Teaching Complexity

Johar M. Ashfaque

1 The Terminology

Let $X = \Sigma^*$ be the set of all strings on the binary alphabet $\Sigma = \{0, 1\}$. X_n denotes the set of all strings of length n or less for $n \geq 0$. A target concept c is a subset of X. Viewed in an other way, a target concept is a function $c : \Sigma^* \to \{0, 1\}$ where $c(x) = 1$ implies x is in the concept and $c(x) = 0$ implies otherwise. A concept class is a non-empty set $F \subseteq 2^X$ of concepts. For a concept $c \in F$ and an integer $n \geq 0$ we define the n-th subclass of F by $F_n = \{c \cap X_n | c \in F\}$. An example on $x \in X$ for a concept c is a pair $\langle x, c(x)\rangle$.

Let F be a concept class. Let S_F be the sample space of F. A mapping $A : S_F \to F$ is called a learn mapping for F. A_F denotes the set of all learn mappings for F. For $A \in A_F$ and $c \in F$, a hypothesis on the sample space of F by A is the concept $h = A(sam_c(\overline{x}))$ where:

$$\overline{x} = (x_1, ..., x_m) \in X^m \tag{1}$$

and:

$$sam_c(\overline{x}) = (\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle) \tag{2}$$

is the sample of size m on $\overline{x}$. In both learning and teaching models, the aim is to identify either exactly or approximately a target concept from the concept class.

2 Vapnik-Chervonenkis dimension

Definition. The goal of the learning algorithm is to learn a good approximation of the target concept, with high probability. This is called *probably approximately correct* learning or *PAC learning*. Concept class F is called learnable if there exists a PAC learning algorithm such that, for any accuracy and confidence parameters of that learning algorithm, there exists a fixed sample size such that, for any concept $c \in F$ and for any probability distribution on domain X, the learning algorithm produces a probably approximately correct hypothesis.

Definition. Consider a function $M(F)$ that assigns a complexity to each concept class F. If $F^{'} \subset F$ implies that $M(F^{'}) \leq M(F)$ we say M is *monotonic*.
Definition. Consider S a subset of X.The projection of F to a subset S is defined as:

$$F|_S = \{c \cap S : c \in F\} \subseteq 2^S \tag{3}$$

if $F|_S = 2^{|S|}$, concept class F *shatters* a subset S.
$S \subseteq X$ is shattered by F if for each subset $S^{'} \subseteq S$ there is a concept $c \in F$ which contains all of $S^{'}$, but none of instances in $S - S^{'}$.

The Vapnik-Chervonenkis dimension of F or $VCD(F)$ is defined to be the maximum size of any subset that can be shattered by F.

$$VCD(F) = \max_{S \subseteq X : |F|_S| = 2^{|S|}} |S| \tag{4}$$

VCD is monotonic.

3 Teaching Dimension

The *teaching set* for c is a set of labeled examples that can distinguish the concept c from other concepts in F. Let $TF(c;F)$ be the family of all teaching sets for concept c; Then *Teaching dimension* of a concept c in the concept class F is defined as:

$$TD(c;F) = \min_{S \in TF(c;F)} |S| \tag{5}$$

Teaching dimension of a concept class F or $TD(F)$ is the minimum number of examples a teacher must reveal to uniquely identify every concept in the class and is defined as:

$$TD(F) = \max_{c \in F} TD(c;F) \tag{6}$$

Also the best case teaching dimension of F is defined as:

$$TD_{min}(F) = \min_{c \in F} TD(c;F) \tag{7}$$

Note that TD is monotonic but TD_{min} is not monotonic.

3.1 VCD vs. TD

Teaching dimension is fundamentally different from the VC dimension. In the following you can find neither of these dimension measures dominates the other. in some cases $TD(F) \ll VCD(F)$ and in other cases $TD(F) \gg VCD(F)$.

$TD(F) \gg VCD(F)$

There is a concept class F that $TD(F) = |F| - 1$ and $VCD(F) = 1$. This concept class has the largest possible teaching dimension. So for every concept class F we have

$$TD(F) \leq |F| - 1 \tag{8}$$

3.2 $TD(F) \ll VCD(F)$

There is a concept class F that $TD(F) < VCD(F)$.

3.3 Elimination of a concept from F

Another key difference between VCD and TD is the effect of removing a concept from the concept class. Let $VCD(F) = d$ and $F^{'} = F - c$, where c is some random concept in F. Then we have $VCD(F^{'}) \geq d-1$. But in contrast for the concept class F which $TD(F) \geq |F| - k$, there is a concept c that if we remove it from F, the TD of $F^{'} = F - c$ would be $TD(F^{'}) \leq k$.

