Machine Learning and Data Analysis
Lecture 9: Infinite Hypothesis Spaces

Filip Železný

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Intelligent Data Analysis lab
http://ida.felk.cvut.cz

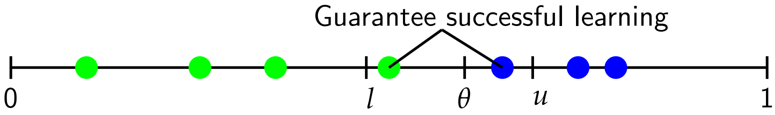
January 10, 2011

PAC Learning Summary

Concept class (efficiently) PAC learnable by a hypothesis class if

- a *consistent* hypothesis can be (efficiently) produced for each sample
- size of hypothesis space at most exponential

Two weeks ago we proved PAC-learnability of threshold hypotheses on $[0;1]$



Here PAC-learnability does not follow from the above principle since there are ∞ threshold hypotheses. Can we extend the above principle to cover infinite hypothesis classes?

An Intuitive Approach

Assume θ has finite precision, say 64 bits. In a digital machine, this is the case anyway.

For threshold hypotheses on $[0, 1]$:

$$\ln |\mathcal{F}| = \ln |2^{64}| = 64 \ln 2$$

For threshold hypotheses

$$f(x) = 1 \text{ iff } \theta_1 x^{(1)} + \theta_2 x^{(2)} > 0$$

on $[0, 1]^2$:

$$\ln |\mathcal{F}| = \ln |2^{2 \cdot 64}| = 128 \ln 2$$

Generally for hypothesis classes with n parameters

$$\ln |\mathcal{F}| = \ln |2^{64n}| = 64n \ln 2 = \mathcal{O}(n)$$

An Intuitive Approach (cont'd)

$\ln |\mathcal{F}|$ linear in number of hypothesis-class parameters and precision of real-number representation

Approach seems viable, allows PAC-learning

Problem:

$$\begin{array}{lll} \mathcal{F}_1\text{:} & f(x) = 1 \text{ iff } \theta_1 x^{(1)} + \theta_2 x^{(2)} > 0 & \text{2 parameters} \\ \mathcal{F}_2\text{:} & f(x) = 1 \text{ iff } |\theta_1 - \theta_2| x^{(1)} + |\theta_3 - \theta_4| x^{(2)} > 0 & \text{4 parameters} \end{array}$$

Different number of parameters but $\mathcal{F}_1 = \mathcal{F}_2$!

Instead of the number of parameters and precision, we will build a different characterization of infinite hypothesis classes.

$\Pi_{\mathcal{F}}$ function

A finite sample from P_X will be called an x-*sample*.

- $x_1, x_2, \ldots$ instead of $(x_1, y_1), (x_2, y_2), \ldots$

Remind the set-notation we earlier introduced for hypotheses:

- $x \in f$ means the same as $f(x) = 1$

$\Pi_{\mathcal{F}}$ function

For any X and $\mathcal{F}$ and a finite x-sample S define

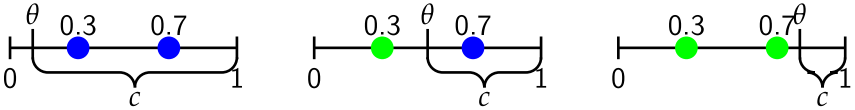
$$\Pi_{\mathcal{F}}(S) = \{f \cap S \mid f \in \mathcal{F}\}$$

We call $f \cap S$ a *labelling* on S. $\Pi_{\mathcal{F}}(S)$ gives all labellings of S possible with hypotheses from $\mathcal{F}$

$\Pi_{\mathcal{F}}$ function: Example

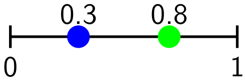
Let $\mathcal{F}$ be threshold hypotheses on $[0, 1]$ and $S = \{0.3, 0.7\}$

$$\Pi_{\mathcal{F}}(S) = \{\{0.3, 0.7\}, \{0.7\}, \{\}\}$$



but

$$\{0.3\} \notin \Pi_{\mathcal{F}}(S)$$



Shattering

Shattering

If $|\Pi_{\mathcal{F}}(S)| = 2^{|S|}$ then S is *shattered* by $\mathcal{F}$.

S is shattered by $\mathcal{F}$ if for *any* subset $S' \subseteq S$ there is a hypothesis $f \in \mathcal{F}$ such that $f \cap S = S'$.

