

The Vapnik-Chervonenkis Dimension

Prof. Dan A. Simovici

UMB

- 1 Growth Functions
- 2 Basic Definitions for Vapnik-Chervonenkis Dimension
- 3 The Sauer-Shelah Theorem
- 4 The Link between VCD and PAC Learning
- 5 The VCD of Collections of Sets

Definition

Let H be a set of hypotheses and let (x_1, \dots, x_m) be a sequence of examples of length m . A hypothesis $h \in H$ induces a classification

$$(h(x_1), \dots, h(x_m))$$

of the components of this sequence. The **growth function** of H is the function $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ gives the number of ways a sequence of examples of length m can be classified by a hypothesis in H :

$$\Pi_H(m) = \max_{(x_1, \dots, x_m) \in \mathcal{X}^m} |\{(h(x_1), \dots, h(x_m)) \mid h \in H\}|$$

Dichotomies

Definition

A **dichotomy** is a hypothesis $h : \mathcal{X} \rightarrow \{-1, 1\}$.

If H consists of dichotomies, then (x_1, \dots, x_m) can be classified in at most 2^m ways.

Trace of a Collection of Sets

Definition

Let \mathcal{C} be a collection of sets and let K be a set. The **trace of \mathcal{C} on K** is the collection

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

Definition

Let \mathcal{C} be a collection of sets. If the trace of \mathcal{C} on K , \mathcal{C}_K equals $\mathcal{P}(K)$, then we say that K is **shattered by \mathcal{C}** .

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 $VCD(\mathcal{C})$.

If $VCD(\mathcal{C}) = d$, then there exists a set K of size d such that for each subset L of K there exists a set $C \in \mathcal{C}$ such that $L = K \cap C$.

A collection \mathcal{C} shatters a set K if and only if \mathcal{C}_K shatters K . This allows us to assume without loss of generality that **both the sets of the collection \mathcal{C} and a set K shattered by \mathcal{C} are subsets of a set U** .

Collections of Sets as Sets of Hypotheses

Let U be a set, K a subset, and let \mathcal{C} be a collection of sets.

Each $C \in \mathcal{C}$ defines a hypothesis $h_C : S \rightarrow \{-1, 1\}$ that is a dichotomy, where

$$h_C(u) = \begin{cases} 1 & \text{if } u \in C, \\ -1 & \text{if } u \notin C. \end{cases}$$

K is shattered by \mathcal{C} if and only if for **every** subset L of K there exists a hypothesis h_C such that L_{pos} consists of the positive examples of h_C .

Finite Collections have Finite VC-Dimension

Let \mathcal{C} be a collection of sets with $VCD(\mathcal{C}) = d$ and let K be a set shattered by \mathcal{C} with $|K| = d$. Since there exist 2^d subsets of K , there are at least 2^d subsets of \mathcal{C} , so $2^d \leq |\mathcal{C}|$. Consequently, $VCD(\mathcal{C}) \leq \log_2 |\mathcal{C}|$. This shows that if \mathcal{C} is finite, then $VCD(\mathcal{C})$ is finite. The converse is false: there exist infinite collections \mathcal{C} that have a finite VC-dimension.

A Tabular Representation of Shattering

If $U = \{u_1, \dots, u_n\}$ is a finite set, then the trace of a collection $\mathcal{C} = \{C_1, \dots, C_p\}$ of subsets of U on a subset K of U can be presented in an intuitive, tabular form.

Let θ be a table containing the rows t_1, \dots, t_p and the binary attributes u_1, \dots, u_n .

Each tuple t_k corresponds to a set C_k of \mathcal{C} and is defined by

$$t_k[u_i] = \begin{cases} 1 & \text{if } u_i \in C_k, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i \leq n$. Then, \mathcal{C} shatters K if the content of the projection $\mathbf{r}[K]$ consists of $2^{|K|}$ distinct rows.

Example

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

$\mathcal{C} = \{\{u_2, u_3\}, \{u_1, u_3, u_4\}, \{u_2, u_4\}, \{u_1, u_2\}, \{u_2, u_3, u_4\}\}$ represented by:

$$T_{\mathcal{C}}$$

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

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

$$r[K] = ((0, 1), (1, 1), (0, 0), (1, 0), (0, 1))$$

contains the all four necessary tuples $(0, 1)$, $(1, 1)$, $(0, 0)$, and $(1, 0)$. On the other hand, it is clear that no subset K of U that contains at least three elements can be shattered by \mathcal{C} because this would require $r[K]$ to contain at least eight tuples. Thus, $VCD(\mathcal{C}) = 2$.

- every collection of sets shatters the empty set;
- if \mathcal{C} shatters a set of size n , then it shatters a set of size p , where $p \leq n$.

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

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

be 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} .

- We have $\Pi_{\mathcal{C}}[m] \leq 2^m$;
- if \mathcal{C} shatters a set of size m , then $\Pi_{\mathcal{C}}[m] = 2^m$.

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 \mathbb{R} be the set of real numbers and let \mathcal{S} be the collection of sets $\{(-\infty, t) \mid t \in \mathbb{R}\}$.

We claim that any singleton is shattered by \mathcal{S} . Indeed, if $S = \{x\}$ is a singleton, then $\mathcal{P}(\{x\}) = \{\emptyset, \{x\}\}$. Thus, if $t \geq x$, 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} . Indeed, 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 $VCD(\mathcal{S}) = 1$.

Example

Consider the collection $\mathcal{I} = \{[a, b] \mid a, b \in \mathbb{R}, a \leq b\}$ of closed intervals. We claim that $VCD(\mathcal{I}) = 2$. To justify this claim, we need to show that there exists a set $S = \{x, y\}$ such that $\mathcal{I}_S = \mathcal{P}(S)$ and no three-element set can be shattered by \mathcal{I} .

For the first part of the statement, consider the intersections

$$[u, v] \cap S = \emptyset, \text{ where } v < x,$$

$$[x - \epsilon, \frac{x+y}{2}] \cap S = \{x\},$$

$$[\frac{x+y}{2}, y] \cap S = \{y\},$$

$$[x - \epsilon, y + \epsilon] \cap S = \{x, y\},$$

which show that $\mathcal{I}_S = \mathcal{P}(S)$.

For the second part of the statement, let $T = \{x, y, z\}$ be a set that contains three elements. 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 .

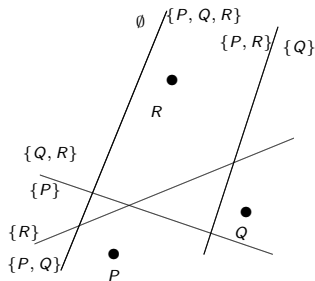
An Example

Let \mathcal{H} be the collection of closed half-planes in \mathbb{R}^2 of the form

$$\{x = (x_1, x_2) \in \mathbb{R}^2 \mid ax_1 + bx_2 - c \geq 0, a \neq 0 \text{ or } b \neq 0\}.$$

We claim that $VCD(\mathcal{H}) = 3$.

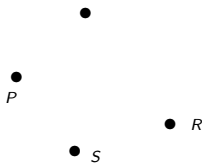
Let P, Q, R be three non-collinear points. Each line is marked with the sets it defines; thus, it is clear that the family of half-planes shatters the set $\{P, Q, R\}$, so $VCD(\mathcal{H})$ is at least 3.



Example (cont'd)

To complete the justification of the claim we need to show that no set that contains at least four points can be shattered by \mathcal{H} .