Definition: Consider a matrix M such that the rows correspond to the hypothesis and the columns correspond to the set of points in the domain. *Dual Class* is the set of hypothesis represented by the transpose of the matrix M. Dehghani claimed that for a given concept class F with $VCD(F) = d$, the dual class has a VC dimension at most $2^{d+1} - 1$.

4 Recursive Teaching Dimension

Recursive teaching dimension (RTD) is another complexity measure. RTD is constructed based on a hierarchy of best case teaching dimension. Consider the concept class F. Then F_1 is defined as $F = F_1$. The subset of all concepts achieving the best case teaching dimension is ${F_1}^{min}$:

$$ {F_1}^{min} = c \in F_1 : TD(c; F_1) = TD_{min}(F_1) \tag{9} $$

By removing the subset ${F_1}^{min}$ from F_1 and repeating the process on $F_2 = F - {F_1}^{min}$ we get ${F_2}^{min}$. Then RTD is defined as the maximum of the best teaching dimensions over the all subsets F_i, $i = 1, ..., t$ where $F_{t+1} = \emptyset$ and we have:

$$ F = F_1 \supset F_2 \supset ... \supset F_t \supset F_{t+1} \tag{10} $$

Formally we have:

$$ RTD(F) = \max_{1 \leq i \leq t} TD_{min}(F_i) \tag{11} $$

Also there is another definition for RTD in maxmin form:

$$ RTD(F) = \max_{F' \subset F} \min_{c \in F'} TD(c, F') \tag{12} $$

4.1 Properties

1. RTD is bounded by TD_{min} and TD:

$$ TD_{min}(F) \leq RTD(F) \leq TD(F) \tag{13} $$

   The left of the inequality is trivial. So TD_{min} and TD are lower and upper bounds of RTD. Since it is NP-hard to compute the optimal teaching sequence problem, the upper bound is not easy to compute. But finding TD_{min} is more convenient. If P is monotonic and $TS_{min}(F) \leq P(F)$ for every concept class F, then $RTD(F) \leq P(F)$ for every concept class F. Then $RTD(F) \leq k$ follows directly from $TD_{min} \leq k$.

2. RTD is monotonic.

4.2 RTD vs. TD and VCD

Suppose $TD_{min}(F) \leq k \cdot VCD(F)$ for all F. Using (12) we have $RTD(F) = TD_{min}(F')$ for some of the subclasses F' of F. By the monotonicity of VCD we have $RTD = O(VCD)$, since

$$ RTD(F) = TD_{min}(F') \leq k \cdot VCD(F') \leq k \cdot VCD(F) \tag{14} $$

In many general cases, RTD is upper bounded by VCD, e.g. classes of $VCD = 1$, intersection-closed classes, finite maximum classes and standard Boolean functions. In the following we will provide some insights into the correspondence between RTD and VCD in these particular classes.

4.3 RTD of Intersection-closed classes

We call a concept class F an intersection-closed class if for all $c, c' \in F$ we have $c \cap c' \in F$. Some of the examples of these classes are d-dimensional boxes, monomials and and vector spaces in R^n. For any subset S of the domain X we define the *closure* of S with respect to F as

$$ \cap \{c : c \in F \quad and \quad S \subset c\} \tag{15} $$

A *spanning set* s for a subset c of concept class F, is a subset of c that the closure of c is equal to the closure of s w.r.t. F. c is the minimal spanning set if none of the subsets of c has the same closure as

c. The minimal spanning set can be used as a compression set for the sample. Let $I(F)$ denotes the size of the largest minimal spanning set w.r.t. F. We have $I(F) \leq VCD(F)$. So the size of the sample compression scheme for an intersection-closed class is at most the VCD of the class. Consider a subset of concept class F as $F^{'}$. Each spanning set for $c \subset F^{'}$ w.r.t. F is also a spanning set for c w.r.t. $F|_{F'}$ (projection of F to $F^{'}$).
There is another class called *nested differences*, which is defined on intersection-closed classes. Consider $c_1, c_2, ..., c_t \in F$, the nested difference class has the form:

$$c_1 - (c_2 - (c_3 - (... - c_t)...)) \tag{16}$$

The depth of this class is t.

