Introduction to Machine Learning

Foundations and Applications

Paul J. Atzberger
University of California Santa Barbara

Hypothesis Class Complexity

Motivations

Hypothesis classes are typically infinite $|\mathcal{H}| = \infty$.

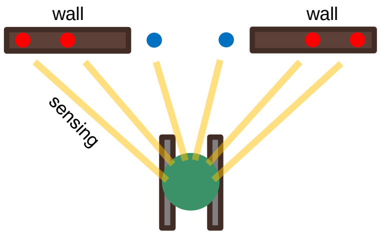
Can we still efficiently learn concepts c?

Yes, recall interval problem and axis-aligned rectangle problem was infinite but PAC-Learnable.

We need a notion of complexity for hypothesis class $\mathcal{H}$ beyond cardinality $|\mathcal{H}|$.

Ultimately, we aim to obtain bounds on the generalization error in terms of the empirical risk.

Obstacle Navigation



Building Identification

Google Maps: UCSB South Hall

Picture Annotation, Facial Recognition

usplash

Rademacher Complexity

Hans Rademacher
1892-1969

Notation and definitions:

$\mathcal{X}$ input space, $\mathcal{Y}$ output space
$\mathcal{C}$ concept class, concept $c(x)$: $\mathcal{X} \to \mathcal{Y}$
$\mathcal{H}$ hypothesis class, hypothesis $h(x)$: $\mathcal{X} \to \mathcal{Y}$.

Issue: Hypothesis classes are typically infinite $|\mathcal{H}| = \infty$. **Can we still efficiently learn concepts c?**

Recall: Axis-aligned rectangle problem has infinite $|\mathcal{H}| = \infty$ but proved is PAC-Learnable.

Need a notion of complexity for hypothesis class $\mathcal{H}$ beyond cardinality $|\mathcal{H}|$.

$$h \in \mathcal{H}, \quad g(x,y) = L(h(x),y)$$

Let **loss function** be denoted L: $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and let G be **family of loss functions** associated with $\mathcal{H}$.

Definition: The **empirical Rademacher complexity** of a family of functions G with $g(z)$: $Z \to [a,b] \subset \mathbb{R}$ and m fixed samples $S = (z_1, z_2, \ldots, z_m)$ is given by

$$\hat{\mathcal{R}}_S(G) = E_\sigma\left[\sup_{g \in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\right]$$

, where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ are uniform random variables in $\{-1,+1\}$.

Rademacher Complexity

Hans Rademacher
1892-1969

Notation and definitions:

$\mathcal{X}$ input space, $\mathcal{Y}$ output space
$\mathcal{C}$ concept class, concept c(x): $\mathcal{X} \rightarrow \mathcal{Y}$
$\mathcal{H}$ hypothesis class, hypothesis h(x): $\mathcal{X} \rightarrow \mathcal{Y}$.

Definition: The **empirical Rademacher complexity** of a family of functions G with g(z): Z $\rightarrow$ [a,b] $\subset \mathbb{R}$ and m fixed samples S = $(z_1, z_2, \ldots, z_m)$ is given by

$$\hat{\mathcal{R}}_S(G) = E_\sigma \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right]$$

, where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ are uniform random variables in {-1,+1}.

Definition: The **Rademacher complexity** of a family of functions G on m samples is

$$\mathcal{R}_m(G) = E_{S \sim D^m} \left[\hat{\mathcal{R}}_S(G) \right]$$

- **Averaged sum term** can be viewed as an inner-product: $\sum \sigma_i \cdot g(z_i) = \boldsymbol{\sigma} \cdot \mathbf{g}_S$.

- **Rademacher complexity** gives a measure of the "richness" of family G in approximating random functions.

$$\mathcal{R}_m(G) = E_{S \sim D^m, \sigma} \left[\sup_{g \in G} \frac{1}{m} \sigma \cdot \mathbf{g}_S \right].$$

Gives a measure of the "correlation" between $\mathbf{g}_S$ and $\boldsymbol{\sigma}$.

Rademacher Complexity

Definition: The **empirical Rademacher complexity** of a family of functions G with g(z): Z → [a,b] ⊂ ℝ and m fixed samples S = $(z_1,z_2,\ldots,z_m)$ is given by

$$\hat{\mathcal{R}}_S(G) = E_\sigma\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right]$$, where **σ** = $(\sigma_1, \sigma_2, \ldots, \sigma_m)$ are random in set {-1,+1}.