Let $\{P, Q, R, S\}$ be a set that contains four points such that no three points of this set are collinear. If S is located inside the triangle P, Q, R , then every half-plane that contains P, Q, R also contains 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. Observe that any half-plane that contains two diagonally opposite points, for example, P and R , contains 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 $VCD(\mathcal{H}) = 3$.

A family of $d + 1$ points in \mathbb{R}^d can be shattered by hyperplanes. Consider the points

$$\mathbf{x}_0 = \mathbf{0}_d, \mathbf{x}_i = \mathbf{e}_i \text{ for } 1 \leq i \leq d.$$

Let $y_0, y_1, \dots, y_d \in \{-1, 1\}$ and let \mathbf{w} be the vector whose i^{th} coordinate is y_i . We have $\mathbf{w}'\mathbf{x}_i = y_i$ for $1 \leq i \leq d$, so

$$\text{sign} \left(\mathbf{w}'\mathbf{x}_i + \frac{y_0}{2} \right) = \text{sign} \left(y_i + \frac{y_0}{2} \right) = y_i.$$

Thus, points \mathbf{x}_i for which $y_i = 1$ are on the positive side of the hyperplane $\mathbf{w}'\mathbf{x} = 0$; the ones for which $y_i = -1$ are on the opposite side, so any family of $d + 1$ points in \mathbb{R}^d can be shattered by hyperplanes.

To obtain an upper bound we need to show that no set of $d + 2$ points can be shattered by half-spaces. For this we need the following result:

Theorem

(Radon's Theorem) Any set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_{d+2}\}$ of $d + 2$ points in \mathbb{R}^d can be partitioned into two sets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.

Proof

Consider the following system with $d + 1$ linear equations and $d + 2$ variables $\alpha_1, \alpha_2, \dots, \alpha_{d+2}$:

$$\sum_{i=1}^{d+2} \alpha_i \mathbf{x}_i = \mathbf{0}_d, \sum_{i=1}^{d+2} \alpha_i = 0.$$

Since the number of variables ($d + 2$) is larger than $d + 1$, the system has a non-trivial solution $\beta_1, \dots, \beta_{d+2}$. Since $\sum_{i=1}^{d+2} \beta_i = 0$ both sets

$$I_1 = \{i | 1 \leq i \leq d + 2, \beta_i > 0\}, I_2 = \{i | 1 \leq i \leq d + 2, \beta_i < 0\}$$

are non-empty sets and

$$X_1 = \{\mathbf{x}_i \mid i \in I_1\}, X_2 = \{\mathbf{x}_i \mid i \in I_2\},$$

form a partition of X .

Proof (cont'd)

Define $\beta = \sum_{i \in I_1} \beta_i$. Since $\sum_{i \in I_1} \beta_i = -\sum_{i \in I_2} \beta_i$, we have

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} \mathbf{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} \mathbf{x}_i.$$

Also,

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} = \sum_{i \in I_2} \frac{-\beta_i}{\beta} = 1,$$

$\frac{\beta_i}{\beta} \geq 0$ for $i \in I_1$ and $\frac{-\beta_i}{\beta} \geq 0$ for $i \in I_2$. This implies that

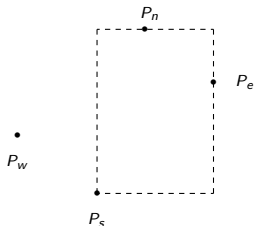
$$\sum_{i \in I_1} \frac{\beta_i}{\beta} \mathbf{x}_i$$

belongs both to the convex hulls of X_1 and X_2 .

Let X be a set of $d + 2$ points in \mathbb{R}^d . By Radon's Theorem it can be partitioned into X_1 and X_2 such that the two convex hulls intersect. When two sets are separated by a hyperplane, their convex hulls are also separated by the hyperplane. Thus, X_1 and X_2 cannot be separated by a hyperplane and X is not shattered.

Example

Let \mathcal{R} be the set of rectangles whose sides are parallel with the axes x and y . There is a set S with $|S| = 4$ that is shattered by \mathcal{R} . 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 . For example, we show the rectangle R_L for the set $\{P_n, P_s, P_e\}$.



Example (cont'd)

This collection cannot shatter a set of points that contains at least five points. Indeed, let S be such that $|S| \geq 5$. 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 S .

The Class of Convex Polygons

Example

Consider the system of all convex polygons in the plane.

For any positive integer m , place m points on the unit circle. Any subset of the points are the vertices of a convex polygon. Clearly that polygon will not contain any of the points not in the subset. This shows that we can shatter arbitrarily large sets, so the VC-dimension is infinite.

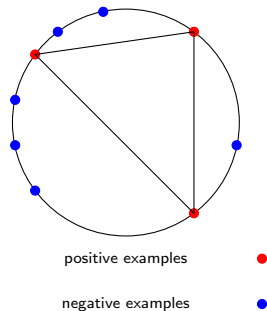
The Case of Convex Polygons with d Vertices

Example

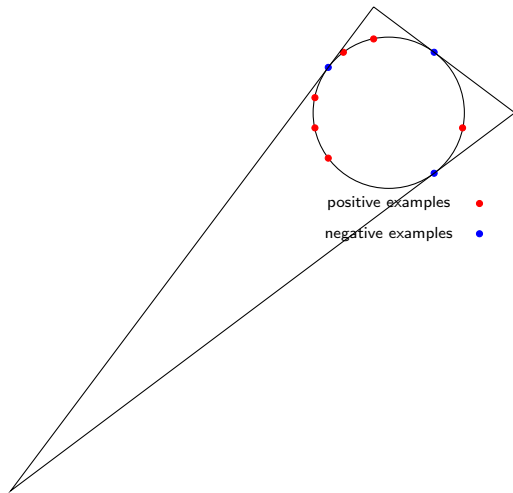
Consider the class of convex polygons that have no more than d vertices in \mathbb{R}^2 and place $2d + 1$ points placed on the circle.

- Label a subset of these points as positive, and the remaining points as negative. Since we have an odd number of points there exists a majority in one of the classes (positive or negative).
- If the negative points are in majority, there are at most d positive points; these are contained by the convex polygon formed by joining the positive points.
- If the positive points are in majority, consider the polygon formed by the tangents of the negative points.

Negative Points in the Majority



Positive Points in the Majority



Example cont'd

- Since a set with $2d + 1$ points can be shattered, the VC dimension of the set of convex polygons with at most d vertices is at least $2d + 1$.
- Note that if all labeled points are located on a circle then it is impossible for a point to be in the convex closure of a subsets of the remaining points. Thus, placing the points on a circle maximizes the number of sets required to shatter the set, so the VC-dimension is indeed $2d + 1$.

Recall that for a collection of sets \mathcal{C} the growth function $\Pi_{\mathcal{C}}$ was defined as

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

for $m \in \mathbb{N}$.

Two completely distinct situations may occur:

- If a collection of sets \mathcal{C} is not a VC class (that is, if the Vapnik-Chervonenkis dimension of \mathcal{C} is infinite), then $\Pi_{\mathcal{C}}[m] = 2^m$ for all $m \in \mathbb{N}$.
- If $VCD(\mathcal{C}) = d$, then $\Pi_{\mathcal{C}}[m]$ is bounded asymptotically by a polynomial of degree d .

Definition

Let \mathcal{C} be a collection of subsets of a finite set S , let $s \in S$, and let $\mathcal{C}_{(s)}$ be the collection $\mathcal{C}_{S-\{s\}}$.

\mathcal{C} has a pair (A, B) at s if $A, B \in \mathcal{C}$, $B \subseteq A$, and $A - B = \{s\}$.

