THE VAPNIK-CHERVONENKIS DIMENSION and LEARNABILITY

Dan A. Simovici

UMB,
Doctoral Summer School
Iasi, Romania

We are given a sequence of points on a two-dimensional grid such that:

- each blue point is inside an unknown shape;
- each red point is outside an unknown shape.

How many points we need until we can say with a “reasonable” degree of certainty what is the shape?

The Complexities of a Grid

We'll see a 45×33-grid containing 1485 points. With this set we can:

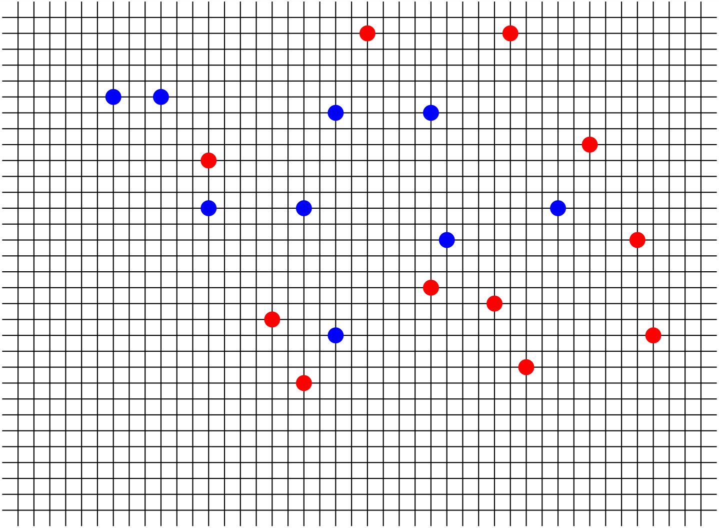
- draw 2^{1485} shapes, which is about

$$10^{500}$$

- define $2^{2^{1485}}$ families of shapes, which is about

$$10^{\frac{10^{500}}{3}}$$

- for comparison, the number of atoms in our Universe is about 10^{80}!



Why it is difficult to determine what is the “right” shape?

We are seeking to determine a concept starting from a series of examples.

- The concept class is too broad.
- We need to limit the class of concepts and to formulate a hypothesis that is consistent with the examples examined.
- Formalization must be introduced such that we know precisely what we are taking about.

Concepts, Positive and Negative Examples

Let $X \subseteq S^+$.

- X is the *example space*;
- A *concept* is a function $C : X \rightsquigarrow \{0, 1\}$ (identifiable with a subset of X);
- If $C(x) = 1$, then x is a *positive example*; if $C(x) = 0$, then c is a *negative example*.

- $POS(C) = \{x \in X \mid C(x) = 1\}$ is the set of *positive examples*;
- $NEG(C) = \{x \in X \mid C(x) = 0\}$ is the set of *negative examples*;
- $\mathrm{dom}(C) = POS(C) \cup NEG(C)$.

Concepts and Hypotheses

Two sets of concepts need to be considered:

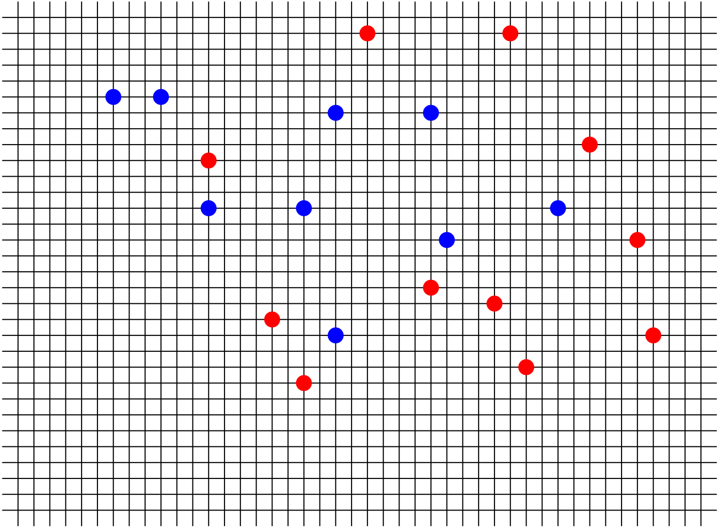
- concepts from real world (the *concept space* $\mathcal{C}$);
- concepts that an algorithm is capable of recognizing (the *hypothesis space* $\mathcal{H}$).

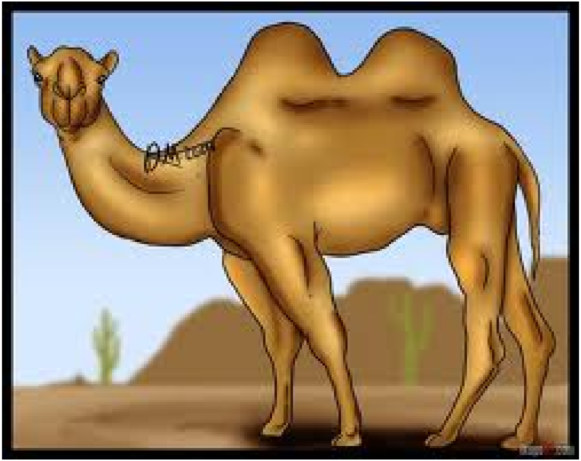
The central problem of ML: For each concept $C \in \mathcal{C}$ find a hypothesis $H \in \mathcal{H}$ which is an approximation of C.

A hypothesis H for a concept C is formed by feeding a sequence of examples (positive and negative) of C to a learning algorithm L.

Lessons to be drawn:

- use concept classes that can be identified in feasible time with a guaranteed level of certainty;
- define the concept class: in our case, two-dimensional drawings of animals.





The Trace of a Collection of Sets

Let U be a set, $K \subseteq U$, and $\mathcal{C}$ be a collection of subsets of U. The *trace* of $\mathcal{C}$ on K is the collection of sets

$$\mathcal{C}_K = \{C \cap K \mid C \in \mathcal{C}\}$$

Set Shattering and the VCD

Let U be a set, $K \subseteq U$, and $\mathcal{C}$ be a collection of subsets of U.
If $\mathcal{C}_K = \mathcal{P}(K)$, then we say that K is *shattered by $\mathcal{C}$*.

Definition

The *Vapnik-Chervonenkis dimension* of the collection $\mathcal{C}$ (called the VC-dimension for brevity) is the largest cardinality of a set K that is shattered by $\mathcal{C}$ and is denoted by $VCD(\mathcal{C})$.