4.4 Properties

1. $TS_{min}(F) \leq I(F)$
2. $RTD(F) \leq I(F)$.
Proof: Suppose that we have a teaching plan for F in topological order. Let the size of the minimal spanning set of the largest concept c be k. Then the only concept that contains this minimal spanning is c and $k \leq I(F)$. These hierarchy is similar to the ones we defined for RTD. So we obtain $RTD(F) \leq I(F)$.
3. Since $I(F) \leq VCD(F)$, we have $RTD(F) \leq VCD(F)$
4. RTD of a nested difference class with concepts from F and depth d is upper bounded by $d \cdot I(F)$.
Proof: Assume a lexicographic ordering on concepts from nested difference class of F. Let c be the largest concept and the stater concept of the nested difference class, which if any other concept was the stater, it would give us a smaller size nested difference class. So by discarding all of the concepts preceding c, we obtain a teaching set for c of size $d \cdot I(F)$. So we have $RTD(nesteddifferenceclassofF) \leq d \cdot I(F)$.

4.5 RTD of Maximum classes

A concept class is called *maximal* if adding any concept to the class increase the VCD of that class. A concept class F of $VCD = d$ on X is called *maximum* if for every subset Y of X, $F|_Y$ contains $\Phi_a(|Y|)$ concepts on Y where:

$$\Phi_d(m) = \begin{cases} \sum_{i=0}^{d} \binom{m}{i} & m \geq d \\ 2^m & m < d \end{cases} \tag{17}$$

A concept class that is maximum on a finite domain X is also maximal on that set. The upper bound coincides with *Sauers* well-known bound on classes with a fixed VC-dimension.
If $VCD(F) = d$ and F is maximum then:

- every subset of X of size $\leq d$ is shattered by F.
- F shatters exactly $|F|$ many subsets of X.

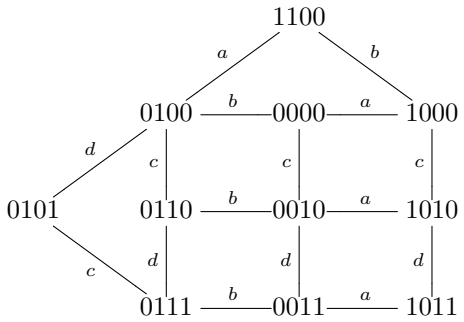
Definition. *One − inclusion graph* is an undirected graph which characterizes the concept class on a set of example points. The vertices are possible labelings of the example points and there is an edge between two concepts if they disagree on a single point. The edges are naturally labeled by the differing points.
Example. We give an example below where $VCD = 2$, $n = 4$ and $|F| = 11$.

a	b	c	d
0	0	0	0
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	0
1	0	1	1
1	1	0	0

It can be seen that the class F is maximum since

$$\sum_{i=0}^{2}\binom{4}{i}=\binom{4}{0}+\binom{4}{1}+\binom{4}{2}=1+4+6=11$$

The one-inclusion graph $G(F)$ is given below:



Example. The following example is a maximal class, since $VCD = 2$, $n = 4$ and $|F| = 10$.

a	b	c	d
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	0
1	0	1	1
1	1	0	0

If we add another concept $< 0, 0, 0, 0 >$ to the concept class, VCD would be 3. So this concept class is a maximal class.

If F is maximum then $RTD(F) = VCD(F)$. In a maximum class, every set of $k \leq VCD(F)$ instances is shattered, which implies $RTD(F) \geq TS_{min}(F) \geq VCD(F)$. Thus we need to focus on the reverse direction where $RTD(F) \leq VCD(F)$.
There is another way to define RTD using *teaching plan*. A teaching plan for F is a sequence:

$$P = ((c_1, s_1), ..., (c_n, s_n)) \tag{18}$$

Which has 2 properties:
1. $n = |F|$ *and* $F = \{c_1, ..., c_n\}$
2. *for all* $t = 1, ..., n,\ s_t \in TD(c_t, \{c_t, ..., c_n\})$
So we can use the order of teaching to define maximum of s_t for each t and this is the ordering we need to define RTD as mentioned before.
Also we define *corner peeling plan* for F as:

$$P = ((c_1, F_1^{'}), ..., (c_n, F_n^{'})) \tag{19}$$

Which has 2 properties:
1. $n = |F|$ and $F = \{c_1, ..., c_n\}$
2. For all $t = 1, ..., n,$, $F_t^{'}$ is a cube in $\{c_t, ..., c_n\}$ which contains c_t and all its neighbors in $G(\{c_t, ..., c_n\})$.

Here $G(c)$ is one-inclusion-graph, which the nodes are the concepts from F and each pair of nodes

are connected if they differ in exactly one coordinate.
If we change the second condition of corner peeling, we get the definition of teaching plan for RTD. the replacement of the second condition is the following:
$2^{'}$. for all $t = 1, ..., n$, $F_t^{'}$ is a cube in $\{c_t, ..., c_n\}$ which contains c_t and whose instances form a teaching set for $\{c_t, ..., c_n\}$.
Then the corner peeling plan is called *strong*. Now we show that every strong corner peeling plan is a corner peeling plan:
Consider the condition 2 is violated. Assume that there is an instance $x \in X \backslash X_t$ and a concept $c \in \{c_t, ..., c_n\}$ which c is not consistent with x. Then s_t is not a teaching set for c_t. As you see the condition $2^{'}$ is violated too.