Example: Rademacher Complexity for family of functions $\mathrm{G} = \{\mathrm{g(z)} = g_0 \in [-\mathrm{c}, c]\}$ (constants).

$$\begin{aligned}\tilde{\mathcal{R}}_S(\mathcal{G}) &= E_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] = E_\sigma\left[\max\left\{\frac{1}{m}\sum_{i=1}^{m}c\sigma_i, -\frac{1}{m}\sum_{i=1}^{m}c\sigma_i\right\}\right]\\ &= E_\sigma\left[\frac{1}{m}c\left|\#\{\sigma_i=+1\}-\#\{\sigma_i=-1\}\right|\right] = \frac{c}{m}E_\sigma\left[\left|\sum_{i=1}^{m}\sigma_i\right|\right] \leq \frac{c\sqrt{m}}{m} = \frac{c}{\sqrt{m}}\end{aligned}$$

Jensen Inequality (ϕ convex):

$$\phi(E[X]) \leq E[\phi(X)]$$

$$(E[|X|])^2 \leq E\left[|X|^2\right]$$

$$\begin{aligned}E\left[\left|\sum_{i=1}^{m}\sigma_i\right|\right] &\leq E\left[\left|\sum_{i=1}^{m}\sigma_i\right|^2\right]^{1/2}\\ &= E\left[\sum_{i,j=1}^{m}\sigma_i\sigma_j\right]^{1/2} = \sqrt{m}\end{aligned}$$

Rademacher Complexity

Hans Rademacher
1892-1969

Notation and definitions:

$\mathcal{X}$ input space, $\mathcal{Y}$ output space
$\mathcal{C}$ concept class, concept c(x): $\mathcal{X} \rightarrow \mathcal{Y}$
$\mathcal{H}$ hypothesis class, hypothesis h(x): $\mathcal{X} \rightarrow \mathcal{Y}$.

Theorem: (expectation bounds g: Z → [0,1]) For family of loss functions G into [0,1] and any $\delta > 0$ we have with probability $1-\delta$ that the following bounds hold uniformly for any $g \in G$,

$$E\left[g(z)\right] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}, \quad \text{(Rademacher bound)}$$

$$E\left[g(z)\right] \leq \underbrace{\frac{1}{m}\sum_{i=1}^{m} g(z_i)}_{\text{empirical estimate}} + \underbrace{2\hat{\mathcal{R}}_S(G)}_{\text{model complexity}} + \underbrace{3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}}_{\text{sampling confidence}}, \quad \text{(Empirical Rademacher bound)}$$

Significance: The expected value E[g] can be bounded by the observed empirical average. This differs at most by the Rademacher Complexity plus a term vanishing as m → ∞.

We shall use for bound on the **generalization error** by the **empirical risk**.

Rademacher Complexity

Hans Rademacher
1892-1969

Notation and definitions:

$\mathcal{X}$ input space, $\mathcal{Y}$ output space
$\mathcal{C}$ concept class, concept c(x): $\mathcal{X} \rightarrow \mathcal{Y}$
$\mathcal{H}$ hypothesis class, hypothesis h(x): $\mathcal{X} \rightarrow \mathcal{Y}$.

Definition: The **Empirical Rademacher Complexity** of *a hypothesis class* $\mathcal{H}$ is

$$\hat{\mathcal{R}}_{S^{\mathcal{X}}}(H) = E_{\sigma}\left[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right]$$, (note: we take h ∈ {-1,1})

Definition: The **Rademacher Complexity** of *a hypothesis class* $\mathcal{H}$ is

$$\mathcal{R}_m(H) = E_{S\sim D^m}\left[\hat{\mathcal{R}}_{S^{\mathcal{X}}}(H)\right]$$, (note: we take h ∈ {-1,1})

Lemma: For the family of 0-1 loss functions $G = \{(x,y) \rightarrow 1_{h(x)\neq y} | h \in H\}$ we have

$$\hat{\mathcal{R}}_S(G) = \frac{1}{2}\hat{\mathcal{R}}_{S^{\mathcal{X}}}(H)$$

- **Allows for working more directly** with the hypothesis space in constructing bounds.

Rademacher Complexity

Hans Rademacher
1892-1969

Notation and definitions:

$\mathcal{X}$ input space, $\mathcal{Y}$ output space
$\mathcal{C}$ concept class, concept c(x): $\mathcal{X} \to \mathcal{Y}$
$\mathcal{H}$ hypothesis class, hypothesis h(x): $\mathcal{X} \to \mathcal{Y}$.