Example

$U = \{u_1, u_2, u_3, u_4\}$ and

$$\mathcal{C} = \{\{u_2, u_3\}, \{u_1, u_3, u_4\}, \{u_2, u_4\}, \{u_1, u_2\}, \{u_2, u_3, u_4\}\}.$$

$K = \{u_1, u_3\}$ is shattered by the collection $\mathcal{C}$ because

$$\begin{aligned}
\{u_1, u_3\} \cap \{u_2, u_3\} &= \{u_3\} \\
\{u_1, u_3\} \cap \{u_1, u_3, u_4\} &= \{u_1, u_3\} \\
\{u_1, u_3\} \cap \{u_2, u_4\} &= \emptyset \\
\{u_1, u_3\} \cap \{u_1, u_2\} &= \{u_1\} \\
\{u_1, u_3\} \cap \{u_2, u_3, u_4\} &= \{u_3\}
\end{aligned}$$

The Tabular Form

$T_{\mathcal{C}}$

u_1	u_2	u_3	u_4
0	1	1	0
1	0	1	1
0	1	0	1
1	1	0	0
0	1	1	1

$K = \{u_1, u_3\}$ is shattered by the collection $\mathcal{C}$ because $\boldsymbol{r}[K] = ((0,1),(1,1),(0,0),(1,0),(0,1))$ contains the all four necessary tuples $(0,1)$, $(1,1)$, $(0,0)$, and $(1,0)$.
No subset K of U with $|K| \geqslant 3$ can be shattered by $\mathcal{C}$ because this would require $|\boldsymbol{r}[K]| \geqslant 8$. Thus, $VCD(\mathcal{C}) = 2$.

The Functional Form

Let $U = \{u_1, \ldots, u_n\}$.

- Each set $C \subseteq U$ can be identified with its signed characteristic function $f_C : U \longrightarrow \{-1, 1\}$, where

$$f_C(x) = \begin{cases} 1 & \text{if } x \in C, \\ -1 & \text{otherwise.} \end{cases}$$

  Thus, $\mathcal{C}$ can be regarded as a collection of function $\mathcal{F} \subseteq \{-1, 1\}^U$.
- K with $|K| = m$ is shattered by $\mathcal{F}$ if for every $(b_1, \ldots, b_m) \in \{-1, 1\}^m$ there exists $f \in \mathcal{F}$ such that

$$(f(u_1), \ldots, f(u_m)) = (b_1, \ldots, b_m).$$

- $VCD(\mathcal{F})$ is the cardinality of the largest subset of X that is shattered by $\mathcal{F}$.

Theorem

Let U be a finite nonempty set and let $\mathcal{C}$ be a collection of subsets of U. If $d = VCD(\mathcal{C})$, then $2^d \leqslant |\mathcal{C}| \leqslant (|U| + 1)^d$.

Vapnik-Chervonenkis classes

For a collection of sets $\mathcal{C}$ and for $m \in \mathbb{N}$, let $\mathcal{C}[m]$ be

$$\mathcal{C}[m] = \max\{|\mathcal{C}_K| \mid |K| = m\}.$$

This is the largest number of distinct subsets of a set having m elements that can be obtained as intersections of the set with members of $\mathcal{C}$. In general, $\mathcal{C}[m] \leqslant 2^m$; however, if $\mathcal{C}$ shatters a set of size m, then $\mathcal{C}[m] = 2^m$.

Definition

A *Vapnik-Chervonenkis class* (or a *VC class*) is a collection $\mathcal{C}$ of sets such that $VCD(\mathcal{C})$ is finite.

Example

Let $\mathcal{S}$ be the collection of sets $\{(-\infty, t) \mid t \in \mathbb{R}\}$. Any singleton is shattered by $\mathcal{S}$. Indeed, if $S = \{x\}$ is a singleton, then $\mathcal{P}(\{x\}) = \{\emptyset, \{x\}\}$. Thus, if $t \geqslant x$, we have $(-\infty, t) \cap S = \{x\}$; also, if $t < x$, we have $(-\infty, t) \cap S = \emptyset$, so $\mathcal{S}_S = \mathcal{P}(S)$.

There is no set S with $|S| = 2$ that can be shattered by $\mathcal{S}$.

Suppose that $S = \{x, y\}$, where $x < y$. Then, any member of $\mathcal{S}$ that contains y includes the entire set S, so $\mathcal{S}_S = \{\emptyset, \{x\}, \{x, y\}\} \neq \mathcal{P}(S)$. This shows that $\mathcal{S}$ is a VC class and $VCD(\mathcal{S}) = 1$.

For $\mathcal{I} = \{[a, b] \mid a, b \in \mathbb{R}, a \leqslant b\}$ we have $VCD(\mathcal{I}) = 2$.
No three-element set can be shattered by $\mathcal{I}$.
Consider the intersections

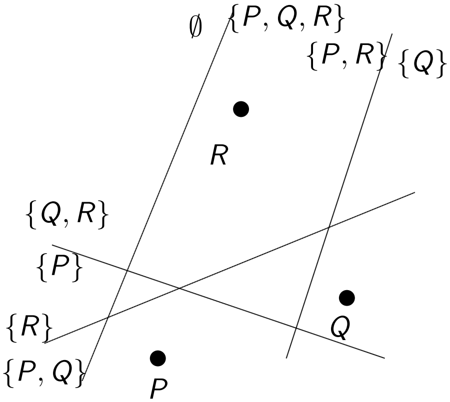
$$\begin{aligned}
&[u, v] \cap S = \emptyset, \text{ where } v < x,\\
&[x - \epsilon, \tfrac{x+y}{2}] \cap S = \{x\},\\
&[\tfrac{x+y}{2}, y] \cap S = \{y\},\\
&[x - \epsilon, y + \epsilon] \cap S = \{x, y\},
\end{aligned}$$

which show that $\mathcal{I}_S = \mathcal{P}(S)$.
No three-element set can be shattered by $\mathcal{I}$. Let $T = \{x, y, z\}$. Any interval that contains x and z also contains y, so it is impossible to obtain the set $\{x, z\}$ as an intersection between an interval in $\mathcal{I}$ and the set T.

Example

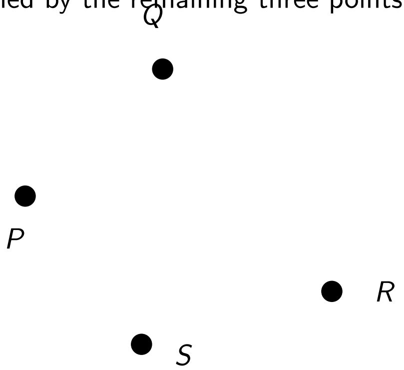
Let $\mathcal{H}$ be the collection of closed half-planes in $\mathbb{R}^2$. We claim that $VCD(\mathcal{H}) = 3$.

Let P, Q, R be three points in $\mathbb{R}^2$ such that they are not located on the same line. Each line is marked with the sets it defines; thus, the family of hyperplanes shatters the set $\{P, Q, R\}$, so $VCD(\mathcal{H})$ is at least 3.