Definition. The size of the largest cube is the dimension of corner peeling plan.

Definition. F is called *shortest path closed* if for every pair of the concepts $c, c^{'} \in F$, the $G(F)$ contains a path of length $H(c, c^{'})$ (Hamming distance between c and $c^{'}$).

Note. Every strong corner peeling plan is a corner peeling plan.

Note. If F is shortest path closed, every corner peeling plan is strong.

As a result it obtains that every maximum class is a VC-dimension corner peeling plan. If F is a shortest path closed concept class, it will have corner peeling plan. As every strong corner peeling plan is a corner peeling plan, then every corner peeling plan for F is strong. Maximum classes are shortest path closed. Then a maximum class has a teaching plan of the same order and we have $RTD(F) \leq VCD(F)$ (since a maximum class is VC-dimension corner peeling plan).
Then we have $RTD = VCD$ for a maximum class.

Suppose $RTD(F) \geq 2$. Then we have $RTD(F) \geq \log|F| / \log|X|$.
Proof: We solve it using Sauer's bound for VCD. Let $RTD(F) = k$ then the following holds:

$$|F| \leq \sum_{i=1}^{k} \binom{|X|}{i} = \phi_k(|X|) \leq |X|^k \tag{20}$$

Using logarithmic form of the left and right sides of the inequality, we have $RTD(F) \geq \log|F| / \log|X|$.

Conjecture: Given a class of concepts F whose teaching dimension is t, it is impossible to add a new concept $c^{'} \notin F$ such that the teaching dimension of the new concept class is $t + 2$ or greater.

This conjecture is interesting because if it's true, this says that adding concept by concept one cannot increase the teaching teaching dimension by a lot. On the other hand, adding in more concepts increases the VC dimension at least logarithmically. This will potentially give us a relation between the two, independent of the size of concept class.
Dehghani attempted to find an concept class such that the RTD is much bigger than the VCD. They wrote a computer code to enumerate the concept classes. For small values of $m \leq 30$ and $n \leq 8$, where m is the number of concepts and n is the number of examples, they generated thousands of random matrices and they were unable to find a single example where RTD exceeded VCD more than 2. This gives and evidence that, for most matrices the RTD is very close to the VCD exactly as conjectured. Currently they are writing a code to systematically enumerate all examples to see if there is a special example where it might break.

5 Self-directed learning

In studying the problem of learning binary relations, Goldman, Rivest, and Schapire in 1993[1] introduced the model of self-directed learning in which the learn selects the order in which the instances are presented

[1]Learning binary relations and total orders

to the learner. The model is defined as follows. The definitions only apply to finite instance spaces with countable sequences. We define a query sequence to be a permutation :

$$\pi = \langle x_1, x_2, ..., x_{|X|} \rangle \tag{21}$$

of the instance space X where x_t is the instance the learner will predict at the t^{th} trial.

Note: By the definition of a query sequence, x_t must be an instance that the learner has not yet considered.

Furthermore, if after the completion of the t^{th} trial, the learner knows with certainty the classification of all instances from X that have not been queried (i.e. $x_{t+1}, ..., x_{|X|}$) then the learning sessions is complete. Self-directed complexity of concept class F denoted as $SDC(F)$ is the smallest number q such that there is some self-directed learning algorithm which can exactly identify any concept $c^* \in F$ without making more than q mistakes. The *optimal mistake bound* or M_opt for the self-directed learning of concept class F is the minimum of the mistake bound over all self-directed learning algorithms for F.

5.1 The chain of inequalities

Then we have the following chain of inequalities:

$$RTD(F) \leq SDC(F) \leq LCP(F) \leq M_{opt}(F) \leq \log |F| \tag{22}$$

Proof: We first prove that $TD_{min}(F) \leq SDC(F)$. Consider z to be the sequence that learner presents in the self-directed learning. Let $\{x_1, x2, ..., x_k\} \subset z$ be the instances on which learner made a mistake and $\{y_1, y2, ..., y_l\} \subset z$ those for which the adversary was not able to reveal another label but the one issued from the learner. Then $x_1, x2, ..., x_k | c_t$ already uniquely identifies the target concept c_t. So if we remove y_i from the sequence, we still can identify c_t from all concepts in F. Therefore y_i cannot gain more information and $TD_{min}(F) \leq SDC(F)$. By monotonicity of SDC we have

$$RTD(F) \leq SDC(F) \tag{23}$$

$SDC(F) \leq M_{opt}(F)$ is obvious regarding the definition.

