OPPA European Social Fund
Prague & EU: We invest in your future.

Machine Learning and Data Analysis
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

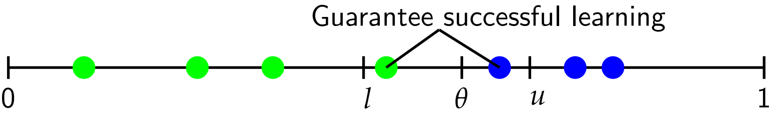
January 6, 2012

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:

$\mathcal{F}_1$:	$f(x) = 1$ iff $\theta_1 x^{(1)} + \theta_2 x^{(2)} > 0$	2 parameters
$\mathcal{F}_2$:	$f(x) = 1$ iff $\lvert\theta_1 - \theta_2\rvert x^{(1)} + \lvert\theta_3 - \theta_4\rvert x^{(2)} > 0$	4 parameters

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

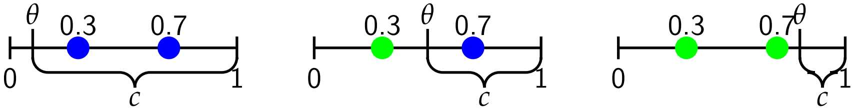
$$\Pi_{\mathcal{F}}(S) = \{f \cap S \mid f \in \mathcal{F}\}$$

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

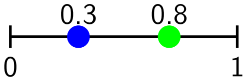
Let $\mathcal{F}$ be threshold hypotheses on $[0,1]$ and $S = \{0.3, 0.7\}$

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



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 $[0, 1]$

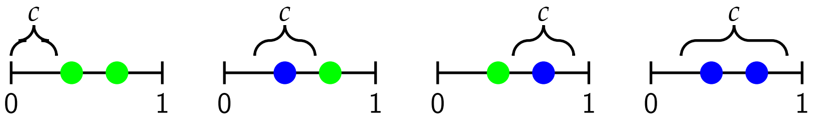
- $\{0.3\}$ is shattered by $\mathcal{F}$
- No x-sample S of cardinality 2 is shattered by $\mathcal{F}$ because $\{\min S\} \subseteq S$, but $S \cap f = \{\min S\}$ for no $f \in \mathcal{F}$.
- Since no x-sample of cardinality 2 is shattered, no x-sample of cardinality > 2 is shattered
- Therefore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

VC Dimension: Examples

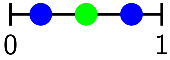
Let $\mathcal{F}$ be intervals $[a, b]$, $0 < a, b < 1$

- $\{0.3, 0.7\}$ 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

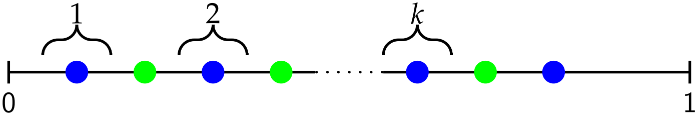
Let $\mathcal{F}$ be unions of k disjoint intervals $[a, b]$

- 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 $S = \{x_1, x_2, \dots, x_{2k+1}\}$ such that $x_i < x_j$ for $i < j$. Then for

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

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

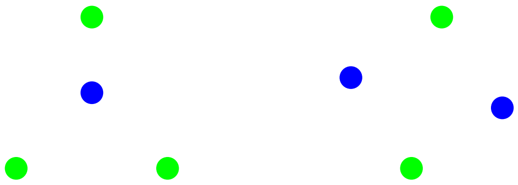
No $2k+1$ points can be shattered



VC Dimension: Examples

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

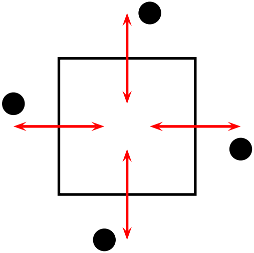
- Some 3 points can be shattered (obvious)
- No 4 points can be shattered. Clear if three of them in line. If not, then two cases possible, and impossible labelling exists in each:



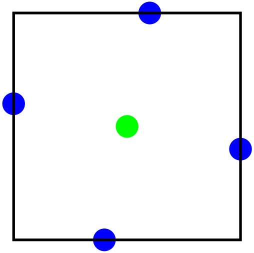
- $\mathcal{V}(\mathcal{F}) = 3$
- similarly shown: $\mathcal{V}(\text{circles in } R^2) = 3$
- Generally, $\mathcal{V}(\text{half-planes in } R^n) = n + 1$