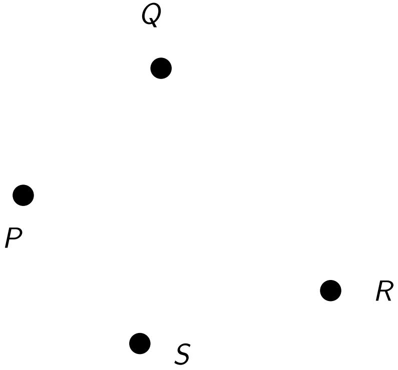


Example (cont'd)

No set that contains at least four points can be shattered by $\mathcal{H}$.
Let $\{P, Q, R, S\}$ be a set in general position.
If S is located inside the triangle P, Q, R, then every half-plane that contains P, Q, R will contain S, so it is impossible to separate the subset $\{P, Q, R\}$. Thus, we may assume that no point is inside the triangle formed by the remaining three points.



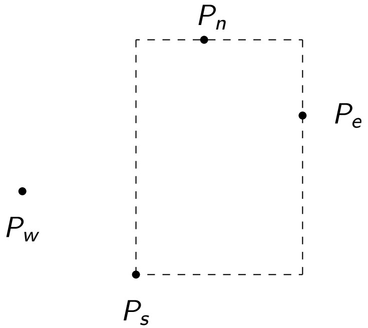
Example (cont'd)



Any half-plane that contains two diagonally opposite points, for example, P and R, will contain either Q or S, which shows that it is impossible to separate the set $\{P, R\}$. Thus, no set that contains four points may be shattered by $\mathcal{H}$, so $VCD(\mathcal{H}) = 3$.

Example

Let $\mathcal{R}$ be the set of rectangles whose sides are parallel with the axes x and y. Each such rectangle has the form $[x_0, x_1] \times [y_0, y_1]$.
There is a set S with $|S| = 4$ that is shattered by $\mathcal{R}$. Indeed, let S be a set of four points in $\mathbb{R}^2$ that contains a unique "northernmost point" P_n, a unique "southernmost point" P_s, a unique "easternmost point" P_e, and a unique "westernmost point" P_w. If $L \subseteq S$ and $L \neq \emptyset$, let R_L be the smallest rectangle that contains L.



Example (cont'd)

This collection cannot shatter a set of points that contains at least five points.
Indeed, let S be a set of points such that $|S| \geqslant 5$ and, as before, let P_n be the northernmost point, etc. If the set contains more than one "northernmost" point, then we select exactly one to be P_n. Then, the rectangle that contains the set $K = \{P_n, P_e, P_s, P_w\}$ contains the entire set S, which shows the impossibility of separating the set K.

Recapitulation

X	$\mathcal{C}$	$VCD(\mathcal{C})$
$\mathbb{R}^2$	convex polygons	∞
$\mathbb{R}^2$	axis-aligned rectangles	4
$\mathbb{R}^2$	convex polygons with d vertices	$2d+1$
$\mathbb{R}^d$	closed half-spaces	$d+1$
$\mathbb{R}^N$	neural networks with	
	N parameters	$O(N \log N)$

- If $\mathcal{C}$ is not a VC class, then $\mathcal{C}[m] = 2^m$ for all $m \in \mathbb{N}$.
- If $VCD(\mathcal{C}) = d$, then $\mathcal{C}[m]$ is bounded asymptotically by a polynomial of degree d.

The number of subsets having at most d elements of a subset having m elements is:

$$\binom{n}{\leqslant k} = \sum_{i=0}^{k} \binom{n}{k}.$$

Theorem

Let $\phi : \mathbb{N}^2 \longrightarrow \mathbb{N}$ be the function defined by

$$\phi(d,m) = \begin{cases} 1 & \text{if } m = 0 \text{ or } d = 0 \\ \phi(d, m-1) + \phi(d-1, m-1) & \text{otherwise.} \end{cases}$$

We have $\phi(d,m) = \binom{m}{\leqslant d}$ for $d, m \in \mathbb{N}$.

Proof by strong induction on $s = i + m$

The base case: $s = 0$ implies $m = 0$ and $d = 0$; the equality is immediate.
Inductive case: suppose that the equality holds for $\phi(d', m')$, where $d' + m' < d + m$. We have

$$
\begin{aligned}
\phi(d,m) &= \phi(d,m-1) + \phi(d-1,m-1) \\
&\quad \text{(by definition)} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&\quad \text{(by inductive hypothesis)} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
&\quad \left(\text{since } \binom{m-1}{-1} = 0\right) \\
&= \sum_{i=0}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \\
&= \sum_{i=0}^{d} \binom{m}{i} = \binom{m}{\leqslant d},
\end{aligned}
$$

Proof by strong induction on $s = d + m$

The base case: for $s = 0$, $d = m = 0$ so $\mathcal{C}$ shatters only the empty set. Thus, $\mathcal{C}[0] = |\mathcal{C}_{\emptyset}| = 1$, and therefore $\mathcal{C}[0] = 1 = \phi(0, 0)$.
Inductive case: suppose that the statement holds for pairs (d', m') such that $d' + m' < s$ and let $\mathcal{C}$ be a collection of subsets of S such that $VCD(\mathcal{C}) = d$.
Let $|K| = m$ and let $k_0 \in K$ be a fixed (but, otherwise, arbitrary) element of K. Consider the trace $\mathcal{C}_{K-\{k_0\}}$. Since $|K - \{k_0\}| = m - 1$, we have, by the inductive hypothesis, $|\mathcal{C}_{K-\{k_0\}}| \leqslant \phi(d, m - 1)$.

Proof by strong induction on $s = d + m$ (cont'd)

Let $\mathcal{C}'$ be the collection of sets given by

$$\mathcal{C}' = \{G \in \mathcal{C}_K \mid k_0 \notin G, G \cup \{k_0\} \in \mathcal{C}_K\}.$$

Observe that $\mathcal{C}' = \mathcal{C}'_{K-\{k_0\}}$ because $\mathcal{C}'$ consists only of subsets of $K - \{k_0\}$. Further, note that the Vapnik-Chervonenkis dimension of $\mathcal{C}'$ is less than d. Indeed, let K' be a subset of $K - \{k_0\}$ that is shattered by $\mathcal{C}'$. Then, $K' \cup \{k_0\}$ is shattered by $\mathcal{C}$. hence $|K'| < d$. By the inductive hypothesis, $|\mathcal{C}'| = |\mathcal{C}_{K-\{k_0\}}| \leq \phi(d-1, m-1)$.

Proof by strong induction on $s = d + m$ (cont'd)

The $\mathcal{C}_K$ can be regarded as the union of two disjoint collections:

- those subsets in $\mathcal{C}_K$ that do not contain the element k_0 ($\mathcal{C}_{K-\{k_0\}}$)
- those subsets of K that contain k_0.