6 Sample Compression

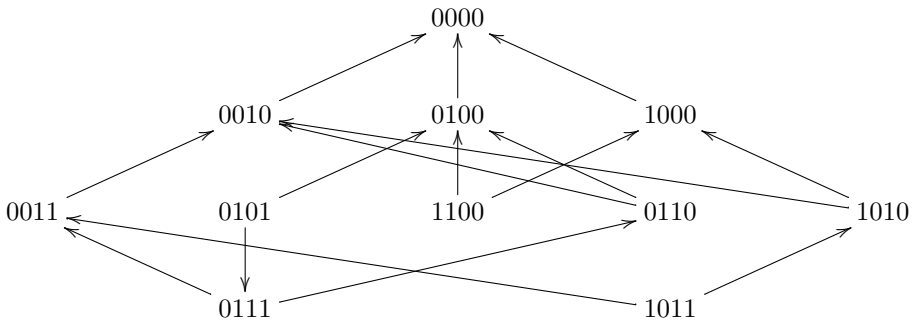
A sample compression scheme takes a long list of samples and compresses it to a short sub-list of samples in a way that allows to invert the compression. In other words, it is a scheme for encoding a set of examples in a small subset of examples. A sample compression scheme of size at most k for a concept class F on X consists of a *compression function* and a *reconstruction function*. The compression function maps every finite sample set to a compression set, a subset of at most k labeled examples. The reconstruction function maps every possible compression set to a hypothesis. In other words, the compression function f maps every F-realizable sample S to a subset of size at most k, called a compression set. The reconstruction function g maps any compression set $f(S)$ to a hypothesis $g(f(S)) \subseteq X$ which is not required to be in the concept class of F but is consistent with the original sample set S. For example consider the class of rectangles in R^2. each concept corresponds to a rectangle.The points within the rectangles have positive labels and the ones outside the rectangles have negative labels. The compression function can save at most four points from any sample set which are inside corners of the rectangle and have positive labels. And the reconstruction function has as a hypothesis the smallest rectangle consistent with these points. This class has a VCD of 4. Sample compression scheme is related to teaching dimension. The size of a scheme is the size of the largest subset resulting from compression of any sample consistent with some concept in the concept class F. Similarly, teachers provide particularly helpful examples to the student in the cooperative models of learning. It means they compress concepts to some specific subsets of examples.
There is a long-standing open conjecture which claims if VCD of a concept class F is d, the sample compression scheme of the concepts in F to be compressed to some subsets has a size no larger than d.

6.1 Compression scheme for maximum classes

Kuzmin and Warmuth, in 2005, were the first to find a compression scheme for maximum classes. They used one-inclusion graph and min-peeling algorithm to represent a compression function. As mentioned before, one-inclusion graph is an undirected graph which characterizes the concept class on a set of example points. The vertices are possible labelings of the example points and there is an edge between two concepts if they disagree on a single point. The edges are naturally labeled by the differing points. Using this definition we have the following algorithm for min-peeling:

Example. In the following example you can find one-inclusion graph and compression scheme for the concept class F. For the concept class F we have $VCD = 2$, $n = 4$ and $|F| = 11$. The representatives for each concept are indicating in the right column, as $r(c)$ function.

a	b	c	d	$r(c)$
0	0	0	0	$\emptyset$
0	0	1	0	$\{c\}$
0	0	1	1	$\{d\}$
0	1	0	0	$\{b\}$
0	1	0	1	$\{c, d\}$
0	1	1	0	$\{b, c\}$
0	1	1	1	$\{b, d\}$
1	0	0	0	$\{a\}$
1	0	1	0	$\{a, c\}$
1	0	1	1	$\{a, d\}$
1	1	0	0	$\{a, b\}$



Compression scheme for linear arrangements

In a d dimensional space, linear arrangement is a collection of oriented hyperplanes in R^d. The cells of the arrangement are the concepts and the planes the dimensions of the concept class. The orientations of the planes are indicated by arrows. All the cells above the plane of a dimension label that dimension with *one* and the cells below label it with *zero*. Let the number of instances be n. Linear arrangements are special maximum classes, because their VC dimension is $min(n, d)$ and they have $\binom{n}{\leq d}$ cells.