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, $\mathcal{V}(\text{convex } d\text{-gons}) = 2d + 1$

PAC Learning with Infinite $\mathcal{F}$: Result

PAC Learning with Infinite $\mathcal{F}$

Let $\mathcal{F}$ be a hypothesis class with a finite $\mathcal{V}(\mathcal{F})$ and $\mathcal{C}$ be concept class, both on X. Let $c \in \mathcal{C}$ be a concept. A hypothesis f consistent with a sample $\{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$ will have $e(f) \leq \epsilon$ with probability at least $1 - \delta$ if

$$m \geq \max\left(\frac{8}{\epsilon}\log_2\frac{2}{\delta}\,,\ \frac{8\mathcal{V}(\mathcal{F})}{\epsilon}\log_2\frac{13}{\epsilon}\right)$$

Therefore any $\mathcal{C}$ is (efficiently) PAC-learnable by $\mathcal{F}$ if there is an (efficient) learner producing a consistent $f \in \mathcal{F}$ for any sample, and $\mathcal{V}(\mathcal{F})$ is polynomial (in the size of examples n).

As we have seen, $\mathcal{V}(\mathcal{F})$ is usually linear in the number of hypothesis class parameters, which corresponds to n.

$\mathcal{V}(\mathcal{F})$: Remarks

- The result can be rewritten into a simpler form

$$m \geq c_0 \left(\frac{\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{1}{\epsilon} + \frac{1}{\epsilon} \log_2 \frac{1}{\delta} \right)$$

  where c_0 is a constant.

- The result holds also for finite $\mathcal{F}$. For some $\mathcal{F}$, it may even provide better bounds than those we derived specially for finite $\mathcal{F}$.
- Finite $\mathcal{V}(\mathcal{F})$ is also a necessary condition for PAC-learning. It can be proved that at least

$$\frac{\mathcal{V}(\mathcal{F}) - 1}{64\epsilon}$$

  examples are needed to PAC-learn a concept class with $\mathcal{F}$ if $\delta \leq 1/15$.

Error Bounds for Infinite $\mathcal{F}$

$\mathcal{V}(\mathcal{F})$ also enables to derive error bounds for inconsistent hypotheses.
$\mathcal{V}(\mathcal{F})$ is 'analogical' to $\ln|\mathcal{F}|$ for finite hypothesis classes.

With probability at least $1-\delta$, for a training set S:

$$|e(f) - e(S,f)| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m}\log_2\frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m}\log_2\frac{1}{\delta}}\right)$$

and if f minimizes training error $e(f,S)$ then with probability at least $1-\delta$:

$$e(f) \leq e(f^*) + \mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m}\log_2\frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m}\log_2\frac{1}{\delta}}\right)$$

where f^* minimizes classification error $e(f)$.

Bias-Variance Trade-off Revisited

Remind: in the finite $\mathcal{F}$ case, by extending $\mathcal{F}$

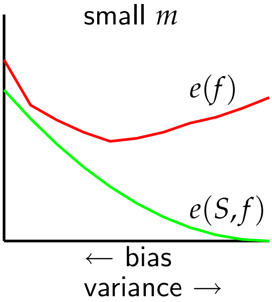
$$e(f) \leq \underbrace{\left(\min_{f \in \mathcal{F}} e(f)\right)}_{\text{'bias': may decrease}} + \underbrace{2\sqrt{\frac{1}{2m} \ln \frac{2|\mathcal{F}|}{\delta}}}_{\text{'variance': will increase}}$$

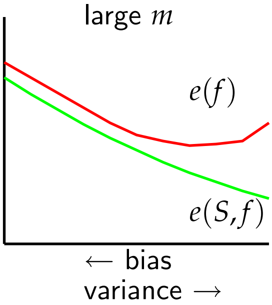
This holds analogically for infinite $\mathcal{F}$

$$e(f) \leq \underbrace{\left(\min_{f \in \mathcal{F}} e(f)\right)}_{\text{'bias': may decrease}} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m} \log_2 \frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m} \log_2 \frac{1}{\delta}}\right)}_{\text{'variance': will increase}}$$

Bias-Variance Trade-off Revisited (cont'd)

Resulting behavior (we have seen this before)





OPPA European Social Fund
Prague & EU: We invest in your future.