If L is a second type of subset, then $L - \{k_0\}$ is clearly a member of $\mathcal{C}'$.
Thus, we have

$$|\mathcal{C}_K| = |\mathcal{C}_{K-\{k_0\}}| + |\mathcal{C}'_{K-\{k_0\}}|$$

This equality implies

$$|\mathcal{C}_K| \leqslant \phi(d, m-1) + \phi(d-1, m-1),$$

the desired conclusion.

Sauer-Shelah Theorem

Theorem

If $\mathcal{C}$ is a collection of subsets of S that is a VC-class such that $VCD(\mathcal{C}) = d$, then $\mathcal{C}[m] \leqslant \phi(d, m)$ for $m \in \mathbb{N}$.

Proof by strong induction on $s = d + m$

The base case: for $s = 0$, $d = m = 0$ so $\mathcal{C}$ shatters only the empty set. Thus, $\mathcal{C}[0] = |\mathcal{C}_\emptyset| = 1$, and therefore $\mathcal{C}[0] = 1 = \phi(0, 0)$.
Inductive case: suppose that the statement holds for pairs (d', m') such that $d' + m' < s$ and let $\mathcal{C}$ be a collection of subsets of S such that $VCD(\mathcal{C}) = d$.
Let $|K| = m$ and let $k_0 \in K$ be a fixed (but, otherwise, arbitrary) element of K. Consider the trace $\mathcal{C}_{K-\{k_0\}}$. Since $|K - \{k_0\}| = m - 1$, we have, by the inductive hypothesis, $|\mathcal{C}_{K-\{k_0\}}| \leqslant \phi(d, m - 1)$.

Proof by strong induction on $s = d + m$ (cont'd)

Let $\mathcal{C}'$ be the collection of sets given by

$$\mathcal{C}' = \{G \in \mathcal{C}_K \mid k_0 \notin G, G \cup \{k_0\} \in \mathcal{C}_K\}.$$

Observe that $\mathcal{C}' = \mathcal{C}'_{K-\{k_0\}}$ because $\mathcal{C}'$ consists only of subsets of $K - \{k_0\}$. Further, note that the Vapnik-Chervonenkis dimension of $\mathcal{C}'$ is less than d. Indeed, let K' be a subset of $K - \{k_0\}$ that is shattered by $\mathcal{C}'$. Then, $K' \cup \{k_0\}$ is shattered by $\mathcal{C}$. hence $|K'| < d$. By the inductive hypothesis, $|\mathcal{C}'| = |\mathcal{C}_{K-\{k_0\}}| \leq \phi(d-1, m-1)$.

Proof by strong induction on $s = d + m$ (cont'd)

The $\mathcal{C}_K$ can be regarded as the union of two disjoint collections:

- those subsets in $\mathcal{C}_K$ that do not contain the element k_0 ($\mathcal{C}_{K-\{k_0\}}$)
- those subsets of K that contain k_0.

If L is a second type of subset, then $L - \{k_0\}$ is clearly a member of $\mathcal{C}'$. Thus, we have

$$|\mathcal{C}_K| = |\mathcal{C}_{K-\{k_0\}}| + |\mathcal{C}'_{K-\{k_0\}}|$$

This equality implies

$$|\mathcal{C}_K| \leqslant \phi(d, m-1) + \phi(d-1, m-1),$$

the desired conclusion.

Lemma

Lemma

We have $\phi(d, m) \leqslant \frac{2m^d}{d!}$

Proof: If $d = 1$, this amounts to $\phi(1, m) = m + 1 \leqslant 2m$, which is obvious. Thus, we assume that $d > 1$.

For $m = d$ we prove that $\phi(d, d) = 2^d \leqslant \frac{2d^d}{d!}$, by induction on d.

The base case: for $d = 1$ the inequality is immediate.

The inductive case: Suppose that $2^d \leqslant \frac{2d^d}{d!}$. We have

$$\begin{aligned} 2^{d+1} = 2 \cdot 2^d &\leqslant \left(\frac{d+1}{d}\right)^d \cdot 2^d \\ &\quad \left(\text{by the well-known inequality } 2 < \left(\frac{d+1}{d}\right)^d\right) \\ &\leqslant \left(\frac{d+1}{d}\right)^d \cdot \frac{2d^d}{d!} \\ &= 2\frac{(d+1)^{d+1}}{(d+1)!}, \end{aligned}$$

which concludes the induction.

Proof (cont'd)

For a given d the argument is by induction on m, where $m \geqslant d$.
The base case: we presented the argument for $m = d$.
The inductive case: Since $\phi(d+1, m+1) = \phi(d+1, m) + \phi(d, m)$, it suffices to show that

$$2\frac{m^d}{d!} + 2\frac{m^{d+1}}{(d+1)!} \leqslant 2\frac{(m+1)^{d+1}}{(d+1)!}.$$

By multiplying both sides by $\frac{1}{2}\frac{d!}{m^d}$ we have the equivalent and immediate inequality

$$1 + \frac{m}{d+1} \leqslant \left(1 + \frac{1}{m}\right)^{d+1},$$

which concludes the proof.

A Second Lemma

Lemma

For $d \geqslant 1$ we have $2\left(\frac{d}{e}\right)^d < d!$.

Proof: The argument is by induction on d.

The base case: for $d = 1$ the proof is immediate.

The inductive case: suppose that the inequality holds for d. Then, for $d + 1$ we have

$$
\begin{aligned}
2\left(\frac{d+1}{e}\right)^{d+1} &= 2\left(\frac{d+1}{d}\right)^{d} \frac{d+1}{d} \frac{d^{d+1}}{e^{d+1}} \\
&\leqslant 2e\frac{d+1}{d} \frac{d^{d+1}}{e^{d+1}} \\
&\quad \left(\text{because } \left(\frac{d+1}{d}\right)^{d} \leqslant e\right) \\
&= 2(d+1)\frac{d^d}{e^d} \leqslant (d+1)! \\
&\quad \text{(by inductive hypothesis).}
\end{aligned}
$$

An Inequality Involving ϕ

Theorem

For all $m \geqslant d \geqslant 1$ we have

$$\phi(d, m) < \left(\frac{em}{d}\right)^d .$$

Proof: the theorem follows by combining the previous two lemmas.

A Corollary of Sauer-Shelah Theorem

Corollary

If $\mathcal{C}$ is a collection of subsets of S that is a VC-class such that $VCD(\mathcal{C}) = d$, then $\mathcal{C}[m] \leqslant \left(\frac{em}{d}\right)^d$ for $m \geq d \geqslant 1$.

Sequences on Sets

- $\mathbf{Seq}_n(S)$ is the set of all sequences of length n on S, also denoted by S^n;
- $\mathbf{Seq}(S) = \bigcup_{n \in \mathbb{N}} \mathbf{Seq}_n(S)$ is the set of all sequences on S, also denoted by S^*;
- $\mathbf{Seq}_0(S)$ consists of the null sequence λ;
- $S^+ = S^* - \{\lambda\}$ is the set of non-null sequences on S.