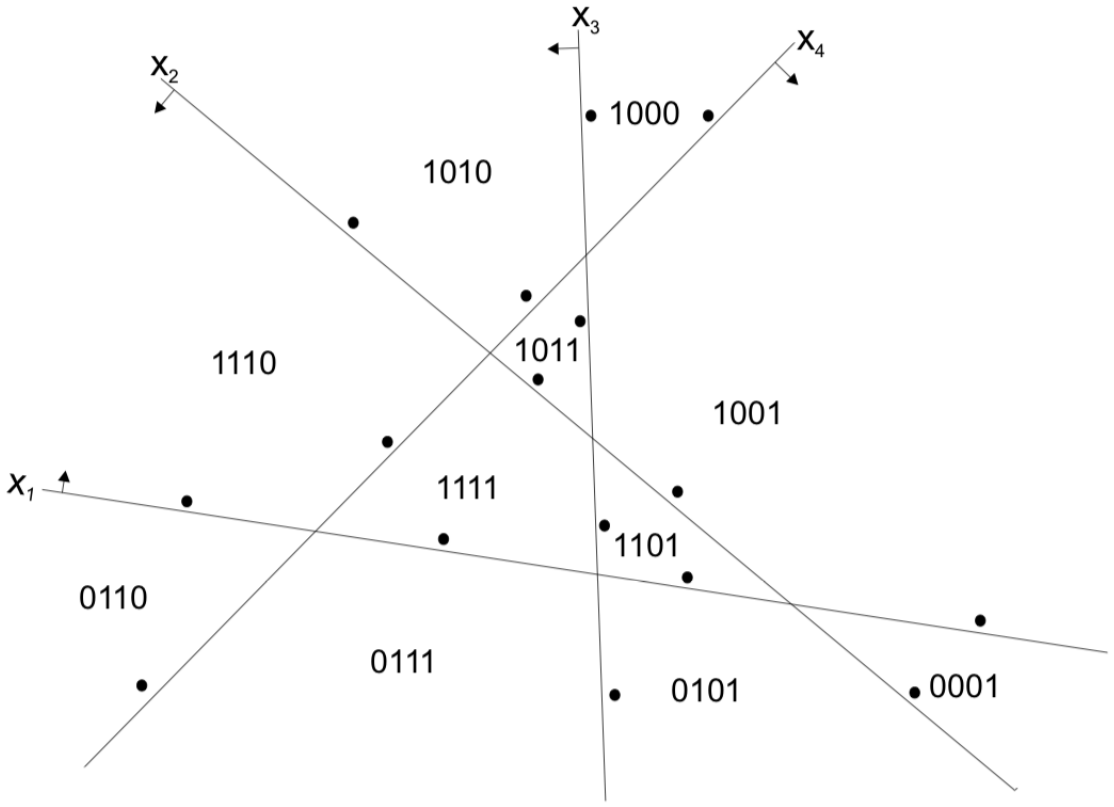


Figure 1: An example linear arrangement. The cells of the arrangement represent the concepts and the planes the dimensions of the class

Open Problems

1. Is RTD linearly upper bounded by VCD?
   Does $RTD \leq \kappa d$ hold for a universal constant κ?
2. No concept class known for which the ratio of RTD over $VCD = 5/3$.
3. The best known result is an exponential upper bound $RTD = \mathcal{O}(d \cdot 2^d)$ due Chen et al. (2016)
4. Quadratic upper bound:
$$RTD = \mathcal{O}(d^2)$$

Moran, Upper Bound for RTD

Let F be a concept class of $VCD(F) = d$. Then there exists $c \in F$ with a teaching set of size at most:

$$d2^{d+3}(\log(4e^2) + \log\log|F|) \tag{24}$$

Moral et al. if $|F| > (4e^2)^{d2^{d+2}}$, there exist two distinct x and $x^{'}$ in X such that the set of concepts $c \in F$ such that $c(x) \neq c(x^{'})$ is much smaller than $|F|$. More precisely they show:

$$|\{c \in F : c(x) = 0 \; and \; c(x^{'} = 1)\}| \leq |F|^{1-1/d2^{d+2}} \tag{25}$$

They add x and $x^{'}$ to the teaching set. They recursively do this until there exists a teaching set of size 1. This theorem implies the RTD is upper bounded by:

$$d2^{d+3}(\log(4e^2) + \log\log|F|) \tag{26}$$

However they leave the following open question:
Let F be a concept class of $VCD(F) = d$. Does there exist any upper bound for the RTD of F which is independent of size F?