Note that if (A, B) is a pair of \mathcal{C} at s , then $B = A - \{s\}$.

Define $\mathcal{P}'(\mathcal{C}, s)$ and $\mathcal{P}''(\mathcal{C}, s)$ as

$$\begin{aligned}\mathcal{P}'(\mathcal{C}, s) &= \{A \in \mathcal{C} \mid (A, B) \text{ is a pair at } s \text{ for some } B \in \mathcal{C}\} \\ &= \{A \in \mathcal{C} \mid s \in A, A - \{s\} \in \mathcal{C}\}, \\ \mathcal{P}''(\mathcal{C}, s) &= \{B \in \mathcal{C} \mid (A, B) \text{ is a pair at } s \text{ for some } A \in \mathcal{C}\} \\ &= \{B \in \mathcal{C} \mid s \notin B, B \cup \{s\} \in \mathcal{C}\}.\end{aligned}$$

Let \mathcal{C} be a collection of subsets of a finite set S , let $s \in S$, and let $\mathcal{C}_{(s)}$ be the collection

$$\mathcal{C}_{(s)} = \{C \cap (S - \{s\}) \mid C \in \mathcal{C}\} = \{C - \{s\} \mid C \in \mathcal{C}\},$$

that is the trace $\mathcal{C}_{S-\{s\}}$ of \mathcal{C} on $S - \{s\}$.

Lemma

The following statements hold for any collection \mathcal{C} of subsets of a set S and any $s \in S$:

- (i) *if $(A_1, B_1), (A_1, B_2), (A_2, B_1), (A_2, B_2)$ are pairs at s of \mathcal{C} , then $A_1 = A_2$ and $B_1 = B_2$;*
- (ii) *we have $|\mathcal{P}'(\mathcal{C}, s)| = |\mathcal{P}''(\mathcal{C}, s)|$;*
- (iii) *$|\mathcal{C}| - |\mathcal{C}_{(s)}| = |\mathcal{P}''(\mathcal{C}, s)|$;*
- (iv) *$\mathcal{C}_{(s)} = \{C \in \mathcal{C} \mid s \notin C\} \cup \{C - \{s\} \mid C \in \mathcal{C} \text{ and } s \in C\}$.*

Proof

For [Part \(i\)](#), by definition we have:

$$B_1 \subseteq A_1, B_2 \subseteq A_1, B_1 \subseteq A_2, B_2 \subseteq A_2,$$

and $A_1 = B_1 \cup \{s\}$, $A_1 = B_2 \cup \{s\}$, $A_2 = B_1 \cup \{s\}$, and $A_2 = B_2 \cup \{s\}$.

Therefore, $A_1 = B_1 \cup \{s\} = A_2$, which implies

$$B_1 = A_1 - \{s\} = A_2 - \{s\} = B_2.$$

For [Part \(ii\)](#), let $f : \mathcal{P}'(\mathcal{C}, s) \longrightarrow \mathcal{P}''(\mathcal{C}, s)$ be the function $f(A) = A - \{s\}$.

It is easy to verify that f is a bijection and this implies

$$|\mathcal{P}'(\mathcal{C}, s)| = |\mathcal{P}''(\mathcal{C}, s)|.$$

Proof (cont'd)

Recall that $\mathcal{C}_{(s)}$ is the collection $\mathcal{C}_{S-\{s\}}$.

To prove [Part \(iii\)](#), $|\mathcal{C}| - |\mathcal{C}_{(s)}| = |\mathcal{P}''(\mathcal{C}, s)|$, let $B \in \mathcal{P}''(\mathcal{C}, s)$. By the definition of $\mathcal{P}''(\mathcal{C}, s)$ we have $B \in \mathcal{C}$, $s \notin B$, and $B \cup \{s\} \in \mathcal{C}$.

Define the mapping

$$g : \mathcal{P}''(\mathcal{C}, s) \longrightarrow \mathcal{C} - \mathcal{C}_{(s)} \text{ as } g(B) = B \cup \{s\}.$$

Note that $B \cup \{s\} \in \mathcal{C} - \mathcal{C}_{(s)}$, so g is a well-defined function. Moreover, g is one-to one because $B_1 \cup \{s\} = B_2 \cup \{s\}$, $s \notin B_1$, and $s \notin B_2$ imply $B_1 = B_2$. Also, g is a surjection because if $D \in \mathcal{C} - \mathcal{C}_{(s)}$, then $s \in D$ and $D - \{s\} \in \mathcal{P}''(\mathcal{C}, s)$. Thus, g is a bijection, which implies $|\mathcal{C}| - |\mathcal{C}_{(s)}| = |\mathcal{C} - \mathcal{C}_{(s)}| = |\mathcal{P}''(\mathcal{C}, s)|$.

Proof (cont'd)

To prove Part (iv),

$$\mathcal{C}_{(s)} = \{C \in \mathcal{C} \mid s \notin C\} \cup \{C - \{s\} \mid C \in \mathcal{C} \text{ and } s \in C\},$$

let $C \in \mathcal{C}_{(s)}$. If $s \notin C$, then C belongs to the first collection of the union; otherwise, that is if $s \in C$, $C - \{s\}$ belongs to the second collection and the equality follows.

Lemma

Let \mathcal{C} be a collection of sets of a non-empty finite set S and let s_0 be an element of S . If $\mathcal{P}''(\mathcal{C}, s_0)$ shatters a subset T of $S - \{s_0\}$, then \mathcal{C} shatters $T \cup \{s_0\}$.

Proof

Since $\mathcal{P}''(\mathcal{C}, s_0)$ shatters T , for every subset U of T there is $B \in \mathcal{P}''(\mathcal{C}, s_0)$ such that $U = T \cap B$. Let W be a subset of $T \cup \{s_0\}$. If $s_0 \notin W$, then $W \subseteq T$ and by the previous assumption, there exists $B \in \mathcal{P}''(\mathcal{C}, s_0)$ such that $W = (T \cup \{s_0\}) \cap B$. If $s_0 \in W$, then there exists $B_1 \in \mathcal{P}''(\mathcal{C}, s_0)$ such that for $W_1 = W - \{s_0\}$ we have $W_1 = T \cap B_1$. By the definition of $\mathcal{P}''(\mathcal{C}, s_0)$, $B_1 \cup \{s_0\} \in \mathcal{C}$ and

$$(T \cup \{s_0\}) \cap (B_1 \cup \{s_0\}) = (T \cap B_1) \cup \{s_0\} = W_1 \cup \{s_0\} = W.$$

Thus, \mathcal{C} shatters $T \cup \{s_0\}$.

We saw that we have $VCD(\mathcal{C}) \leq \log_2 |\mathcal{C}|$. For collections of subsets of finite sets we have a stronger result.

Theorem

Let \mathcal{C} be a collection of sets of a non-empty finite set S with $VCD(\mathcal{C}) = d$. We have

$$2^d \leq |\mathcal{C}| \leq (|S| + 1)^d.$$

Proof

The first inequality, reproduced here for completeness, was discussed earlier.

For the second inequality the argument is by induction on $|S|$. The basis case, $|S| = 1$ is immediate.