Theorem: (bound on generalization error for 0-1 loss) For 0 - 1 loss $G = \{(x,y) \to 1_{h(x)\neq y} | h \in H\}$ and any $\delta > 0$ we have with probability $1 - \delta$ that the following bounds hold uniformly for any $g \in G$,

$$R(h) \leq \hat{R}_S(h) + \mathcal{R}_m(H) + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$, (Rademacher bound)

$$R(h) \leq \underbrace{\hat{R}_S(h)}_{\text{empirical estimate}} + \underbrace{\hat{\mathcal{R}}_{S^{\mathcal{X}}}(H)}_{\text{model complexity}} + \underbrace{3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}}_{\text{sampling confidence}}$$, (Empirical Rademacher bound)

Significance: The **generalization error** can be bounded by the observed **empirical risk**. This differs most by the Rademacher Complexity plus a term vanishing as m → ∞.

- This shows we can use Rademacher complexity in place of $|\mathcal{H}|$ to obtain bounds on **generalization error** to obtain **scaling in m**.

Rademacher Complexity

Theorem: (bound on generalization error for 0-1 loss) For 0 - 1 loss $G = \{(x,y) \to 1_{h(x)\neq y} | h \in H\}$ and any $\delta > 0$ we have with probability $1-\delta$ that the following bounds hold uniformly for any $g \in G$,

$$R(h) \leq \hat{R}_S(h) + \hat{\mathcal{R}}_{S^x}(H) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}},$$ (Empirical Rademacher bound)

Example: Rademacher Complexity for family of functions $\mathrm{H} = \{\mathrm{h(x)} = h_0 \in [-\mathrm{c}, c], c = 1\}$ (constants).

$$\tilde{\mathcal{R}}_S(\mathcal{G}) = E_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] \leq \frac{c\sqrt{m}}{m} = \frac{c}{\sqrt{m}}$$ (from previous derivation)

$$\tilde{\mathcal{R}}_S(\mathcal{H}) = 2\tilde{\mathcal{R}}(\mathcal{G}) = \frac{2c}{\sqrt{m}}$$

Rademacher Complexity

Theorem: (bound on generalization error for 0-1 loss) For 0 - 1 loss $G = \{(x, y) \to 1_{h(x) \neq y} | h \in H\}$ and any $\delta > 0$ we have with probability $1 - \delta$ that the following bounds hold uniformly for any $g \in G$,

$$R(h) \leq \hat{R}_S(h) + \hat{\mathcal{R}}_{S^{\mathcal{X}}}(H) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$, (Empirical Rademacher bound)

Theorem (Massert's lemma): Let $A \subseteq \mathbb{R}^n$ be a finite set of vectors with $r = \max_{a \in A} \|a\|_2$ then

$$\hat{\mathcal{R}}_S(A) = E_\sigma \left[\sup_{\mathbf{a} \in A} \frac{1}{m} \sum_{i=1}^{m} \sigma_i a_i \right] \leq \frac{r\sqrt{2\log|A|}}{m}$$

Hypothesis class $\mathcal{H}$ and m samples consider the set A = $\{(h(x_1), h(x_2), \ldots, h(x_m)) : h \in \mathcal{H}\}$.

Finite hypothesis class we have $|A| \leq |\mathcal{H}|$.

Note: Result similar to prior complexity bound for finite consistent case

$$R(h_S) \leq \frac{1}{m}\left(\log|H| + \log\frac{1}{\delta}\right)$$

Massert's Lemma significantly generalizes this result since |A| is now allowed to grow with m for $|\mathcal{H}| = \infty$.

Alternatively, combinatorial measures like complexity |A| may be easier to estimate than Rademacher complexity.

Growth Function

Definition: The **growth function** $\Pi_H: \mathbb{N} \to \mathbb{N}$ for a hypothesis class $\mathcal{H}$ is defined as

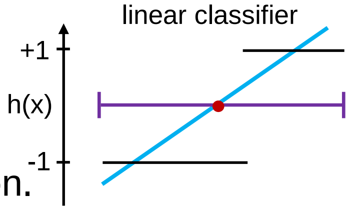
$$\Pi_H(m) = \max_{\{x_1, x_2, \ldots, xm\} \subseteq X} \left|\{\left(h(x_1), h(x_2), \ldots, h(x_m)\right): h \in \mathcal{H}\}\right|$$

Counts maximum number of distinct m-vectors $\left(h(x_1), h(x_2), \ldots, h(x_m)\right)$ that can be generated by the hypothesis class $\mathcal{H}$.