7 Chen, Upper Bound for RTD

Theorem: Let $F \subset \{0,1\}^n$ be a class with $VCD(F) = d$, then:

$$RTD(F) \leq 2^{d+1}(d-2) + d + 4 \tag{27}$$

Following we prove this theorem using the observation that the VC dimension of a concept class does not increase after a concept is removed.
Lemma: Let $F \subset \{0,1\}^n$ be a class with $VCD(F) = d$, then:

$$TD_{min}(F) \leq 2^{d+1}(d-2) + d + 4 \tag{28}$$

Every class F with $VCD(F) = 1$ must have a concept $c \in F$ with a teaching set of size 1. Let ${F_b}^i$ denotes the set of concepts $c \in F$ such that $c_i = b$. By picking an index i and a bot b, we assume $\forall i, b\ {F_b}^i$ is not empty and has the smallest size. Since $VCD(F) = 1$, ${F_b}^i$ is a singleton set. If not, it has two different concepts which don't shatter.
Now we prove the Lemma by induction on d. Let:

$$f(d) = \max_{F:VCD(F)\leq d} TD_{min}(F) \tag{29}$$

Our goal is to prove the following upper bound for $f(d)$:

$$f(d) \leq 2^{d+1}(d-2) + d + 4 \tag{30}$$

The base case of $d = 1$ follows directly from Kulhmann which denotes every class F with $VCD(F) = 1$ must have a concept $c \in F$ with a teaching set of size 1. For the induction step we show that for some $d > 1$, $f(d) \leq 2^{d+1}(d-2) + d + 4$, assuming it holds for $d-1$. Let k be the difference between $f(d)$ and $f(d-1)$ so we have:

$$k = 2^d(d-1) + 1 \tag{31}$$

if $n \leq k$, we are done, because:

$$TD_{min}(F) \leq n \leq k = 2^d(d-1) + 1 \leq 2^{d+1}(d-2) + d + 4 \tag{32}$$

Where the last inequality holds for all $d \geq 1$.
Now, all we need is to show the inequality holds for $n > k$.
For any set of k indices $Y \subset [n]$ that $|Y| = k$, and any $b \in 0, 1^k$, we define:

$${F^Y}_b = \{c \in F : c \cap Y = b\} \tag{33}$$

Among all possible Y and b we choose Y^* and b^* such that ${F_{b^*}}^{Y^*}$ is nonempty and has the smallest size among all nonempty ${F_b}^Y$. For notational convenience, we write F_b to denote ${F_b}^Y$. Without loss of generality, we assume $b^* = 0$ and $Y^* = [k]$ is the all-zero string.
If F_{b^*} has VCD at most $d-1$, then:

$$TS_{min}(F) \leq k + f(d-1) \leq 2^{d+1}(d-2) + d + 4 \tag{34}$$

Using the inductive hypothesis . This is because according to the definition of f one of the concepts $c \in F_{b^*}$ has a teaching set $T \subset [n] \backslash Y^*$ of size at most $f(d-1)$ to distinguish it from other concepts of F_{B^*}. Since T and Y^* are disjoint sets, Thus:

$$|T \cup Y^*| = |T| \cup |Y^*| = |T| \cup [k] \tag{35}$$

Is the teaching set of c in the original class F of size at most $k + f(d-1)$.

$$|T| \cup [k] \leq k + f(d-1) \tag{36}$$

This finishes the induction and the proof of the Lemma.

8 Hu, Upper Bound on RTD

Definition: We say a concept class $F \subset \{0,1\}^n$ is an $(x,y) - class$ for positive integers x, y, if for any $A \subset [n]$ such that $|A| \leq x$, $|\{c|_A : c \in F\}| \leq y$.

Definition: Define $f(x,y) = \sup_F TD_{min}(F)$, where the supremum is taken over all finite $(x,y) - class$ F.
Kuhlmann proved $f(2,3) = 1$, and Moran proved $f(3,6) \leq 3$.

Theorem: For any concept class $F \subset \{0,1\}^n$ with $VCD(F) = d$,

$$RTD(F) = O(d^2) \tag{37}$$

Hu gave two descriptions for the theorem above, informal and formal descriptions. The informal description states that the key idea is to analyze $f(x,y)$, the largest possible best-case teaching dimension for $(x,y) - classes$.The first step is to show a recursive formula for $f(x,y)$. For a monotone increasing function $\phi(x)$ that grows substantially slower than 2^x, we have:

$$f(x+1, \phi(x+1)) \leq f(x, \phi(x)) + O(x) \tag{38}$$

This recursive formula leads to a quadratic upper bound:

$$f(x, \phi(x)) \leq O(x^2) \tag{39}$$

Then, the problems reduces to find a proper function $\phi(.)$. Let $\phi(x) = \alpha^x$ for certain $\alpha \in (1,2)$. For any finite concept class F with $VCD(F) = d$, F must be an $(x,y) - class$ for some x not much larger than d. In fact, it suffices when x is a constant times of d. Using this relation between VCD and $f(x,y)$, it denotes that the best-case teaching dimension of F is upper bounded by $O(d^2)$. Then it yields $RTD(F) = O(d^2)$. The second description by Hu which gives the formal proof is based on the next Lemma.