Suppose that the inequality holds for collections of subsets with no more than n elements and let S be a set containing $n + 1$ elements. Let s_0 be an arbitrary but fixed element of S . By Part (iii) of Lemma given on slide 31, we have

$$|\mathcal{C}| = |\mathcal{C}_{(s_0)}| + |\mathcal{P}''(\mathcal{C}, s_0)|.$$

The collection $\mathcal{C}_{(s_0)}$ consists of subsets of $S - \{s_0\}$. Since $VCD(\mathcal{C}) = d$, it is clear that $VCD(\mathcal{C}_{(s_0)}) \leq d$ and, by inductive hypothesis $|\mathcal{C}_{(s_0)}| \leq (|S - \{s_0\}| + 1)^d$.

Proof (cont'd)

We claim that $VCD(\mathcal{P}''(\mathcal{C}, s_0)) \leq d - 1$. Suppose that

$$\mathcal{P}''(\mathcal{C}, s_0) = \{B \in \mathcal{C} \mid s_0 \notin B, B \cup \{s_0\} \in \mathcal{C}\}$$

shatters a set T , where $T \subseteq S$ and $|T| \geq d$. Then, by Lemma located on slide 35, \mathcal{C} would shatter $T \cup \{s_0\}$; since $|T \cup \{s_0\}| \geq d + 1$, this would lead to a contradiction. Therefore, we have $VCD(\mathcal{P}''(\mathcal{C}, s_0)) \leq d - 1$ and, by the inductive hypothesis, $|\mathcal{P}''(\mathcal{C}, s_0)| \leq (|S - \{s_0\}| + 1)^{d-1}$. These inequalities imply

$$\begin{aligned} |\mathcal{C}| &= |\mathcal{C}_{(s_0)}| + |\mathcal{P}''(\mathcal{C}, s_0)| \\ &\leq (|S - \{s_0\}| + 1)^d + (|S - \{s_0\}| + 1)^{d-1} \\ &= (|S - \{s_0\}| + 1)^{d-1}(|S - \{s_0\}| + 2) \\ &\leq (|S| + 1)^d. \end{aligned}$$

Lemma

Let \mathcal{C} be a collection of subsets of a finite set S , let $s \in S$, and let $\mathcal{C}_{(s)}$ be the collection $\mathcal{C}_{S-\{s\}}$. If $VCD(\mathcal{P}''(\mathcal{C}, s)) = n - 1$ in $S - \{s\}$, then $VCD(\mathcal{C}) = n$.

Proof

Since $VCD(\mathcal{P}''(\mathcal{C}, s)) = n - 1$ in $S - \{s\}$, there exists a set $T \subseteq S - \{s\}$ with $|T| = n - 1$ that is shattered by $\mathcal{P}''(\mathcal{C}, s)$. Let \mathcal{G} be the subcollection of \mathcal{C} defined by:

$$\begin{aligned}\mathcal{G} &= \mathcal{P}'(\mathcal{C}, s) \cup \mathcal{P}''(\mathcal{C}, s) \\ &= \{A \in \mathcal{C} \mid A - \{s\} \in \mathcal{C}\} \cup \{B \in \mathcal{C} \mid B \cup \{s\} \in \mathcal{C}\}.\end{aligned}$$

We claim that \mathcal{G} shatters $T \cup \{s\}$. Two cases may occur for a subset W of $T \cup \{s\}$:

- If $s \notin W$, then $W \subseteq T$ and, since T is shattered by $\mathcal{P}''(\mathcal{C}, s)$ it follows that there exists $G \in \mathcal{G}$ such that $W = G \cap (T \cup \{s\})$.
- If $s \in W$, let $G' \in \mathcal{P}''(\mathcal{C}, s)$ be set such that $W - \{s\} = T \cap G'$, which exists by the previous argument. Then, we have $G' \in \mathcal{P}''(\mathcal{C}, s)$, so $G' \cup \{s\} \in \mathcal{P}'(\mathcal{C}, s)$ and $(G' \cap \{s\}) \cap (T \cup \{s\}) = W$.

Thus, \mathcal{G} shatters $T \cup \{s\}$ and so does \mathcal{C} .

For $n, k \in \mathbb{N}$ and $0 \leq k \leq n$ define the number $\binom{n}{\leq k}$ as

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

Clearly, $\binom{n}{\leq 0} = 1$ and $\binom{n}{\leq n} = 2^n$.

Theorem

Let $\phi : \mathbb{N}^2 \rightarrow \mathbb{N}$ 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

$$\phi(d, m) = \binom{m}{\leq d}$$

for $d, m \in \mathbb{N}$.

Proof

The argument is by strong induction on $s = d + m$. The base case, $s = 0$, implies $m = 0$ and $d = 0$, and the equality is immediate. Suppose that the equality holds for $\phi(d', m')$, where $d' + m' < d + m$. We have

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

which gives the desired conclusion.

The Sauer-Shelah Theorem

Theorem

Let \mathcal{C} be a collection of sets such that $VCD(\mathcal{C}) = d$. If $m \in \mathbb{N}$ is a number such that $d \leq m$, then $\Pi_{\mathcal{C}}[m] \leq \phi(d, m)$.

Proof: The argument is by strong induction on $s = d + m$. For the base case, $s = 0$ we have $d = m = 0$ and this means that the collection \mathcal{C} shatters only the empty set. Thus, $\Pi_{\mathcal{C}}[0] = |\mathcal{C}_{\emptyset}| = 1$, and this implies $\Pi_{\mathcal{C}}[0] = 1 = \phi(0, 0)$.

Proof (cont'd)

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 $VCD(\mathcal{C}) = d$ and K be a set with $|K| = m$.

Let k_0 be a fixed (but, otherwise, arbitrary) element of K . Since $|K - \{k_0\}| = m - 1$, by inductive hypothesis, we have $|\mathcal{C}_{K - \{k_0\}}| \leq \phi(d, m - 1)$.

Define

$$\mathcal{D} = \mathcal{P}''(\mathcal{C}_K, k_0) = \{B \in \mathcal{C}_K \mid k_0 \notin B \text{ and } B \cup \{k_0\} \in \mathcal{C}_K\}.$$

By a previous Lemma, $VCD(\mathcal{D}) \leq d - 1$, so $|\mathcal{D}| \leq \phi(d - 1, m - 1)$.

By Part (ii) of another Lemma,

$$|\mathcal{C}_K| = |\mathcal{C}_{K - \{k_0\}}| + |\mathcal{D}| \leq \phi(d, m - 1) + \phi(d - 1, m - 1) = \phi(d, m).$$

Lemma

For $d \in \mathbb{N}$ and $d \geq 2$ we have

$$2^{d-1} \leq \frac{d^d}{d!}.$$

Proof: The argument is by induction on d . In the basis step, $d = 2$ both members are equal to 2.

Suppose the inequality holds for d . We have

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

because

$$\left(1 + \frac{1}{d}\right)^d \geq 1 + d \frac{1}{d} = 2.$$

This concludes the proof of the inequality.

Lemma

We have $\phi(d, m) \leq 2 \frac{m^d}{d!}$ for every $m \geq d$ and $d \geq 1$.

Proof: The argument is by induction on d and n . If $d = 1$, then $\phi(1, m) = m + 1 \leq 2m$ for $m \geq 1$, so the inequality holds for every $m \geq 1$, when $d = 1$.

Proof (cont'd)

If $m = d \geq 2$, then $\phi(d, m) = \phi(d, d) = 2^d$ and the desired inequality follows immediately from a previous Lemma.