Upper bound on the number of distinct ways m points can be classified by $\mathcal{H}$.

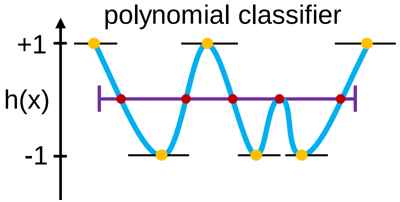
Example: $\mathcal{X}$ = {-2,-1,1}, $\mathcal{Y}$ = {-1,1}, $\mathcal{H}$ = {$h_1(x)$ = sign(x), $h_2(x)$ = sign(x – 1.5)}, h_1: -1,-1,1; h_2: -1,-1,-1. For m=2, most variation for x_1=-1, x_2=1, with $\Pi_H(2) = |\{(-1,+1),(-1,-1)\}|$ = 2. In general, we have $\Pi_H(m)$ = 2.

Remark: For finite hypothesis class always have $\Pi_H(m) \leq |\mathcal{H}|$.



Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y}$ = {-1,1}, $\mathcal{H}$ = {h(x) = sign(p(x)) with p(x) polynomial degree n}.
Now $|\mathcal{H}| = \infty$ and we have $\Pi_H(m) \leq$ r(m)2^{n+1}, r = poly. Follows from Lagrange interpolation.

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y}$ = {-1,1}, $\mathcal{H}$ = {h(x) = sign(x- a) with $a \in \mathbb{R}$} half-space classifiers.
Now $|\mathcal{H}| = \infty$ and we have $\Pi_H(m) = m + 1$. $\#\{h(x_i) = -1\}, \quad i \in \{1, \ldots, m\}$



Growth Function

Definition: The **growth function** $\Pi_H : \mathbb{N} \to \mathbb{N}$ for a hypothesis class $\mathcal{H}$ is defined as

$$\Pi_H(m) = \max_{\{x_1, x_2, \ldots, xm\} \subseteq X} \left| \left\{ \left(h(x_1), h(x_2), \ldots, h(x_m) \right) : h \in \mathcal{H} \right\} \right|$$

- **Counts maximum number** of distinct m-vectors $\left(h(x_1), h(x_2), \ldots, h(x_m)\right)$ that can be generated by the hypothesis class $\mathcal{H}$.

- **Upper bound** on the number of distinct ways m points can be classified by $\mathcal{H}$.

Theorem (Massert's Lemma): The **Rademacher complexity** is bounded by the **growth function** as

$$\mathcal{R}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\left(\Pi_{\mathcal{H}}(m)\right)}{m}}$$

Theorem (bound on generalization error for 0-1 loss): For any $\delta > 0$ we have with probability $1 - \delta$ that the following bounds hold uniformly for any $\mathrm{h} \in \mathcal{H}$,

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2\log\left(\Pi_{\mathcal{H}}(m)\right)}{m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

Note: Bound is now distribution D independent depending only on combinatorial features of $\mathcal{H}$.

VC-Dimension


Vladimir Vapnik

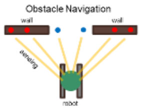

Alexey Chervonenkis

Definition: For a sample set S = $(x_1,x_2,\ldots,x_m)$ of size m, a **dichotomy** is one of the possible ways to label the set $(y_1,y_2,\ldots,y_m)$.

Definition: A set S of size m is said to be **shattered** by the hypothesis class $\mathcal{H}$ if for each dichotomy $\boldsymbol{y}$ there is an $h \in \mathcal{H}$ so that $(h(x_1) = y_1, h(x_2) = y_2,\ldots, h(x_m) = y_m)$.

Example: $\mathcal{X} = \{-2,-1,1\}$, $\mathcal{Y} = \{-1,1\}$, H = $\{h_1(x) = \text{sign}(x), h_2(x) = \text{sign}(x - 1.5)\}$.
h_1: -1,-1,1; h_2: -1,-1,-1. Now for $x_1 = -2$ with dichotomy $y_1 = 1$ can not be obtained from either h_1 or h_2 this hypothesis class fails to shatter even $\mathcal{X}^m$ for m = 1.

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{-1,1\}$, $\mathcal{H} = \{\text{h} : \text{h}(\text{x}) = \text{sign}(x - a)\cdot\text{sign}(b - x)$ for some $\text{a}, b \in \mathbb{R}\}$ the set of intervals [a,b].