Example: let $\mathcal{F}$ be threshold hypotheses on $[0, 1]$

- $\{0.3\}$ and $\{0.7\}$ are shattered by $\mathcal{F}$
- $\{0.3, 0.7\}$ is not shattered by $\mathcal{F}$

VC Dimension

VC Dimension

The *Vapnik-Chervonenkis* dimension of $\mathcal{F}$, denoted $\mathcal{V}(\mathcal{F})$, is the largest d such that some x-sample of cardinality d is shattered by $\mathcal{F}$. If no such d exists, then $\mathcal{V}(\mathcal{F}) = \infty$.

Example: let $\mathcal{F}$ be threshold hypotheses on $[0,1]$

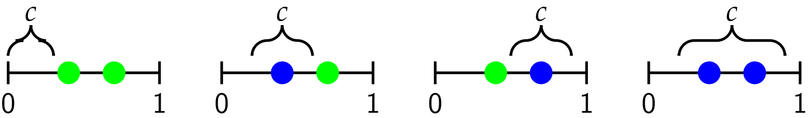
- $\{0.3\}$ is shattered by $\mathcal{F}$
- No x-sample S of cardinality 2 is shattered by $\mathcal{F}$ because $\{\min S\} \subseteq S$, but $S \cap f = \{\min S\}$ for no $f \in \mathcal{F}$.
- Since no x-sample of cardinality 2 is shattered, no x-sample of cardinality > 2 is shattered
- Therefore $\mathcal{V}(\mathcal{F}) = 1$.

VC Dimension: Examples

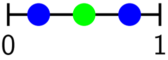
Let $\mathcal{F}$ be intervals $[a, b]$, $0 < a, b < 1$

- $\{0.3, 0.7\}$ is shattered by $\mathcal{F}$
- No x-sample of cardinality 3 or higher is shattered by $\mathcal{F}$ because $\{\min S, \max S\} \subseteq S$ but $S \cap f = \{\min S, \max S\}$ for no $f \in \mathcal{F}$.
- Therefore $\mathcal{V}(\mathcal{F}) = 2$.

Two points shattered



No three points can be shattered, the middle one can never be left out



VC Dimension: Examples

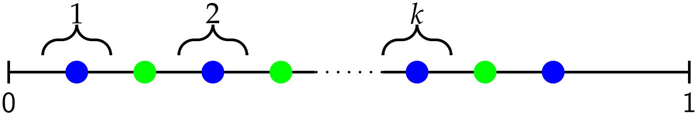
Let $\mathcal{F}$ be unions of k disjoint intervals $[a, b]$

- An x-sample of $2k$ elements shattered by $\mathcal{F}$
- No x-sample of cardinality $2k+1$ or higher is shattered by $\mathcal{F}$. Let $S = \{x_1, x_2, \ldots, x_{2k+1}\}$ such that $x_i < x_j$ for $i < j$. Then for

$$S' = \{x_1, x_3, \ldots x_{2k+1}\}$$

  $S' \subseteq S$ but $S' = S \cap c$ for no $f \in \mathcal{F}$.
- Therefore $\mathcal{V}(\mathcal{F}) = 2k$.

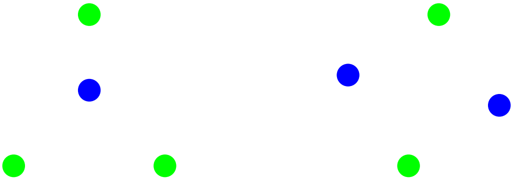
No $2k+1$ points can be shattered



VC Dimension: Examples

Let $\mathcal{F}$ be half-planes in R^2

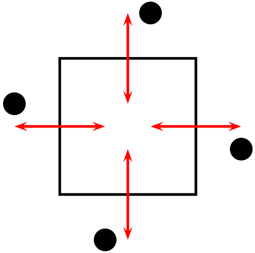
- Some 3 points can be shattered (obvious)
- No 4 points can be shattered. Clear if three of them in line. If not, then two cases possible, and impossible labelling exists in each:



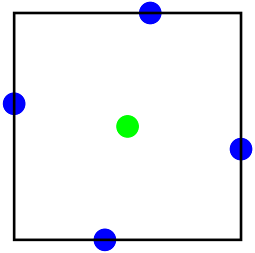
- $\mathcal{V}(\mathcal{F}) = 3$
- similarly shown: $\mathcal{V}(\text{circles in } R^2) = 3$
- Generally, $\mathcal{V}(\text{half-planes in } R^n) = n + 1$