Suppose that the inequality holds for $m > d \geq 1$. We have

$$\begin{aligned}
 \phi(d, m+1) &= \phi(d, m) + \phi(d-1, m) \\
 &\quad \text{(by the definition of } \phi) \\
 &\leq 2 \frac{m^d}{d!} + 2 \frac{m^{d-1}}{(d-1)!} \\
 &\quad \text{(by inductive hypothesis)} \\
 &= 2 \frac{m^{d-1}}{(d-1)!} \left(1 + \frac{m}{d} \right).
 \end{aligned}$$

Proof (cont'd)

It is easy to see that the inequality

$$2 \frac{m^{d-1}}{(d-1)!} \left(1 + \frac{m}{d}\right) \leq 2 \frac{(m+1)^d}{d!}$$

is equivalent to

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

and, therefore, is valid. This yields immediately the inequality of the lemma.

The Asymptotic Behavior of the Function ϕ

Theorem

The function ϕ satisfies the inequality:

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

for every $m \geq d$ and $d \geq 1$.

Proof: By a previous Lemma, $\phi(d, m) \leq 2 \frac{m^d}{d!}$. Therefore, we need to show only that

$$2 \left(\frac{d}{e}\right)^d < d!.$$

Proof (cont'd)

The argument is by induction on $d \geq 1$. The basis case, $d = 1$ is immediate. Suppose that $2 \left(\frac{d}{e}\right)^d < d!$. We have

$$\begin{aligned} 2 \left(\frac{d+1}{e}\right)^{d+1} &= 2 \left(\frac{d}{e}\right)^d \left(\frac{d+1}{d}\right)^d \frac{d+1}{e} \\ &= \left(1 + \frac{1}{d}\right)^d \frac{1}{e} \cdot 2 \left(\frac{d}{e}\right)^d (d+1) < 2 \left(\frac{d}{e}\right)^d (d+1), \end{aligned}$$

because

$$\left(1 + \frac{1}{d}\right)^d < e.$$

The last inequality holds because the sequence $\left(\left(1 + \frac{1}{d}\right)^d\right)_{d \in \mathbb{N}}$ is an increasing sequence whose limit is e . Since $2 \left(\frac{d+1}{e}\right)^{d+1} < 2 \left(\frac{d}{e}\right)^d (d+1)$, by inductive hypothesis we obtain:

$$2 \left(\frac{d+1}{e}\right)^{d+1} < (d+1)!.$$

Corollary

If m is sufficiently large we have $\phi(d, m) = O(m^d)$.

The statement is a direct consequence of the previous theorem.

Denote by \oplus the symmetric difference of two sets.

Theorem

Let \mathcal{C} a family of sets and $C_0 \in \mathcal{C}$. Define the family Δ_{C_0} as

$$\Delta_{C_0}(\mathcal{C}) = \{T \mid T = C_0 \oplus C \text{ where } C \in \mathcal{C}\}.$$

We have $VCD(\mathcal{C}) = VCD(\Delta_{C_0}(\mathcal{C}))$.

Proof

Let S be a set, $\mathcal{S} = \mathcal{C}_S$ and $\mathcal{S}_0 = (\Delta_{C_0}(\mathcal{C}))_S$. Define $\psi : \mathcal{S} \longrightarrow \mathcal{S}_0$ as $\psi(S \cap C) = S \cap (C_0 \oplus C)$. We claim that ψ is a bijection.

If $\psi(S \cap C) = \psi(S \cap C')$ for $C, C' \in \mathcal{C}$, then $S \cap (C_0 \oplus C) = S \cap (C_0 \oplus C')$. Therefore,

$$(S \cap C_0) \oplus (S \cap C) = (S \cap C_0) \oplus (S \cap C'),$$

which implies $S \cap C = S \cap C'$, so ψ is injective.

On the other hand, if $U \in \mathcal{S}_0$ we have $U = S \cap (C_0 \oplus C)$, so $U = \psi(S \cap C)$, hence ψ is a surjection. Thus, \mathcal{S} and \mathcal{S}_0 have the same number of sets, which implies that a set S is shattered by \mathcal{C} if and only if it is shattered by $\Delta_{C_0}(\mathcal{C})$.

Classes with Infinte VCDs Are not PAC-learnable

Theorem

A concept class \mathcal{C} with $VCD(\mathcal{C}) = \infty$ is not PAC-learnable.

Proof: Assume that \mathcal{C} is PAC-learnable. Let \mathcal{A} be a training algorithm and let m be the sample size needed to learn \mathcal{C} with accuracy ϵ and certainty $1 - \delta$. In other words, after seeing m examples, \mathcal{A} produces a hypothesis h with $P(\hat{R}(h) \leq \epsilon) \geq 1 - \delta$.

Since $VCD(\mathcal{C}) = \infty$, there exists a sample S of length $2m$ that is shattered by \mathcal{C} . Define the probability distribution \mathcal{D} such that the probability of each example x_i of S is $\frac{1}{2m}$ and the probability of other examples is 0. Since S is shattered, we can choose a target concept c_t such that

$$P(c_t(x_i) = 0) = P(c_t(x_i) = 1) = \frac{1}{2}$$

for every x_i in S (as if the labels $c_t(x_i)$ are determined by a coin flip).

Proof (cont'd)

\mathcal{A} selects an iid sample of m instances S' such that $S' \subseteq S$ and outputs a consistent hypothesis h . The probability of error for each $x_i \notin S'$ is

$$P(c_t(x_i) \neq h(x_i)) = \frac{1}{2}$$

because we could select the labels of the points not seen by \mathcal{A} arbitrarily. Regardless of h we have

$$E(\hat{R}(h)) = m \cdot 0 \cdot \frac{1}{2m} + m \cdot \frac{1}{2} \cdot \frac{1}{2m} = \frac{1}{4}.$$

(We have $2m$ points to sample such that the error of half of them is 0 as h is consistent on S').

Proof (cont'd)

Thus, for any sample size m , if \mathcal{A} produces a consistent hypothesis, then the expectation of the error will be $\frac{1}{4}$.

However, since with probability at least $1 - \delta$ we have that $\hat{R}(h) \leq \epsilon$, it follows that

$$E(\hat{R}(h)) \leq (1 - \delta)\epsilon + \delta \cdot \beta,$$

where $\epsilon < \beta \leq 1$. Note that

$$(1 - \delta)\epsilon + \delta \cdot \beta \leq (1 - \delta)\epsilon + \delta = \epsilon + \delta - \epsilon\delta < \epsilon + \delta$$

It suffices to take

$$\epsilon + \delta < \frac{1}{4}$$

to obtain a contradiction!

Hypothesis Consistency in Set-Theoretical Terms

Let C be a concept over the set of examples \mathcal{X} and let S be a sample drawn from \mathcal{X} according to a probability distribution \mathcal{D} .

- A hypothesis C_0 regarded here as a set, is consistent with S if $C_0 \cap S = C \cap S$. Equivalently, $S \cap (C_0 \oplus C) = \emptyset$.
- C_0 is inconsistent with S if $S \cap (C_0 \oplus C) \neq \emptyset$, that is, S **hits** the error region $C_0 \oplus S$.

On slide 54 we established that $VCD(\mathcal{C}) = VCD(\Delta_{C_0}(\mathcal{C}))$, where

$$\Delta_{C_0}(\mathcal{C}) = \{T \mid T = C_0 \oplus C \mid C \in \mathcal{C}\}.$$

Define now

$$\begin{aligned} \Delta_{C_0, \epsilon}(\mathcal{C}) &= \{T \mid T \in \Delta_{C_0}(\mathcal{C}) \mid P(T) \geq \epsilon\} \\ &= \{T \mid T = C_0 \oplus C, C \in \mathcal{C} \text{ and } P(T) \geq \epsilon\}. \end{aligned}$$

- $\Delta_{C_0}(\mathcal{C})$ is the set of **error regions relative to the hypothesis C_0** ;
- $\Delta_{C_0, \epsilon}(\mathcal{C})$ is the set of **error regions relative to the hypothesis C_0** having the probability at least equal to ϵ .