Now for m = 2 for any two points $\text{x}1, \text{x}2 \in \mathbb{R}$ we have $\mathcal{H}$ ***shatters*** $\mathcal{X}^2$ by taking [a,b] to contain points with y_i = 1 and exclude any point with y_i = -1.

However, for $\text{m} \geq 3$ we *can not match* all **dichotomies**. Take for example $x_1 < x_2 < x_3$ with the labels $\text{y}_1 = +1, \text{y}_2 = -1, \text{y}_3 = +1$ then there is no interval containing both x_1 and x_3 but excluding x_2. Therefore, there exists dichotomies when m = 3 that no $h \in \mathcal{H}$ can classify correctly.

VC-Dimension

Definition: The **Vapnik-Chervonenkis dimension** is defined as

$$VCdim(\mathcal{H}) = \max\{m : \Pi_H(m) = 2^m\}$$


Vladimir Vapnik

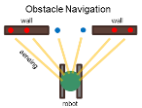

Alexey Chervonenkis

- The **VC-dimension** measures the size of the largest set that can be **shattered** by the hypothesis class $\mathcal{H}$.

- When $VCdim(\mathcal{H}) = d$ this means all sets of size *d* can be fully **shattered** by $\mathcal{H}$.

- For finite $|\mathcal{H}| < \infty$ hypothesis space we have $VCdim(\mathcal{H}) \leq \log(|\mathcal{H}|)$.

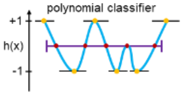
Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{-1,1\}$, $\mathcal{H} = \{\mathrm{h} : \mathrm{h}(\mathrm{x}) = \mathrm{sign}(x - a) \cdot \mathrm{sign}(b - x)$ for some $\mathrm{a}, b \in \mathbb{R}\}$ the set of intervals [a,b].



For m = 2 for any two points $\mathrm{x}_1, \mathrm{x}_2 \in \mathbb{R}$ we have $\mathcal{H}$ ***shatters*** $\mathcal{X}^2$ by taking [a,b] to contain points with y_i = 1 and exclude any point with y_i = -1.

However, for $\mathrm{m} \geq 3$ we *can not match* all **dichotomies**. Take for example $x_1 < x_2 < x_3$ with the labels $\mathrm{y}_1 = +1, \mathrm{y}_2 = -1, \mathrm{y}_3 = +1$ then there is no interval containing both x_1 and x_3 but excluding x_2. **Therefore,** $VCdim(\mathcal{H}) = 2$.

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{-1,1\}$, $\mathcal{H} = \{\mathrm{h} : \mathrm{h}(\mathrm{x}) = \mathrm{sign}(\mathrm{p}(\mathrm{x}))$ polynomial $\mathrm{p}(\mathrm{x})$ of degree n}. We have $\mathcal{H}$ ***shatters*** $\mathcal{X}^m$ for $m = n + 1$. This follows from Lagrange interpolation. However, can not shatter for m > n + 1, so $\mathrm{d} = VCdim(\mathcal{H}) = n + 1$.



VC-Dimension


Vladimir Vapnik


Alexey Chervonenkis

Definition: The **Vapnik-Chervonenkis dimension** is defined as

$$VCdim(\mathcal{H}) = \max\{m : \Pi_H(m) = 2^m\}$$

- The VC-dimension measures the size of the largest set that can be **shattered** by the hypothesis class $\mathcal{H}$.

- When $VCdim(\mathcal{H}) = d$ this means all sets of size *d* can be fully **shattered** by $\mathcal{H}$.

Theorem (bound on generalization error for 0-1 loss): When $VCdim(\mathcal{H}) = d$, for any $\delta > 0$ we have with probability $1 - \delta$ that the following bounds hold uniformly for any $\mathrm{h} \in \mathcal{H}$,

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

- Note the ratio of m/d governs the bound. This corresponds to the overall form

$$\text{"} R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right) \text{"}$$

- This shows we **need sample size** m >> d to obtain **small bound**. Provides useful complexity when $|\mathcal{H}| = \infty$.

VC Dimension

Example: $VCdim(\mathcal{H})$ axis-aligned rectangles.

Claim: $VCdim(\mathcal{H}) = 4$.