Samples

Definition

A *sample of length m*, where $m \geqslant 1$, is an m-tuple $\boldsymbol{s} = ((x_1, b_1), \ldots, (x_m, b_m)) \in \mathbf{Seq}_m\,(X \times \{0, 1\})$ that satisfies the coherence condition: $x_i = x_j$ implies $b_i = b_j$ for $1 \leqslant i, j \leqslant m$.

- $\boldsymbol{s} = ((x_1, b_1) \ldots, (x_m, b_m))$ is a *training sample* for a target concept T if $b_i = T(x_i)$ for $1 \leqslant i \leqslant m$.
- A hypothesis H is *consistent with* $\boldsymbol{s}$ if $H(x_i) = b_i$ for $1 \leqslant i \leqslant m$.

Rays in $\mathbb{R}$

Definition

A *θ-ray* is a set

$$Y_\theta = \{x \in \mathbb{R} \mid x \geqslant \theta\}.$$

Special case:

$$Y_\infty = \emptyset.$$

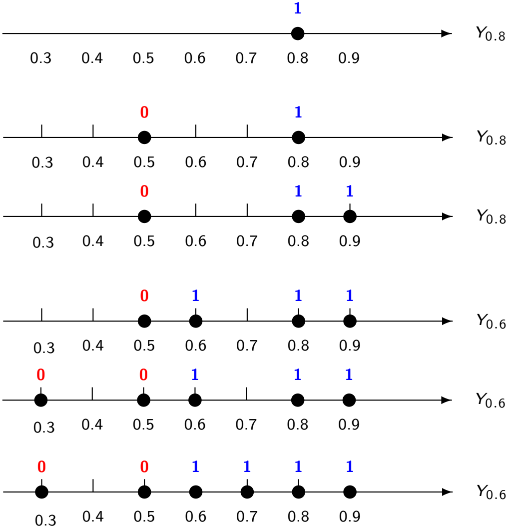
The *space of rays* is $\mathcal{H}_{rays} = \{Y_\theta \mid \theta \in \mathbb{R}\}$.

L - An Algorithm for Learning Rays

Input: a training sample $\boldsymbol{s} = ((x_1, b_1) \ldots, (x_m, b_m))$, where $x_i \in \mathbb{R}$ and $b_i \in \{0, 1\}$ for $1 \leqslant i \leqslant m$.
Output: a hypothesis in H_{rays}.
Algorithm:

$\lambda = \infty$;
`for` $i = 1$ `to` m `do`
  if $b_i = 1$ and $x_i < \lambda$ then $\lambda = x_i$;
$L(\boldsymbol{s}) = Y_\lambda$;

Sequence of hypotheses



Probably Approximative Correct Learning

Main features of the model:

- a training sample $\boldsymbol{s}$ of length m for a target concept C is generated by drawing from the probability space $\mathfrak{X} = (X, \mathcal{E}, P)$ according to some fixed probability distribution;
- a learning algorithm L produces a hypothesis $L(\boldsymbol{s})$ intended to approximate t;
- as m increases the expectation is that the error of using $L(\boldsymbol{s})$ instead of C decreases.

The Error of a Hypothesis

Assumptions:

- the target concept C belongs to a hypothesis space $\mathcal{H}$ available to the learner;
- the error of a hypothesis H with respect to C is

$$err_P(H, C) = P(\{x \in X \mid H(x) \neq C(x)\})$$

- $\{x \in X \mid H(x) \neq C(x)\} \in \mathcal{E}$.

Probabilistic Framework

Given $\mathfrak{X} = (X, \mathcal{E}, P)$, consider the product probability space $\mathfrak{X}^m = (X^m, \mathcal{E}^m, P^{(m)})$.

- the components of a sample $\boldsymbol{s} = (x_1, \ldots, x_m)$ are regarded as m independent random variables, identically distributed;
- $S(m, C)$: the set of training samples of size m for a target concept C;
- the probability on the product space $P^{(m)}$ will be still denoted by P.

Probably Approximately Correct Algorithms

L. Valiant:

- δ: a confidence parameter;
- ϵ: accuracy parameter;

An algorithm L is *probably approximately correct* (PAC) if given $\delta \in (0,1)$ and $\epsilon \in (0,1)$, there is a positive integer $m_0 = m_0(\delta, \epsilon)$ such that for any target concept $C \in \mathcal{H}$ and for any probability P on X, $m \geqslant m_0$ implies

$$P(\{\boldsymbol{s} \in S(m,t) \mid err_P(L(\boldsymbol{s}), C) < \epsilon\}) > 1 - \delta.$$

Essential Feature: m_0 depends only on δ and ϵ.

Learning Rays is PAC

- target concept Y_θ, δ, ϵ and P:
- $\boldsymbol{s}$: a training sample of length m;
- error set: $L(\boldsymbol{s}) = y_\lambda$ and $[\theta, \lambda)$;
- define $\beta_0 = \sup\{\beta \mid P([\theta, \beta)) < \epsilon\}$.

Note that

$$\begin{aligned} P([\theta, \beta_0)) &\leqslant \epsilon \\ P([\theta, \beta_0]) &\geqslant \epsilon \end{aligned}$$

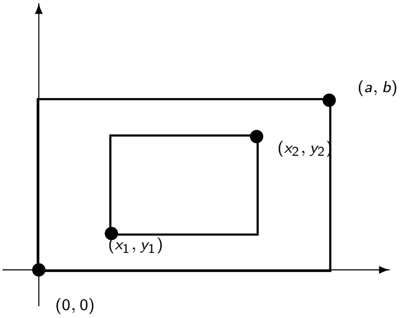
If $\lambda \leqslant \beta_0$ then

$$err_P(L(\boldsymbol{s}), Y_\theta) = err_P(Y_\lambda, Y_\theta) = P([\theta, \lambda)) \leqslant P([\theta, \beta_0)) \leqslant \epsilon.$$

The Hypothesis Space of Axis-Aligned Rectangles

Axis-aligned close rectangles (referred to as rectangles) are specified by the coordinates of their southwestern and northeastern corners; (x_1, y_1) and (x_2, y_2). Such a rectangle is denoted by $[x_1, y_1; x_2, y_2]$, The PAC-learnability was analyzed by Kearns and Vazirani.

$$[x_1, y_1; x_2, y_2] = \{(x, y) \in \mathbb{R}^2 \mid x \in [x_1, x_2], y \in [y_1, y_2]\}.$$



A Learning Strategy

- Construct an axis-aligned rectangle that gives the tightest fit to the positive rectangles.
- This strategy will yield a hypothesis R such that $R \subseteq R_0$.
- If no positive example exist, $R = \emptyset$.
- Error is $P(R - R_0)$.

Other Possible Learning Strategies:

- constructing the largest rectangle that excludes all negative examples, or
- constructing a rectangle located at mid-distance between the positive and the negative examples.