Definition

A set S is an ϵ -net for $\Delta_{C_0}(\mathcal{C})$ if every set T in $\Delta_{C_0, \epsilon}(\mathcal{C})$ is hit by S , that is, for every error region $T \in \Delta_{C_0, \epsilon}(\mathcal{C})$ we have $S \cap T \neq \emptyset$.

Claim:

*If the sample S forms an ϵ -net for $\Delta_{C_0}(\mathcal{C})$ and the learning algorithm outputs a hypothesis (represented here by a set $C_0 \in \mathcal{C}$) that is consistent with S , then this hypothesis **must have error less than ϵ** .*

Indeed, since

- $C_0 \oplus C \in \Delta_{C_0}(\mathcal{C})$ was not hit by S (otherwise, C_0 would not be consistent with S), and
- S is an ϵ -net for $\Delta_{C_0}(\mathcal{C})$,

we must have $C_0 \oplus C \notin \Delta_{C_0, \epsilon}(\mathcal{C})$ and therefore $\hat{R}(C_0) \leq \epsilon$.

Thus, if we can bound the probability that a random sample S does not form an ϵ -net for $\Delta_{C_0, \epsilon}(\mathcal{C})$, then we have bounded the probability that for a hypothesis C_0 consistent with S we have $\hat{R}(C_0) > \epsilon$.

Example

Suppose that \mathcal{C} is finite. For any fixed set $C_0 \oplus C \in \Delta_{C_0, \epsilon}(\mathcal{C})$, the probability that we fail to hit $C_0 \oplus C$ in m random examples is at most $(1 - \epsilon)^m$. Thus, the probability that we fail to hit some $C_0 \oplus C \in \Delta_{C_0, \epsilon}(\mathcal{C})$ is bounded above by $|\mathcal{C}|(1 - \epsilon)^m$.

The Double Sample Theorem

Theorem

Let \mathcal{C} be a concept class with $VCD(\mathcal{C}) = d$.

Let \mathcal{A} be any algorithm that given a set S of m labeled examples $\{(x_i, c(x_i)) \mid 1 \leq i \leq m\}$ sampled iid according to some fixed but unknown distribution \mathcal{D} over the instance space \mathcal{X} produces as output a hypothesis h that is consistent with c . Then, \mathcal{A} is a PAC algorithm and

$$m \geq k_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right).$$

for some positive constant k_0 .

Proof

- Draw a sample S_1 of size m from \mathcal{D} and let A be the event that the elements of S_1 fail to form an ϵ -net for $\Delta_{C_0, \epsilon}(\mathcal{C})$.
- If A occurs, then S_1 misses some region T , where

$$T \in \Delta_{C_0, \epsilon}(\mathcal{C}).$$

Fix this region T and draw an additional sample S_2 of size m from \mathcal{D} .

Let V be a binomial random variable that gives the number of hits of T by the sample S_2 . We have $E(V) = m\epsilon$ and $\text{var}(V) = m\epsilon(1 - \epsilon)$ because the probability of an element of S_2 hitting T is ϵ .

By Chebyshev's Inequality applied to V we have

$$P(|V - m\epsilon| \geq a) \leq \frac{m\epsilon(1 - \epsilon)}{a^2}.$$

Taking $a = \frac{\epsilon m}{2}$ it follows that

$$\begin{aligned} P(|V - m\epsilon| \geq \frac{\epsilon m}{2}) &\leq \frac{4(1 - \epsilon)}{\epsilon m} \\ &\leq \frac{4}{\epsilon m} \leq \frac{1}{2}, \end{aligned}$$

provided that $m \geq \frac{8}{\epsilon}$.

Thus, if $m \geq \frac{8}{\epsilon}$,

$$P\left(|V - \epsilon m| \leq \frac{\epsilon m}{2}\right) \geq \frac{1}{2}.$$

The inequality

$$|V - \epsilon m| \leq \frac{\epsilon m}{2}$$

is equivalent to $\frac{\epsilon m}{2} \leq V \leq \frac{3\epsilon m}{2}$, which implies $P(V \geq \frac{\epsilon m}{2}) \geq \frac{1}{2}$.

Proof (cont'd)

To summarize: we have calculated the probability that S_2 will hit T many times **given** that T was fixed using the previous sampling, that is, given that S_1 does not form an ϵ -net.

Let B be the event that

S_1 does not form an ϵ -net and that S_2 hits T at least $\frac{\epsilon m}{2}$ times.

Then, we have shown that for $m = O(1/\epsilon)$ we have $P(B|A) \geq \frac{1}{2}$.

Proof (cont'd)

Since $P(B|A) \geq \frac{1}{2}$ we have

$$P(B) = P(B|A)P(A) \geq \frac{1}{2}P(A).$$

Our goal of bounding $P(A)$ is equivalent to finding δ such that $P(B) \leq \frac{\delta}{2}$ because this would imply $P(A) \leq \delta$.

Proof (cont'd)

Let $S = S_1 \cup S_2$ be a random sample of $2m$. Note that since the samples are iid obtaining S is equivalent of sampling S_1 and S_2 separately and let T be a fixed set such that $|T| \geq \frac{\epsilon m}{2}$.

Consider a random partition of S into S_1 and S_2 and consider the probability that $S_1 \cap T = \emptyset$.

An Equivalent Problem: we have $2m$ balls each colored red or blue with exactly ℓ red balls, where $\ell \geq \frac{\epsilon m}{2}$. Divide the $2m$ balls into groups of equal size m , namely S_1 and S_2 . Find an upper bound on the probability that all ℓ red balls fall in S_2 (that is, the probability that $S_1 \cap R = \emptyset$).

Proof (cont'd)

Yet Another Equivalent Problem: Divide $2m$ non-colored balls into S_1 and S_2 , choose ℓ to be colored red, and compute the probability that all red balls fall in S_2 . The probability of this taking place is:

$$\frac{\binom{m}{\ell}}{\binom{2m}{\ell}}$$

Note that

$$\frac{\binom{m}{\ell}}{\binom{2m}{\ell}} = \prod_{i=0}^{\ell-1} \frac{m-i}{2m-i} \leq \prod_{i=0}^{\ell-1} \frac{1}{2} = \frac{1}{2^\ell} = 2^{-\frac{\ell m}{2}}.$$

This is the probability for a fixed S and T . The probability that this occurs for some $T \in \Delta_{C_0, \epsilon}(S)$ such that $|T| \geq \frac{\epsilon m}{2}$ can be computed by summing over all T and applying the union bound:

$$\begin{aligned} P(B) &\leq |\Pi_{\Delta_{C_0, \epsilon}(S)}| \left(\frac{\epsilon m}{2}\right) 2^{-\frac{\epsilon m}{2}} \leq |\Pi_{\Delta_{C_0}(S)}| \left(\frac{\epsilon m}{2}\right) 2^{-\frac{\epsilon m}{2}} \\ &\leq \left(\frac{2\epsilon m}{d}\right)^d 2^{-\frac{\epsilon m}{2}} \leq \frac{\delta}{2}. \end{aligned}$$

Proof (cont'd)

The last inequality implies

$$m \geq k_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right).$$

for some positive constant k_0 .