VC Dimension: Examples

Let $\mathcal{F}$ be rectangles in R^2



Some four points can be shattered



Five can never be shattered

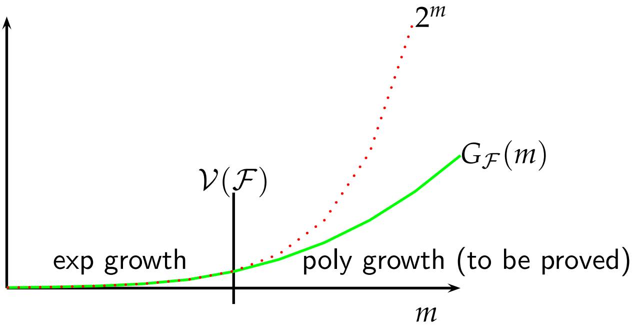
- $\mathcal{V}(\mathcal{F}) = 4$
- More generally, $\mathcal{V}(\text{convex tetragons}) = 9$
- More generally, $\mathcal{V}(\text{convex } d\text{-gons}) = 2d + 1$

Function $G_{\mathcal{F}}$

Function $G_{\mathcal{F}}$

$$G_{\mathcal{F}}(m) = \max\{|\Pi_{\mathcal{F}}(S)| \;:\; |S| = m\}$$

For a given m, $G_{\mathcal{F}}(m)$ returns the maximum number of ways an x-sample of size m can be labeled by hypotheses from $\mathcal{F}$.



Function $\Phi(k,m)$

Define:

$$\Phi(k,m) = \sum_{i=0}^{k} \binom{m}{i} = \begin{cases} 1 \text{ if } k = 0 \text{ or } m = 0 \\ \Phi(k,m-1) + \Phi(k-1,m-1) \end{cases}$$

The second equality may be shown by induction ('Pascal's triangle').

For $m > k$, it holds $0 \leq k/m < 1$ and

$$\left(\frac{k}{m}\right)^k \sum_{i=0}^{k} \binom{m}{i} \leq \sum_{i=0}^{k} \left(\frac{k}{m}\right)^i \binom{m}{i}$$

$$\leq \sum_{i=0}^{m} \left(\frac{k}{m}\right)^i \binom{m}{i} = \left(1 + \frac{k}{m}\right)^m \leq e^k$$

Dividing by $\left(\frac{k}{m}\right)^k$, we get that $\Phi(k,m)$ grows polynomially in m

$$\Phi(k,m) \leq e^k \left(\frac{m}{k}\right)^k \leq \left(\frac{me}{k}\right)^k$$

Bounding $G_{\mathcal{F}}(m)$ by $\Phi\left(\mathcal{V}(\mathcal{F}), m\right)$

We prove the polynomial bound

$$G_{\mathcal{F}}(m) \leq \Phi\left(\mathcal{V}(\mathcal{F}), m\right)$$

by induction on m and $\mathcal{V}(\mathcal{F})$.

Base case:

- if $m = 0$ then
  $$G_{\mathcal{F}}(0) = 1 = \Phi(\mathcal{V}(\mathcal{F}), 0)$$
  since there is only one subset of $\{\}$.
- if $\mathcal{V}(\mathcal{F}) = 0$ then
  $$G_{\mathcal{F}}(m) = 1 = \Phi(0, m)$$
  since if only $\{\}$ can be shattered then all points in any x-sample must be labeled the same by any $f \in \mathcal{F}$.

Bounding $G_{\mathcal{F}}(m)$ by $\Phi(\mathcal{V}(\mathcal{F}), m)$ (cont'd)

Induction step (assume an arbitrary S with m elements):

$$|\Pi_{\mathcal{F}}(S)| = |\Pi_{\mathcal{F}}(S \setminus \{x\})| + |\Delta\mathcal{S}|$$

where by definition of the G function (slide 13) and then by the induction assumption

$$|\Pi_{\mathcal{F}}(S \setminus \{x\})| \leq G_{\mathcal{F}}(m-1) \leq \Phi(\mathcal{V}(\mathcal{F}), m-1) \tag{1}$$

What about the difference term $|\Delta\mathcal{S}|$?