Lemma: For any positive integer x, y, z such that $y \leq 2^x - 1$ and $z \leq 2y + 1$, the following inequality holds:

$$f(x+1, z) \leq f(x,y) + \left\lceil \frac{(y+1)(x-1)+1}{2y - z + 2} \right\rceil \tag{40}$$

Using the recursive formula established in this Lemma, we're able to give upper bound on the best-case teaching complexity for all $(x,y) - classes$.

Lemma: For every $\alpha \in (1,2)$, and every positive integer x,

$$f(x, \lfloor \alpha^x \rfloor) \leq \frac{(x-1)^2}{4 - 2\alpha} + \frac{3 - 2\alpha}{4 - 2\alpha} \cdot (x-1) \tag{41}$$

Let $y = \lfloor \alpha^x \rfloor)$ and $z = \lfloor \alpha^{(x+1)} \rfloor)$, we have:

$$f(x+1, \lfloor \alpha^{x+1} \rfloor) \leq f(x, \lfloor \alpha^x \rfloor) + \left\lceil \frac{(\lfloor \alpha^x \rfloor + 1)(x-1)+1}{2\lfloor \alpha^x \rfloor - \lfloor \alpha^{x+1} \rfloor + 2} \right\rceil \tag{42}$$

Since:

$$\left\lceil \frac{(\lfloor \alpha^x \rfloor + 1)(x-1)+1}{2\lfloor \alpha^x \rfloor - \lfloor \alpha^{x+1} \rfloor + 2} \right\rceil \leq \frac{x-1}{2 - \frac{\lfloor \alpha^{x+1} \rfloor}{\lfloor \alpha^x + 1 \rfloor}} + \frac{\frac{1}{\lfloor \alpha^x \rfloor + 1}}{2 - \frac{\lfloor \alpha^{x+1} \rfloor}{\lfloor \alpha^x \rfloor + 1}} \leq \tag{43}$$

$$\frac{x-1}{2 - \frac{\lfloor \alpha^{x+1} \rfloor}{\lfloor \alpha^x + 1 \rfloor}} + 1 \leq \frac{x-1}{2-\alpha} + 1 = \frac{x+1-\alpha}{2-\alpha} \tag{44}$$

So we have:

$$f(x+1, \lfloor \alpha^{x+1} \rfloor) \leq f(x, \lfloor \alpha^x \rfloor) + \frac{x+1-\alpha}{2-\alpha} \tag{45}$$

By applying the above inequality recursively and using the equality $f(1,1) = 0$, we obtain:

$$f(x, \lfloor \alpha^x \rfloor) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1) \tag{46}$$

Now we show that if $VCD(F) = d$, for x not much larger than d, F must be an $(x, \lfloor \alpha^x \rfloor) - class$.

Lemma: Given $\alpha \in (1,2)$, define:

$$\lambda^* = inf\{\lambda \geq 1 : \lambda ln(\alpha) - ln(\lambda) - 1 \geq 0\} \tag{47}$$

Then for any concept class $F \subset \{0,1\}^n$ with $VCD(F) = d$, F is an $(x\lfloor \alpha^x \rfloor) - class)$ for every integer $x \geq \lambda^* d$.
Now we can give the main conclusion:

Theorem: For any concept class $F \subset \{0,1\}^n$ with $VCD(F) = d$,

$$RTD(F) \leq 39.3752d^2 - 3.6330d \tag{48}$$

To prove this theorem using the Lemmas mentioned before, we have for any $\alpha \in (1,2)$ and any $x \geq \lambda^* d$, the following holds:

$$TD_{min}(F) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1) \tag{49}$$

VCD of a concept class does not increase after a concept is removed, so we have:

$$RTD(F) \leq \frac{(x-1)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot (x-1) \tag{50}$$

The optimum coefficients in the quadratic bound are $\lambda^* = 4.71607$ and $\alpha = (e\lambda^*)^{1/\lambda^*} = 1.71757$. So $x = \lceil \lambda^* d \rceil$. Finally we conclude that:

$$RTD(F) \leq \frac{(\lambda^* d)^2}{4-2\alpha} + \frac{3-2\alpha}{4-2\alpha} \cdot \lambda^* d \leq 39.3752d^2 - 3.6330d \tag{51}$$