Let $u : B_2^k \rightarrow B_2$ be a Boolean function of k arguments and let C_1, \dots, C_k be k subsets of a set U . Define the set $u(C_1, \dots, C_k)$ as the subset C of U whose indicator function is $I_C = u(I_{C_1}, \dots, I_{C_k})$.

Example

If $u : B_2^2 \rightarrow B_2$ is the Boolean function $u(a_1, a_2) = a_1 \vee a_2$, then $u(C_1, C_2)$ is $C_1 \cup C_2$; similarly, if $u(x_1, x_2) = x_1 \oplus x_2$, then $u(C_1, C_2)$ is the symmetric difference $C_1 \oplus C_2$ for every $C_1, C_2 \in \mathcal{P}(U)$.

Let $u : B_2^k \rightarrow B_2$ and $\mathcal{C}_1, \dots, \mathcal{C}_k$ are k family of subsets of U , the family of sets $u(\mathcal{C}_1, \dots, \mathcal{C}_k)$ is

$$u(\mathcal{C}_1, \dots, \mathcal{C}_k) = \{u(C_1, \dots, C_k) \mid C_1 \in \mathcal{C}_1, \dots, C_k \in \mathcal{C}_k\}.$$

Theorem

Let $\alpha(k)$ be the least integer a such that $\frac{a}{\log(ea)} > k$.
 If $\mathcal{C}_1, \dots, \mathcal{C}_k$ are k collections of subsets of the set U such that
 $d = \max\{VCD(\mathcal{C}_i) \mid 1 \leq i \leq k\}$ and $u : B_2^k \rightarrow B_2$ is a Boolean function,
 then

$$VCD(u(\mathcal{C}_1, \dots, \mathcal{C}_k)) \leq \alpha(k) \cdot d.$$

Proof

Let S be a subset of U that consists of m elements. The collection $(\mathcal{C}_i)_S$ is not larger than $\phi(d, m)$. For a set in the collection $W \in u(\mathcal{C}_1, \dots, \mathcal{C}_k)_S$ we can write $W = S \cap u(\mathcal{C}_1, \dots, \mathcal{C}_k)$, or, equivalently,

$$1_W = 1_S \cdot u(1_{\mathcal{C}_1}, \dots, 1_{\mathcal{C}_k}).$$

There exists a Boolean function g_S such that

$$1_S \cdot u(1_{\mathcal{C}_1}, \dots, 1_{\mathcal{C}_k}) = g_S(1_S \cdot 1_{\mathcal{C}_1}, \dots, 1_S \cdot 1_{\mathcal{C}_k}) = g_S(1_{S \cap \mathcal{C}_1}, \dots, 1_{S \cap \mathcal{C}_k}).$$

Since there are at most $\phi(d, m)$ distinct sets of the form $S \cap \mathcal{C}_i$ for every i , $1 \leq i \leq k$, it follows that there are at most $(\phi(d, m))^k$ distinct sets W , hence $u(\mathcal{C}_1, \dots, \mathcal{C}_k)[m] \leq (\phi(d, m))^k$.

Proof (cont'd)

By a previous theorem,

$$u(\mathcal{C}_1, \dots, \mathcal{C}_k)[m] \leq \left(\frac{em}{d}\right)^{kd}.$$

We observed that if $\Pi_{\mathcal{C}}[m] < 2^m$, then $VCD(\mathcal{C}) < m$. Therefore, to limit the Vapnik-Chervonenkis dimension of the collection $u(\mathcal{C}_1, \dots, \mathcal{C}_k)$ it suffices to require that $\left(\frac{em}{d}\right)^{kd} < 2^m$.

Let $a = \frac{m}{d}$. The last inequality can be written as $(ea)^{kd} < 2^{ad}$; equivalently, we have $(ea)^k < 2^a$, which yields $k < \frac{a}{\log(ea)}$. If $\alpha(k)$ is the least integer a such that $k < \frac{a}{\log(ea)}$, then $m \leq \alpha(k)d$, which gives our conclusion.

Example

If $k = 2$, the least integer a such that $\frac{a}{\log(ea)} > 2$ is $k = 10$, as it can be seen by graphing this function; thus, if $\mathcal{C}_1, \mathcal{C}_2$ are two collection of concepts with $VCD(\mathcal{C}_1) = VCD(\mathcal{C}_2) = d$, the Vapnik-Chervonenkis dimension of the collections $\mathcal{C}_1 \vee \mathcal{C}_2$ or $\mathcal{C}_1 \wedge \mathcal{C}_2$ is not larger than $10d$.

Lemma

Let S, T be two sets and let $f : S \rightarrow T$ be a function. If \mathcal{D} is a collection of subsets of T , U is a finite subset of S and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $|\mathcal{C}_U| \leq |\mathcal{D}_{f(U)}|$.

Proof: Let $V = f(U)$ and denote $f \upharpoonright_U$ by g . For $D, D' \in \mathcal{D}$ we have

$$\begin{aligned} & (U \cap f^{-1}(D)) \oplus (U \cap f^{-1}(D')) \\ &= U \cap (f^{-1}(D) \oplus f^{-1}(D')) = U \cap (f^{-1}(D \oplus D')) \\ &= g^{-1}(V \cap (D \oplus D')) = g^{-1}(V \cap D) \oplus g^{-1}(V \cap D'). \end{aligned}$$

Thus, $C = U \cap f^{-1}(D)$ and $C' = U \cap f^{-1}(D')$ are two distinct members of \mathcal{C}_U , then $V \cap D$ and $V \cap D'$ are two distinct members of $\mathcal{D}_{f(U)}$. This implies $|\mathcal{C}_U| \leq |\mathcal{D}_{f(U)}|$.

Theorem

Let S, T be two sets and let $f : S \longrightarrow T$ be a function. If \mathcal{D} is a collection of subsets of T and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $VCD(\mathcal{C}) \leq VCD(\mathcal{D})$. Moreover, if f is a surjection, then $VCD(\mathcal{C}) = VCD(\mathcal{D})$.

Proof

Suppose that \mathcal{C} shatters an n -element subset $K = \{x_1, \dots, x_n\}$ of S , so $|\mathcal{C}_K| = 2^n$. By a previous Lemma we have $|\mathcal{C}_K| \leq |\mathcal{D}_{f(U)}|$, so $|\mathcal{D}_{f(U)}| \geq 2^n$, which implies $|f(U)| = n$ and $|\mathcal{D}_{f(U)}| = 2^n$, because $f(U)$ cannot have more than n elements. Thus, \mathcal{D} shatters $f(U)$, so $VCD(\mathcal{C}) \leq VCD(\mathcal{D})$.

Suppose now that f is surjective and $H = \{t_1, \dots, t_m\}$ is an m element set that is shattered by \mathcal{D} . Consider the set $L = \{u_1, \dots, u_m\}$ such that $u_i \in f^{-1}(t_i)$ for $1 \leq i \leq m$. Let U be a subset of L . Since H is shattered by \mathcal{D} , there is a set $D \in \mathcal{D}$ such that $f(U) = H \cap D$, which implies $U = L \cap f^{-1}(D)$. Thus, L is shattered by \mathcal{C} and this means that $VCD(\mathcal{C}) = VCD(\mathcal{D})$.

Definition

The *density* of \mathcal{C} is the number

$$\text{dens}(\mathcal{C}) = \inf\{s \in \mathbb{R}_{>0} \mid \Pi_{\mathcal{C}}[m] \leq c \cdot m^s \text{ for every } m \in \mathbb{N}\},$$

for some positive constant c .