- For all $s \in \Pi_{\mathcal{F}}(S \setminus \{x\})$, there is 1 corresponding labelling
  1. $s \in \Pi_{\mathcal{F}}(S)$
- For some $s \in \Pi_{\mathcal{F}}(S \setminus \{x\})$, there are 2 corresponding labellings
  1. $s \in \Pi_{\mathcal{F}}(S)$
  2. $s \cup \{x\} \in \Pi_{\mathcal{F}}(S)$

Thus $\Delta\mathcal{S}$ should include exactly the $s \in \Pi_{\mathcal{F}}(S \setminus \{x\})$ that have 2 corresponding labellings in $\Pi_{\mathcal{F}}(S)$.

Bounding $G_{\mathcal{F}}(m)$ by $\Phi(\mathcal{V}(\mathcal{F}), m)$ (cont'd)

Therefore:

$$\Delta\mathcal{S} = \{s \in \Pi_{\mathcal{F}}(S) \mid x \notin s,\ s \cup \{x\} \in \Pi_{\mathcal{F}}(S)\}$$

Note that

$$\Delta\mathcal{S} = \Pi_{\Delta\mathcal{S}}(S \setminus \{x\})$$

($\Delta\mathcal{S}$ in the subscript acts as a hypothesis class, which is OK!)

Illustrative example with $\mathcal{F} = \{f \mid f(x) = 1 \text{ iff } x < \theta, \theta \in [0,1]\}$:

- $S = \{0.1, 0.2, 0.3\}$, $x = 0.3$
- $\Pi_{\mathcal{F}}(S) = \{\{\}, \{0.1\}, \{0.1, 0.2\}, \{0.1, 0.2, 0.3\}\}$
- $\Pi_{\mathcal{F}}(S \setminus \{x\}) = \{\{\}, \{0.1\}, \{0.1, 0.2\}\}$
- $\Delta\mathcal{S} = \{\{0.1, 0.2\}\}$
- $\Pi_{\Delta\mathcal{S}}(S \setminus \{x\}) = \Pi_{\{\{0.1,0.2\}\}}(\{0.1, 0.2\}) = \{0.1, 0.2\} = \Delta\mathcal{S}$

Bounding $G_{\mathcal{F}}(m)$ by $\Phi\left(\mathcal{V}(\mathcal{F}), m\right)$ (cont'd)

What about $\mathcal{V}(\Delta\mathcal{S})$?

1. Remind definition: $\Delta\mathcal{S} = \{s \in \Pi_{\mathcal{F}}(S) \mid x \notin s,\ s \cup \{x\} \in \Pi_{\mathcal{F}}(S)\}$
2. $\Delta\mathcal{S} \subseteq \Pi_{\mathcal{F}}(S)$ (from 1).
3. Let T be a sample shattered by $\Delta\mathcal{S}$.
4. $x \notin T$ (from 3 and 1)
5. $|T \cup \{x\}| = |T| + 1$ (from 4)
6. For all $t \subseteq T$, $t \in \Delta\mathcal{S}$ (from 3)
7. For all $t \subseteq T$, $t \in \Pi_{\mathcal{F}}(S)$ (from 6 and 2)
8. For all $t \subseteq T$, $t \cup \{x\} \in \Pi_{\mathcal{F}}(S)$ (from 6 and 1)
9. $\mathcal{F}$ shatters $T \cup \{x\}$ (from 3,7, and 8)
10. $\mathcal{V}(\mathcal{F}) \geq \mathcal{V}(\Delta\mathcal{S}) + 1$ (from 3 and 9)

Bounding $G_{\mathcal{F}}(m)$ by $\Phi\left(\mathcal{V}(\mathcal{F}), m\right)$ (cont'd)

Remind that

$$\Delta\mathcal{S} = \Pi_{\Delta\mathcal{S}}(S \setminus \{x\})$$

by definition of the G function (slide 13)

$$|\Pi_{\Delta\mathcal{S}}(S \setminus \{x\})| \leq G_{\Delta\mathcal{S}}(m-1)$$

we proved that

$$\mathcal{V}(\Delta\mathcal{S}) \leq \mathcal{V}(\mathcal{F}) - 1$$

by induction assumption

$$G_{\Delta\mathcal{S}}(m-1) \leq \Phi(\mathcal{V}(\mathcal{F}) - 1, m-1)$$

so

$$|\Delta\mathcal{S}| = |\Pi_{\Delta\mathcal{S}}(S \setminus \{x\})| \leq \Phi(\mathcal{V}(\mathcal{F}) - 1, m-1) \tag{2}$$

Bounding $G_{\mathcal{F}}(m)$ by $\Phi(\mathcal{V}(\mathcal{F}), m)$ (cont'd)

