a sample $S \in \mathcal{S}$. Using $|X - \overline{S}| \geq (d-1)/2$,

$$\begin{aligned}
\mathop{\mathrm{E}}_{f \sim U}[R_D(h_S, f)] &= \sum_f \sum_{x \in X} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
&\geq \sum_f \sum_{x \notin \overline{S}} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
&= \sum_{x \notin \overline{S}} \bigg(\sum_f 1_{h(x) \neq f(x)} \Pr[f] \bigg) \Pr[x] \\
&= \frac{1}{2} \sum_{x \notin \overline{S}} \Pr[x] \geq \frac{1}{2} \frac{d-1}{2} \frac{8\epsilon}{d-1} = 2\epsilon.
\end{aligned}$$

- Since the inequality holds for all $S \in \mathcal{S}$, it also holds in expectation: $\mathrm{E}_{S, f \sim U}[R_D(h_S, f)] \geq 2\epsilon$. This implies that there exists a labeling f_0 such that $\mathrm{E}_S[R_D(h_S, f_0)] \geq 2\epsilon$.
- Since $\Pr_D[X - \{x_0\}] \leq 8\epsilon$, we also have $R_D(h_S, f_0) \leq 8\epsilon$. Thus,

$$2\epsilon \leq \mathop{\mathrm{E}}_{S}[R_D(h_S, f_0)] \leq 8\epsilon \Pr_{S \in \mathcal{S}}[R_D(h_S, f_0) \geq \epsilon] + (1 - \Pr_{S \in \mathcal{S}}[R_D(h_S, f_0) \geq \epsilon])\epsilon.$$

- Collecting terms in $\Pr_{S \in \mathcal{S}}[R_D(h_S, f_0) \geq \epsilon]$, we obtain:

$$\Pr_{S \in \mathcal{S}}[R_D(h_S, f_0) \geq \epsilon] \geq \frac{1}{7\epsilon}(2\epsilon - \epsilon) = \frac{1}{7}.$$

- Thus, the probability over all samples S (not necessarily in $\mathcal{S}$) can be lower bounded as

$$\Pr_{S}[R_D(h_S, f_0) \geq \epsilon] \geq \Pr_{S \in \mathcal{S}}[R_D(h_S, f_0) \geq \epsilon] \Pr[\mathcal{S}] \geq \frac{1}{7} \Pr[\mathcal{S}].$$

- This leads us to seeking a lower bound for $\Pr[\mathcal{S}]$. The probability that more than $(d-1)/2$ points be drawn in a sample of size m verifies the Chernoff bound for any $\gamma > 0$:

$$1 - \Pr[\mathcal{S}] = \Pr[S_m \geq 8\epsilon m(1+\gamma)] \leq e^{-8\epsilon m \frac{\gamma^2}{3}}.$$

- Thus, for $\epsilon = (d-1)/(32m)$ and $\gamma = 1$,

$$\Pr[S_m \geq \tfrac{d-1}{2}] \leq e^{-(d-1)/12} \leq e^{-1/12} \leq 1 - 7\delta,$$

for $\delta \leq .01$. Thus, $\Pr[\mathcal{S}] \geq 7\delta$ and

$$\Pr_S[R_D(h_S, f_0) \geq \epsilon] \geq \delta.$$

Agnostic PAC Model

- Definition: concept class C is PAC-learnable if there exists a learning algorithm L such that:
  - for all $c \in C, \epsilon > 0, \delta > 0$, and all distributions D,

$$\Pr_{S \sim D}\left[R(h_S) - \inf_{h \in H} R(h) \leq \epsilon\right] \geq 1 - \delta,$$

  - for samples S of size $m = poly(1/\epsilon, 1/\delta)$ for a fixed polynomial.

VCDim Lower Bound - Non-Realizable Case

(Anthony and Bartlett, 1999)

- Theorem: let H be a hypothesis set with VC dimension $d > 1$. Then, for any learning algorithm L,

$$\exists D \text{ over } X \times \{0, 1\},$$

$$\Pr_{S \sim D^m}\left[R_D(h_S) - \inf_{h \in H} R_D(h) > \sqrt{\frac{d}{320m}}\right] \geq 1/64.$$

- Equivalently, for any learning algorithm, the sample complexity verifies

$$m \geq \frac{d}{320\epsilon^2}.$$

References

- Martin Anthony, Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press. 1999.

- Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, Volume 36, Issue 4, 1989.

- A. Ehrenfeucht, David Haussler, Michael Kearns, Leslie Valiant. A general lower bound on the number of examples needed for learning. Proceedings of *1st COLT*. pp. 139-154, 1988.

- Koltchinskii, Vladimir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics, 30(1), 2002.

- Pascal Massart. Some applications of concentration inequalities to statistics. Annales de la Faculte des Sciences de Toulouse, IX:245–303, 2000.

- N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory* (A), 13:145-147, 1972.

References

- Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, 1982.
- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.
- Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.
- Vladimir N. Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow (in Russian). 1974.
- Vladimir N. Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and its Appl.*, vol. 16, no. 2, pp. 264-280, 1971.