Two steps:

(i) lower bound $VCdim(\mathcal{H}) \geq 4$
(ii) upper bound $VCdim(\mathcal{H}) < 5$

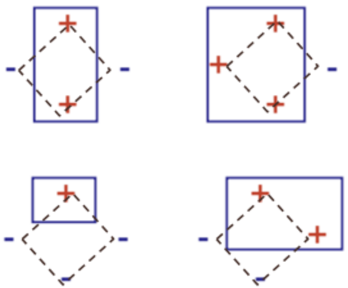
Lower bound: Place 4 points into a diamond configuration.
All cases can clearly be handled.

Upper bound: Place 5 points with 4 on rectangle and the 5th point in the interior.
No axis-aligned rectangle that can correctly classify these points for all labels.
Hence, $VCdim(\mathcal{H}) < 5$.

Characterizes the complexity of the infinite dimensional hypothesis space $\mathcal{H}$.

VC-dimension bounds provide a sampling complexity for learning the axis-aligned rectangle.

Cases for 4 points



Mori 2012

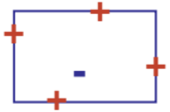
Building Identification

Google Maps: UCSB South Hall

Picture Annotation, Facial Recognition

usplash

Case for 5 points



VC-Dimension: Hyperplanes

Example: Learning separating hyperplane in $\mathbb{R}^N$ (related to SVM). For data $\{(x_i, y_i)\}$ with $x_i \in \mathbb{R}^N$ and $y_i \in \{-1,1\}$. Ideally, find $\mathbf{w}$, b so that $\mathrm{sign}(\boldsymbol{w}^T\boldsymbol{x}_i + b) = y_i$.

Hypothesis class:

$\mathcal{H} = \{\mathrm{h}: \mathrm{h}(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + \mathrm{b})$ with $\boldsymbol{w} \in \mathbb{R}^N$, $\mathrm{b} \in \mathbb{R}\}$.

What is the $\boldsymbol{VCdim}(\mathcal{H})$?

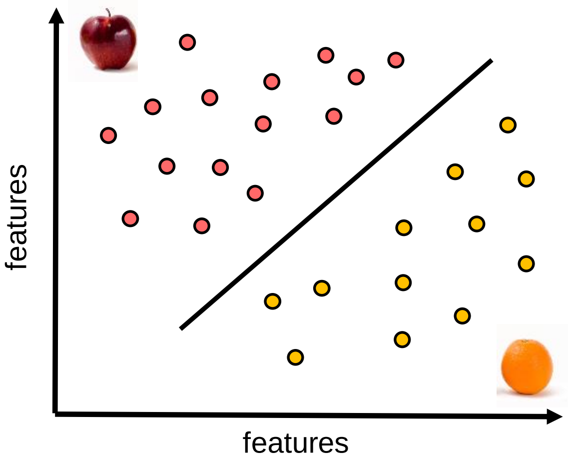
Claim: $VCdim(\mathcal{H}) = N + 1$

Two steps:

(i) lower bound $VCdim(\mathcal{H}) \geq N + 1$.
(ii) upper bound $VCdim(\mathcal{H}) < N + 2$.

Image Database

Linear Classifier



Lower bound: For N + 1 points, let $\mathbf{x}_0 = (0,0,\ldots,0)$ origin, $\mathbf{x}_i = (0, \ldots 1, \ldots 0, 0) = \mathbf{e_i}$, with i^{th} component one. For any labels $y_i \in \{-1,1\}$, let $\mathbf{w} = (y_1, y_2, \ldots, y_N)$ and $b = \frac{y0}{2}$ which defines the classifier $\mathrm{h}(\mathbf{x_i}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x_i} + \mathrm{b}) = \mathrm{sign}\left(y_i + \frac{y0}{2}\right) = y_i$. This verifies any N + 1 labels can be classified correctly.

Hyperplanes $\mathcal{H}$ **shatters** this N + 1 point-set so $VCdim(\mathrm{H}) \geq \mathrm{N}+1$.

VC-Dimension: Hyperplanes

Example: Learning separating hyperplane in $\mathbb{R}^N$ (related to SVM). For data $\{(x_i,y_i)\}$ with $x_i \in \mathbb{R}^N$ and $y_i \in \{-1,1\}$. Ideally, find $\mathbf{w}$, b so that $\mathrm{sign}(\boldsymbol{w}^T\boldsymbol{x}_i + b) = y_i$.