L - An Algorithm for Learning Axis-Aligned Rectangles

Input: a training sample $\boldsymbol{s} = ((x_1, y_1), b_1) \ldots, (x_m, y_m), b_m))$.
Output: a hypothesis R in R.
Algorithm:

```
R = [u_1, v_1; u_1, v_1];
i = 1;
for i = 1 to m do
   if (b_i = 1)
      R = R ⋆ (x_i, y_i);
   endif;
   i = i + 1
endfor
L(s) = R;
```

The Hypothesis Space

- $\mathcal{H}$: hypothesis space defined on an example space X;
- L: learning algorithm for $\mathcal{H}$; L is consistent if for any training sample $\boldsymbol{s}$ for a target concept $T \in \mathcal{H}$, the output hypothesis H of L agrees with T on examples in $\boldsymbol{s}$, that is, $H(x_i) = T(x_i)$ for $1 \leqslant i \leqslant |s|$.
- $S(m, T)$: set of samples of length m for the target concept T.
- $\mathcal{H}[\boldsymbol{s}]$: set of hypothesis consistent with $\boldsymbol{s}$:

$$\mathcal{H}[\boldsymbol{s}] = \{H \in \mathcal{H} \mid H(x_i) = T(x_i) \text{ for } 1 \leqslant i \leqslant m\}.$$

Definitions

- P: a probability distribution on X; T is a target concept;
- define $err(H, T) = P\{x \in X \mid H(x) \neq T(x)\}$;
- for $\boldsymbol{s} = ((x_1, b_1), \ldots, (x_m, b_m))$ define

$$err_{\boldsymbol{s}}(H, T) = \frac{1}{m} \cdot |\{i \mid b_i = T(x_i) \neq H(x_i)\}|$$

- the set of ϵ-bad hypotheses for T is

$$\text{BAD}_\epsilon(T) = \{H \in \mathcal{H} \mid err(H, T) \geqslant \epsilon\}.$$

A consistent learning algorithm L for $\mathcal{H}$ when presented with a sample $\boldsymbol{s}$ produces a hypothesis $H \in \mathcal{H}$ that is consistent with $\boldsymbol{s}$, that is a hypothesis in $\mathcal{H}[s]$.
The PAC property requires that such an output is unlikely to be ϵ-bad.

Potential Learnability of a Hypothesis Space

Definition

A hypothesis space $\mathcal{H}$ is *potentially learnable* if given $\delta, \epsilon \in (0, 1)$ there is a positive integer $m_0 = m_0(\delta, \epsilon)$ such that, $m \geqslant m_0$ implies

$$P(\boldsymbol{s} \in S(m, T) \mid \mathcal{H}[\boldsymbol{s}] \cap \mathrm{BAD}_{\epsilon}(T) = \emptyset) > 1 - \delta$$

for any probability distribution P and any target T.

Theorem

If $\mathcal{H}$ is potentially learnable and L is a consistent learning algorithm, then L is PAC.

Proof: If L is consistent, then $L(\boldsymbol{s}) \in \mathcal{H}[\boldsymbol{s}]$. Thus, the condition $\mathcal{H}[\boldsymbol{s}] \cap \text{BAD}_\epsilon(T) = \emptyset$ means that $err(T, L(\boldsymbol{s})) < \epsilon$.

Potential Learnability of Finite Hypotheses Space

Theorem

If $\mathcal{H}$ is finite, then it is potentially learnable.

Proof

Suppose that $\mathcal{H}$ is finite and ϵ, δ, T and P are given.
Claim: $P(\mathcal{H}[\boldsymbol{s}] \cap \mathrm{BAD}_\epsilon(T) \neq \emptyset)$ can be made sufficiently small if m, the size of $\boldsymbol{s}$ is large enough.

- $P\{x \in X \mid H(x) = T(x)\} = 1 - err(T, H) \leqslant 1 - \epsilon$.
- Probability that any one ϵ-bad hypothesis is in $\mathcal{H}[\boldsymbol{s}]$: $P^m\{\boldsymbol{s} \mid H(x_i) = T(x_i) \text{ for } 1 \leqslant i \leqslant m\} \leqslant (1-\epsilon)^m$.
- Probability that there is some ϵ-bad hypothesis in $\mathcal{H}[\boldsymbol{s}]$: $P^m(\boldsymbol{s} \mid \mathcal{H}[\boldsymbol{s}] \cap \mathrm{BAD}_\epsilon(T) \neq \emptyset) \leqslant |\mathcal{H}|(1-\epsilon)^m$.
- If

$$m \geqslant m_0 = \left\lceil \frac{1}{\epsilon} \ln \frac{|\mathcal{H}|}{\delta} \right\rceil,$$

then

$$|\mathcal{H}|(1-\epsilon)^m \leqslant |\mathcal{H}|(1-\epsilon)^{m_0} < |\mathcal{H}|e^{-\epsilon m_0} \leqslant |\mathcal{H}|e^{\ln(\delta/|\mathcal{H}|)} = \delta.$$

Theorem

If a hypothesis space $\mathcal{H}$ has infinite VCD, then $\mathcal{H}$ is not potentially learnable.

Proof

Suppose that $VCD(\mathcal{H}) = \infty$. There exists a sample $\boldsymbol{z}$ of length $2m$ which is shattered by $\mathcal{H}$.

- $E_{\boldsymbol{z}}$: set of examples of $\boldsymbol{z}$.
- P a probability on X such that

$$P(x) = \begin{cases} \frac{1}{2m} & \text{if } x \in E_{\boldsymbol{z}}, \\ 0 & \text{otherwise.} \end{cases}$$

  With probability 1 a random sample $\boldsymbol{x}$ of length m is a sample of examples from $E_{\boldsymbol{z}}$.
- Since $\boldsymbol{z}$ is shattered by $\mathcal{H}$, there exists $H \in \mathcal{H}$ such that $H(x_i) = T(x_i)$ for $1 \leqslant i \leqslant m$ and $H(x_i) \neq T(x_i)$ for $m+1 \leqslant i \leqslant 2m$. Thus, $err(H, T) \geqslant \frac{1}{2}$. Thus, any positive m and any target concept T, there is P such that $P(\boldsymbol{s} \in S(m, T) \mid \mathcal{H}[\boldsymbol{s}] \cap \mathrm{BAD}_\epsilon(T) = \emptyset) = 0$.

There is no positive integer $m_0 = m_0(0.5, 0.5)$ such that $m > m_0$ such that $P(\boldsymbol{s} \in S(m, T) \mid \mathcal{H}[\boldsymbol{s}] \cap \mathrm{BAD}_{0.5}(T) = \emptyset) > 0.5$.