Theorem

Let S, T be two sets and let $f : S \longrightarrow T$ be a function. If \mathcal{D} is a collection of subsets of T and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $\text{dens}(\mathcal{C}) \leq \text{dens}(\mathcal{D})$. Moreover, if f is a surjection, then $\text{dens}(\mathcal{C}) = \text{dens}(\mathcal{D})$.

Proof: Let L be a subset of S such that $|L| = m$. Then, $|\mathcal{C}_L| \leq |\mathcal{D}_{f(L)}|$. In general, we have $|f(L)| \leq m$, so $|\mathcal{D}_{f(L)}| \leq \mathcal{D}[m] \leq cm^s$. Therefore, we have $|\mathcal{C}_L| \leq |\mathcal{D}_{f(L)}| \leq \mathcal{D}[m] \leq cm^s$, which implies $\text{dens}(\mathcal{C}) \leq \text{dens}(\mathcal{D})$. If f is a surjection, then, for every finite subset M of T such that $|M| = m$ there is a subset L of S such that $|L| = |M|$ and $f(L) = M$. Therefore, $\mathcal{D}[m] \leq \Pi_{\mathcal{C}}[m]$ and this implies $\text{dens}(\mathcal{C}) = \text{dens}(\mathcal{D})$.

If \mathcal{C}, \mathcal{D} are two collections of sets such that $\mathcal{C} \subseteq \mathcal{D}$, then $VCD(\mathcal{C}) \leq VCD(\mathcal{D})$ and $\text{dens}(\mathcal{C}) \leq \text{dens}(\mathcal{D})$.

Theorem

Let \mathcal{C} be a collection of subsets of a set S and let $\mathcal{C}' = \{S - C \mid C \in \mathcal{C}\}$. Then, for every $K \in \mathcal{P}(S)$ we have $|\mathcal{C}_K| = |\mathcal{C}'_K|$.

Proof

We prove the statement by showing the existence of a bijection $f : \mathcal{C}_K \rightarrow \mathcal{C}'_K$. If $U \in \mathcal{C}_K$, then $U = K \cap C$, where $C \in \mathcal{C}$. Then $S - C \in \mathcal{C}'$ and we define $f(U) = K \cap (S - C) = K - C \in \mathcal{C}'_K$. The function f is well-defined because if $K \cap C_1 = K \cap C_2$, then

$$K - C_1 = K - (K \cap C_1) = K - (K \cap C_2) = K - C_2.$$

It is clear that if $f(U) = f(V)$ for $U, V \in \mathcal{C}_K$, $U = K \cap C_1$, and $V = K \cap C_2$, then $K - C_1 = K - C_2$, so $K \cap C_1 = K \cap C_2$ and this means that $U = V$. Thus, f is injective. If $W \in \mathcal{C}'_K$, then $W = K \cap C'$ for some $C' \in \mathcal{C}$. Since $C' = S - C$ for some $C \in \mathcal{C}$, it follows that $W = K - C$, so $W = f(U)$, where $U = K \cap C$.

Corollary

Let \mathcal{C} be a collection of subsets of a set S and let $\mathcal{C}' = \{S - C \mid C \in \mathcal{C}\}$. We have $\text{dens}(\mathcal{C}) = \text{dens}(\mathcal{C}')$ and $\text{VCD}(\mathcal{C}) = \text{VCD}(\mathcal{C}')$.

Theorem

For every collection of sets we have $\text{dens}(\mathcal{C}) \leq \text{VCD}(\mathcal{C})$. Furthermore, if $\text{dens}(\mathcal{C})$ is finite, then \mathcal{C} is a VC-class.

Proof: If \mathcal{C} is not a VC-class the inequality $\text{dens}(\mathcal{C}) \leq \text{VCD}(\mathcal{C})$ is clearly satisfied. Suppose now that \mathcal{C} is a VC-class and $\text{VCD}(\mathcal{C}) = d$. By Sauer-Shelah Theorem we have $\Pi_{\mathcal{C}}[m] \leq \phi(d, m)$; then, we obtain $\Pi_{\mathcal{C}}[m] \leq \left(\frac{em}{d}\right)^d$, so $\text{dens}(\mathcal{C}) \leq d$.

Suppose now that $\text{dens}(\mathcal{C})$ is finite. Since $\Pi_{\mathcal{C}}[m] \leq cm^s \leq 2^m$ for m sufficiently large, it follows that $\text{VCD}(\mathcal{C})$ is finite, so \mathcal{C} is a VC-class.

Let \mathcal{D} be a finite collection of subsets of a set S . The partition $\pi_{\mathcal{D}}$ was defined as consisting of the nonempty sets of the form $\{D_1^{a_1} \cap D_2^{a_2} \cap \cdots \cap D_r^{a_r}, \text{ where } (a_1, a_2, \dots, a_r) \in \{0, 1\}^r\}$.

Definition

A collection $\mathcal{D} = \{D_1, \dots, D_r\}$ of subsets of a set S is *independent* if the partition $\pi_{\mathcal{D}}$ has the maximum numbers of blocks, that is, it consists of 2^r blocks.

If \mathcal{D} is independent, then the Boolean subalgebra generated by \mathcal{D} in the Boolean algebra $(\mathcal{P}(S), \{\cap, \cup, ^-, \emptyset, S\})$ contains 2^{2^r} sets, because this subalgebra has 2^r atoms. Thus, if \mathcal{D} shatters a subset T with $|T| = p$, then the collection \mathcal{D}_T contains 2^p sets, which implies $2^p \leq 2^{2^r}$, or $p \leq 2^r$.

Definition

Let \mathcal{C} be a collection of subsets of a set S . The **independence number of \mathcal{C}** $I(\mathcal{C})$ is:

$$I(\mathcal{C}) = \sup\{r \mid \{C_1, \dots, C_r\} \\ \text{is independent for some finite } \{C_1, \dots, C_r\} \subseteq \mathcal{C}\}.$$

Theorem

Let S, T be two sets and let $f : S \longrightarrow T$ be a function. If \mathcal{D} is a collection of subsets of T and $\mathcal{C} = f^{-1}(\mathcal{D})$ is the collection $\{f^{-1}(D) \mid D \in \mathcal{D}\}$, then $I(\mathcal{C}) \leq I(\mathcal{D})$. Moreover, if f is a surjection, then $I(\mathcal{C}) = I(\mathcal{D})$.

Proof: Let $\mathcal{E} = \{D_1, \dots, D_p\}$ be an independent finite subcollection of \mathcal{D} . The partition $\pi_{\mathcal{E}}$ contains 2^r blocks. The number of atoms of the subalgebra generated by $\{f^{-1}(D_1), \dots, f^{-1}(D_p)\}$ is not greater than 2^r . Therefore, $I(\mathcal{C}) \leq I(\mathcal{D})$; from the same supplement it follows that if f is surjective, then $I(\mathcal{C}) = I(\mathcal{D})$.

Theorem

If \mathcal{C} is a collection of subsets of a set S such that $VCD(\mathcal{C}) \geq 2^n$, then $I(\mathcal{C}) \geq n$.

Proof: Suppose that $VCD(\mathcal{C}) \geq 2^n$, that is, there exists a subset T of S that is shattered by \mathcal{C} and has at least 2^n elements. Then, the collection \mathcal{C}_T contains at least 2^{2^n} sets, which means that the Boolean subalgebra of $\mathcal{P}(T)$ generated by \mathcal{C}_T contains at least 2^n atoms. This implies that the subalgebra of $\mathcal{P}(S)$ generated by \mathcal{C} contains at least this number of atoms, so $I(\mathcal{C}) \geq n$.