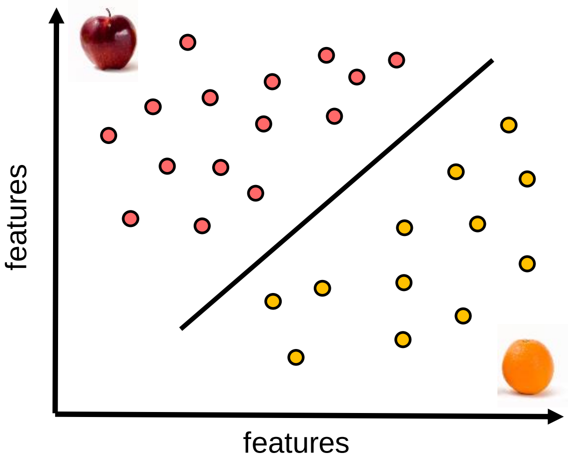
Upper bound: $VCdim(\mathcal{H}) <$ N+2. Must show for any N + 2 points for some labels there is no hyperplane classifier.

Theorem (Radon): In $\mathbb{R}^N$ a set of N + 2 points always can be partitioned into two disjoint subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ that have intersecting convex hulls $C(\mathcal{X}_1) \bigcap C(\mathcal{X}_2) \neq \emptyset$.

Implication: Let labels of $\mathcal{X}_1$ be say +1 and $\mathcal{X}_2$ be -1 then there is no separating hyperplane (it would separate the convex hulls).

Image Database

Linear Classifier

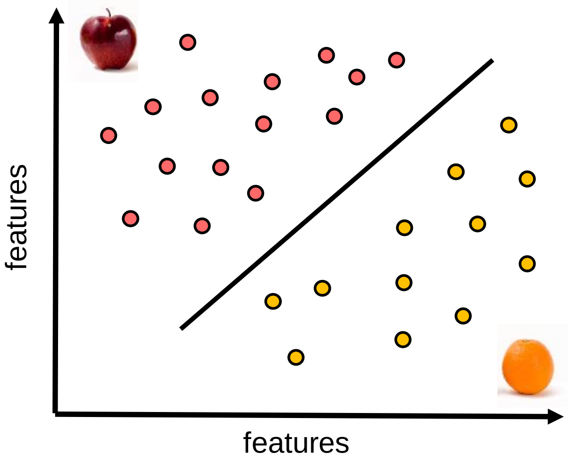


Proof (Radon): Consider the set of N + 1 linear equations in N + 2 unknowns: $\sum_{i=1}^{N+2} \alpha_i \mathbf{x}_i = \mathbf{0}$ and $\sum_{i=1}^{N+2} \alpha_i = 0$
Non-trivial null-space so equations have non-zero solution $\beta_1, \ldots, \beta_{d+2}$ with $\sum_{i=1}^{N+2} \beta_i = 0$. Let $I_1 = \{\mathrm{i} : \beta_\mathrm{i} > 0\}$

and $I_2 = \{\mathrm{i} : \beta_\mathrm{i} \leq 0\}$, then both non-empty. Let $\mathbf{x}^* = \sum_{i_1 \in I_1} \frac{\beta_{i_1}}{\beta} \mathbf{x}_{i_1} = \sum_{i_2 \in I_2} \frac{-\beta_{i_2}}{\beta} \mathbf{x}_{i_2}$ with $\beta = \sum_{i_1 \in I_1} \beta_{i_1}$.

We have $\sum_{i_1 \in I_1} \frac{\beta_{i_1}}{\beta} = \sum_{i_2 \in I_2} \frac{-\beta_{i_2}}{\beta} = 1$ and $\frac{\beta_{i1}}{\beta} \geq 0$, $\frac{-\beta_{i2}}{\beta} \geq 0$, so $x^* \in C(\mathcal{X}_1) \bigcap C(\mathcal{X}_2) \neq \emptyset$ so convex hulls intersect. ■

VC-Dimension: Hyperplanes



Example: Learning separating hyperplane in $\mathbb{R}^N$ (related to SVM). For data $\{(x_i,y_i)\}$ with $x_i \in \mathbb{R}^N$ and $y_i \in \{-1,1\}$. Ideally, find **w**, b so that $\mathrm{sign}(\boldsymbol{w}^T \boldsymbol{x}_i + b) = y_i$.

Hypothesis class:

$\mathcal{H} = \{h : h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$ with $\boldsymbol{w} \in \mathbb{R}^N$, $b \in \mathbb{R}\}$.

What is the $VCdim(\mathcal{H})$?

Claim: $VCdim(\mathcal{H})$ = N + 1

Shows in separable case that we have bound on generalization error

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

Turns out we can do even better in bounding sampling complexity for SVM. Want independent of feature dimension N, for bounded features (future lectures).