An Example of a Hypothesis Space of Infinite VDC

Let $\mathcal{U}$ be the collection of finite union of closed intervals of $\mathbb{R}$.
Let $\boldsymbol{z}$ be a sample and let $E_{\boldsymbol{z}}$ be the set of example in $\boldsymbol{z}$. If $A \subseteq E_{\boldsymbol{z}}$, define U_A to the union of closed intervals, such that each interval contains exactly one element of A. Then, $U_A \cap E_{\boldsymbol{z}} = A$, so $\mathcal{U}$ shatters A.

Lemma

Lemma

For $c > 0$ and $x > 0$ we have

$$\ln x \leqslant \left(\ln \frac{1}{c} - 1\right) + cx.$$

Proof: Let $f : \mathbb{R}_{>0} \longrightarrow \mathbb{R}$ be the function $f(x) = \ln \frac{1}{c} - 1 + cx - \ln x$. Note that $\lim_{x \to 0} f(x) = +\infty$ and $\lim_{x \to \infty} f(x) = +\infty$. Also,

$$f'(x) = c - \frac{1}{x} \text{ and } f''(x) = \frac{1}{x^2},$$

so f has a minimum for $x = \frac{1}{c}$. Since $f\left(\frac{1}{c}\right) = 0$, the inequality follows.

Let T be a target concept, $T \in \mathcal{C}$. The class of *error regions with respect to T* is

$$\Delta(\mathcal{C}, T) = \{C \oplus T \mid C \in \mathcal{C}\}.$$

Also, for $\epsilon \geqslant 0$, let

$$\Delta_{\epsilon}(\mathcal{C}, T) = \{E \in \Delta(\mathcal{C}, T) \mid P(E) \geqslant \epsilon\}.$$

Theorem

$VDC(\Delta(\mathcal{C}, T)) = VDC(\mathcal{C})$ for any $T \in \mathcal{C}$.

Proof: Let K be a fixed concept and let $\phi : \mathcal{C}_K \longrightarrow \Delta(\mathcal{C}, T)_K$ be

$$\phi(C \cap T) = (C \oplus T) \cap K$$

for $C \in \mathcal{C}$. ϕ is a bijection, for if

$$(C_1 \oplus T) \cap K = (C_2 \oplus T) \cap K,$$

we have $C_1 \cap K_1 = C_2 \cap K_2$.

ϵ-Nets

Definition

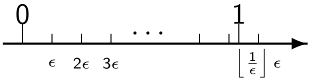
A set S is an ϵ-net for $(\mathcal{C}, T)$ if for every $R \in \Delta_\epsilon(\mathcal{C}, T)$ we have $S \cap R \neq \emptyset$. (S hits R)

- Equivalently: S is an ϵ-net for $\Delta(\mathcal{C}, T)$ if $R \in \Delta(\mathcal{C}, T)$ and $P(R) \geqslant \epsilon$ imply $S \cap R \neq \emptyset$.
- S fails to be an ϵ-net for $(\mathcal{C}, T)$ if there exists an error region $R \in \Delta_\epsilon(\mathcal{C}, T)$ such that $S \cap R = \emptyset$, so if there exists an error region that is missed by S.

Example

- $X = [0,1]$ equipped with the uniform probability P;
- $\mathcal{C} = \{[a,b] \mid a,b \in [0,1]\} \cup \{\emptyset\}$;
- if $T = \emptyset$, $\Delta(\mathcal{C}, C) = \mathcal{C}$;
- $$S = \left\{ k\epsilon \,\middle|\, 1 \leqslant k \leqslant \left\lceil \frac{1}{\epsilon} \right\rceil \right\}$$

  is an ϵ-net for $\Delta(\mathcal{C}, \emptyset)$.



A Property of ϵ-Nets

Theorem

Let $\boldsymbol{s}$ be a sequence of examples. If there exits an ϵ-net for $\Delta_\epsilon(\mathcal{C}, T)$ and the output of the learning algorithm L is a hypothesis $H = L(\boldsymbol{s}) \in \mathcal{C}$ that is consistent with $\boldsymbol{s}$, then the error of H must be less than ϵ.

Proof: Since H is consistent with $\boldsymbol{s}$, $T \oplus H$ was not hit by $E_{\boldsymbol{s}}$ (otherwise H would not be consistent with $\boldsymbol{s}$). Thus, $T \oplus H \not\in \Delta_\epsilon(\mathcal{C}, T)$, so $err(H) = P(T \oplus H) \leqslant \epsilon$.

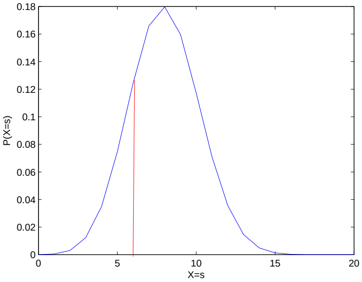
Main Theorem

Theorem

(Blumer et al.) *Let $\mathcal{C}$ be a concept class such that $VCD(\mathcal{C}) = d$. Then, $\mathcal{C}$ is potentially learnable.*

The proof consists in proving that $m \geqslant \frac{4}{\epsilon}\left(d \log \frac{12}{\epsilon} + \log \frac{2}{\delta}\right)$.

Chernoff's Bound



X is a binomial variable that corresponds to m drawings and probability of success is p. Then,

$P(X \leqslant s) \leqslant e^{-\frac{-\beta^2 mp}{2}}$,
where $\beta = 1 - \frac{s}{mp}$.

Draw a sequence $\boldsymbol{x}$ of m-random samples.

- A: takes place when $\boldsymbol{x}$ misses some $R \in \Delta_\epsilon(\mathcal{C}, T)$, that is, when $\boldsymbol{x}$ fails to form an ϵ-net for $\Delta(\mathcal{C}, T)$;
- fix R and draw a second m-sample $\boldsymbol{y}$;
- B: the combined event that takes place when:
  - we draw a sequence $\boldsymbol{xy}$ of length $2m$,
  - A occurs on $\boldsymbol{x}$,
  - and $\boldsymbol{y}$ has at least $\frac{m\epsilon}{2}$ hits in R in $\Delta_\epsilon(\mathcal{C}, T)$.

Let "success" be defined as occurring when an error occurs, that is, when $H(x) \neq T(x)$.

- let $p = P\left(\{x \in X \mid H(x) \neq T(x)\}\right)$;
- let $\ell_{p,m,s}$ be the probability of having at most s successes in m drawings; by the Chernoff bound for the binomial distribution we have

$$\ell_{p,m,s} \leqslant e^{-\frac{\beta^2 mp}{2}},$$

where $s = (1 - \beta)mp$.