Returning to the induction step:

$$|\Pi_{\mathcal{F}}(S)| = |\Pi_{\mathcal{F}}(S \setminus x)| + |\Delta \mathcal{S}|$$

We have proved (Eq. 1 and Eq. 2):

$$\begin{aligned} |\Pi_{\mathcal{F}}(S \setminus \{x\})| &\leq \Phi(\mathcal{V}(\mathcal{F}), m-1) \\ |\Delta \mathcal{S}| &\leq \Phi(\mathcal{V}(\mathcal{F}) - 1, m-1) \end{aligned}$$

Using the above and the definition of Φ (slide 14) we have

$$|\Pi_{\mathcal{F}}(S)| \leq \Phi(\mathcal{V}(\mathcal{F}), m-1) + \Phi(\mathcal{V}(\mathcal{F}) - 1, m-1) = \Phi(\mathcal{V}(\mathcal{F}), m)$$

Since S was arbitrary, we proved the polynomial bound for $G_{\mathcal{F}}(m)$:

$$G_{\mathcal{F}}(m) \leq \Phi(\mathcal{V}(\mathcal{F}), m)$$

Error regions

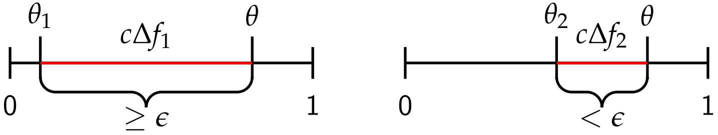
Denote $c\Delta f = \{x \in X \mid c(x) \neq f(x)\}$ and define for $c \in \mathcal{C}$, $\epsilon \in R$:

$$\Delta_\epsilon(c) = \{c\Delta f \mid f \in \mathcal{F}, \sum_{x \in c\Delta f} P_X(x) \geq \epsilon\}$$

Notes:

- Replace $\sum$ by $\int$ for continuous X
- $\Delta_\epsilon(c)$ does not have $\mathcal{F}$ in the subscript but it depends on it!

Example for a treshold concept c (with threshold θ) and $\epsilon = 0.5$, with $\mathcal{F} = \{f_1, f_2\}$ (thresholds θ_1, θ_2), assuming uniform P_X:



$c\Delta f_1 \in \Delta_\epsilon(c)$ $\qquad$ $c\Delta f_2 \notin \Delta_\epsilon(c)$

Error regions

Note that for any $\mathcal{F}$, any $c \in \mathcal{C}$ and any x-sample S

$$\begin{aligned}\Pi_{\mathcal{F}}(S) =& \{f \cap S \mid f \in \mathcal{F}\} \\ \Pi_{\Delta_0(c)}(S) =& \{(c\Delta f) \cap S \mid f \in \mathcal{F}\}\end{aligned}$$

There is a bijective mapping

$$f \cap S \Leftrightarrow (c\Delta f) \cap S$$

between $\Pi_{\mathcal{F}}(S)$ and $\Pi_{\Delta_0(c)}(S)$. Thus

$$|\Pi_{\Delta_0(c)}(S)| = |\Pi_{\mathcal{F}}(S)|$$

and therefore

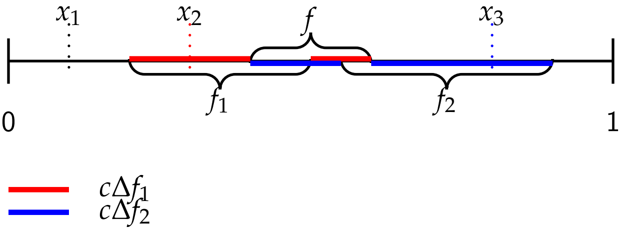
$$\mathcal{V}(\Delta_0(c)) = \mathcal{V}(\mathcal{F})$$

We will need this observation later. (Remind: $\mathcal{V}(\Delta_0(c))$ depends on $\mathcal{F}$!)

ϵ-net

For any $\epsilon \in R$, an x-sample S is an ϵ-net for a concept $c \in \mathcal{C}$ and hypothesis class $\mathcal{F}$ if every region $r \in \Delta_\epsilon(c)$ contains a point from S, i.e $r \cap S \neq \{\}$.

Example for interval hypotheses, with $\mathcal{F} = \{f_1, f_2\}$:



$\{x_1, x_2\}$ is not an ϵ-net.
$\{x_2, x_3\}$ is an ϵ-net.
$\{x_1, x_2, x_3\}$ is an ϵ-net.