Will discuss further these results later when we cover **Support Vector Machines.**

VC Dimension: Lower Bounds

Lower Bounds: Given assumptions of PAC-Learning and $VCdim(\mathcal{H})$.
What is lower bound on generalization error given m samples?

Theorem: Under assumptions of PAC for $\mathrm{d} = VCdim(\mathcal{H}) > 1$ given any learning algorithm $\mathcal{A}$ there always exists a distribution D and concept $f \in \mathcal{C}$ so that for m samples

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

Shows that at least 1% of the time you will always have generalization error bigger than $\frac{d-1}{32m}$.

Characterizes the worse-case generalization errors given complexity of $\mathcal{H}$.

Consequence: If $VCdim(\mathcal{H})$ = ∞ then task is **not PAC-Learnable.**

Building Identification

Google Maps: UCSB South Hall

Picture Annotation, Facial Recognition

usplash

VC Dimension: Lower Bounds

Example: Consider hypothesis class of all polynomials
$\mathcal{H}$ = {h: h(x) = sign(p(x)) any polynomial of finite degree}.

Complexity: $VCdim(\mathcal{H}) = \infty$ (recall for n degree polynomial VCdim = n+1).

Consequence: **Concepts from $\mathcal{H}$** are **not** PAC-Learnable.

Why? At least 1% of the time you will always have generalization error bigger than $\frac{d-1}{32m}$ so make $d = \lceil 31.7m + 1 \rceil$ (since $VCdim(\mathcal{H}) = \infty$ can take any d > 1) then we have

$$\Pr_{S\sim D^m}\left[R_D(h_S, f) > \frac{d-1}{32m}\right] \geq 1/100 \longrightarrow \Pr_{S\sim D^m}\{R_D(h_S, f) > 0.99\} \geq 1/100$$

Shows no matter how many samples m used, 1% of the time the generalization error is greater than 99%.

Not enough information from finite data alone to distinguish unknown function in $\mathcal{H}$ without further assumptions (i.e. could miss local variations). Need other approaches (i.e. regularization, level of smoothness).

Consequence, if $VCdim(\mathcal{H})$ **=** ∞ then task is **not PAC-Learnable.**

Complexity: Rademacher, Growth Functions, VC-Dimension

Complexity Bounds Theory and Practice

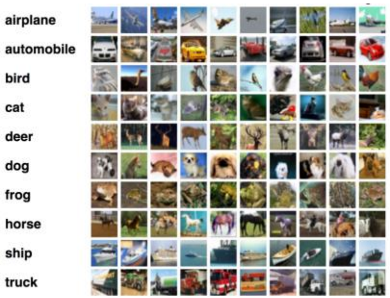
Significance: Complexity measures give some guarantees to assess generalization errors based on observed empirical risk.

In practice, often challenging since models have large complexity and we want to avoid overfitting data by only minimizing empirical risk. Training methods often also have further regularizations or stochasticity (Kernel-SVM, N-Nets, SGD, Dropout).

Many extensions to the introduced ideas here but PAC + complexity bounds provide a good starting point for theory and intuition.

How can we use complexity measures in practice to inform our design of learning algorithms, training methods, and assess expected performance?

Image Classification

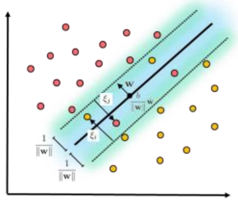


Abdellatif Abdelfattah

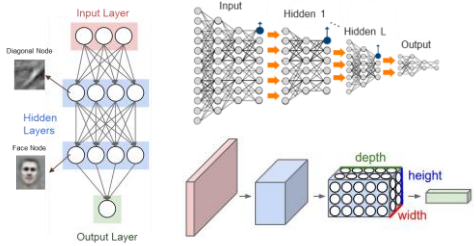
Robotics and Control

MIT and Boston Dynamics

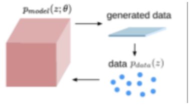
Support Vector Machines



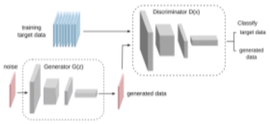
Neural Networks and Deep Learning



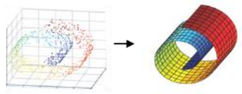
Generative Methods



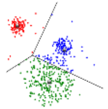
GANs



Manifold Learning



Clustering Methods



Forecasting

washingtonpost.com