- if we have at most s successes in m drawings, then

$$err_{\boldsymbol{y}}(H, T) = \frac{1}{m} \cdot |\{i \mid H(y_i) \neq T(y_i)\}| \leqslant \frac{s}{m},$$

  so $m \cdot err_{\boldsymbol{y}}(H, T)$ is a binomially distributed random variable with probability of success $err_{\boldsymbol{y}}(H, T) > \epsilon$;
- by applying these definitions we have:

$$\begin{aligned} & P\left(\left\{\boldsymbol{y} \mid err_{\boldsymbol{y}}(H, T) \leqslant \frac{\epsilon}{2}\right\}\right) \\ &= P\left(\left\{\boldsymbol{y} \mid m \cdot err_{\boldsymbol{y}}(H, T) \leqslant \frac{m \cdot \epsilon}{2}\right\}\right) \\ &= \ell\left(\epsilon, m, \frac{m \cdot \epsilon}{2}\right). \end{aligned}$$

Application of Chernoff's Bound

- $\beta = 1 - \frac{s}{m\epsilon} = 1 - \frac{\frac{m \cdot \epsilon}{2}}{m\epsilon} = \frac{1}{2}$
- $\ell(\epsilon, m, \frac{m \cdot \epsilon}{2}) \leqslant e^{-\frac{m\epsilon}{8}}$
- for $m \geqslant \frac{8}{\epsilon}$, $\ell(\epsilon, m, \frac{m \cdot \epsilon}{2}) \leqslant \frac{1}{\epsilon}$,

$$P\left(\left\{\mathbf{y} \mid err_{\mathbf{y}}(H, T) \leq \frac{\epsilon}{2}\right\}\right) \leqslant \frac{1}{\epsilon},$$

which implies for any $H \in \text{BAD}_\epsilon(T)$:

$$P\left(\left\{\mathbf{y} \mid err_{\mathbf{y}}(H, T) > \frac{\epsilon}{2}\right\}\right) \leqslant 1 - \frac{1}{\epsilon} > \frac{1}{2}.$$

The link between $P(A)$ and $P(B)$

- Since $P(B|A) \geqslant \frac{1}{2}$ and $B \subseteq A$, we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)} \geqslant \frac{1}{2},$$

  so $2P(B) \geqslant P(A)$.
- An upper bound on $P(A)$ can be found through an upper bound on $P(B)$.

Another way of looking to this problem

- draw randomly $2m$ balls;
- fix a region R in $R \in \Delta_\epsilon(\mathcal{C}, T)$ such that $|R| \geqslant \frac{\epsilon m}{2}$;
- randomy divide these into $\boldsymbol{x}$ and $\boldsymbol{y}$;
- analyze the probability that none of the x_i is in R with respect to the random division into $\boldsymbol{x}$ and $\boldsymbol{y}$;
- summing up over all possible fixed R and applying the union bound we obtain a bound on $P(B)$.

Reduction to A Combinatorial Problem

- an urn with $2m$ balls colored red or blue with ℓ red balls;
- divide the balls randomly into two groups S_1 and S_2 of equal size m;
- find an upper bound on the probability that all ℓ red balls fall in S_2;

A Combinatorial Problem (cont'd)

- there are $\binom{2m}{\ell}$ ways to paint $2m$ balls in red;
- if the red balls occur only in S_2 there are $\binom{m}{\ell}$ ways to paint in red these balls;
- the probability that all ℓ red balls belong to S_2 is

$$\begin{aligned} \frac{\binom{m}{\ell}}{\binom{2m}{\ell}} &= \frac{\frac{m!}{\ell!(m-\ell)!}}{\frac{(2m)!}{\ell!(2m-\ell)!}} \\ &= \frac{m!(2m-\ell)!}{(m-\ell)!(2m)!} \\ &= \prod_{i=0}^{\ell} \frac{m-i}{2m-i} \leqslant \prod_{i=0}^{\ell} \frac{1}{2} = 2^{-\ell}. \end{aligned}$$

Thus

$$\begin{aligned} P(B) &\leq \phi(d, 2m) 2^{-\frac{m\epsilon}{2}} \\ &\leq \left(\frac{2em}{d}\right)^{d} \cdot 2^{-\frac{m\epsilon}{2}} \end{aligned}$$

by the corollary of Sauer-Shelah Theorem

Therefore,

$$P(A) \leqslant 2P(B) \leqslant 2\left(\frac{2em}{d}\right)^{d} \cdot 2^{-\frac{m\epsilon}{2}}$$

The following statements are equivalent:

- $2\left(\frac{2em}{d}\right)^d \cdot 2^{-\frac{m\epsilon}{2}} \leq \delta$;
- $d \ln\left(\frac{2e}{d}\right) + d \ln m - \frac{\epsilon m}{2} \ln 2 \leqslant \ln \frac{\delta}{2}$;
- $\frac{\epsilon m}{2} \ln 2 - d \ln m \geqslant d \ln \frac{2e}{d} + \ln \frac{2}{\delta}$;
- choosing $c = \frac{\epsilon \ln 2}{4d}$ and $x = m$ in the inequality $\ln x \leqslant \left(\ln \frac{1}{c} - 1\right) + cx$ proven in the lemma,

$$d \ln m \leqslant d \left(\ln \frac{4d}{\epsilon \ln 2} - 1\right) + \frac{\epsilon \ln 2}{4} m.$$

Combining the inequalities

$$\begin{aligned} \frac{\epsilon m}{2} \ln 2 &\geqslant d \ln m + d \ln \frac{2e}{d} + \ln \frac{2}{\delta} \\ d \ln m &\leqslant d \left(\ln \frac{4d}{\epsilon \ln 2} - 1 \right) + \frac{\epsilon \ln 2}{4} m \end{aligned}$$

it follows that it suffices to have

$$\begin{aligned} \frac{\epsilon m}{4} \ln 2 &\geqslant d \left(\ln \frac{4d}{\epsilon \ln 2} - 1 \right) + d \ln \frac{2e}{d} + \ln \frac{2}{\delta} \\ &= d \ln \frac{8e}{\epsilon \ln 2} + \ln \frac{2}{\delta} - d = d \ln \frac{8}{\epsilon \ln 2} + \ln \frac{2}{\delta}. \end{aligned}$$

Since $\frac{8}{\ln 2} < 12$ the inequality

$$\frac{\epsilon m}{4} \ln 2 \geqslant d \ln \frac{8}{\epsilon \ln 2} + \ln \frac{2}{\delta}$$

can be satisfied by taking m such that

$$\frac{\epsilon m}{4} \ln 2 \geqslant d \ln \frac{12}{\epsilon} + \ln \frac{2}{\delta},$$

so $m \geqslant \frac{4}{\epsilon} \left(d \log \frac{12}{\epsilon} + \log \frac{2}{\delta} \right)$, which concludes the proof.

Where to look further...

- M. Anthony and N. Biggs: Computational Learning Theory, Cambridge, 1997
- V.N. Vapnik: Statistical Learning Theory, J. Wiley, 1998
- M. Vidyasagar: Learning and generalization with applications to neural networks, Springer Verlag, 2003
- D. Simovici and C. Djeraba: Mathematical Tools for Data Mining, Springer, 2008