Machine Learning: Foundations **Spring Semester, 2002**

Lecture 6: May 2, 2002

Lecturer: Yishay Mansour *Scribe: Nir Andelman* [1]

6.1 The PAC Model - Review

In the PAC Model we assume there exists a distribution D on the examples that the learner receives; i.e. when choosing an instance from the sample it is drawn according to D. We assume that D is:

1. Fixed throughout the learning process.
2. Unknown to the learner.
3. The instances are chosen independently.

The target concept is specified as a computable function c_t, thus our instances are of the form $<x, c_t(x)>$. Our goal is to find a function h which approximates c_t with respect to D, in the following sense. Let

$$error(h) = \ Prob_D[c_t(x) \neq h(x)].$$

We would like to ensure that $error(h)$ is below a certain threshold ϵ,which is given as a parameter to the algorithm. This parameter is a measure of the accuracy of the learning process.

As a measure of our confidence in the outcome of the learning process, we add another parameter δ. We require that the following hold:

$$Prob[error(h) < \epsilon] \geq 1 - \delta.$$

The PAC algorithm has two inputs: the accuracy parameter ϵ and the confidence parameter δ. It also has access to instances using $EX(D, c_t)$, which generates a random example, using the distribution D, and labelled by c_t.

We say that an algorithm A *learns* a family of concepts $\mathcal{C}$ if for **any** $c_t \in \mathcal{C}$ and **any** distribution D on the instances in $\mathcal{C}$, A outputs a function h, such that the probability that $error(h) < \epsilon$ is at least $1 - \delta$.

[1] Based on scribe written by Vladimir Goldner

A PAC algorithm is *efficient* if its running time is polynomial in $\frac{1}{\epsilon}$, $\ln\frac{1}{\delta}$, the input size and the size of the target concept c_t.

6.2 THE VC-DIMENSION

6.2.1 Motivation

Let us consider the following question: How many random examples does a learning algorithm need to draw before it has sufficient information to learn an unknown target concept chosen from the concept class C? For the case of a finite concept class C, we proved a lower bound on the number of examples required for PAC learning:

$$m \geq \frac{1}{\epsilon}\ln\frac{|\mathcal{C}|}{\delta}$$

We would like to be able to handle infinite concept classes, perhaps even not enumerable. We saw already some examples:

- concept of axes-parallel rectangles,
- concept of C_θ for $\theta \in [0,1]$: $C_\theta(x) = 0$ for $x < \theta$ and $C_\theta(x) = 1$ for $x \geq \theta$.

Here we saw that the number of examples sufficient for PAC learning is $\frac{1}{\epsilon}\ln\frac{1}{\delta}$ (between the leftist '1' and the rightist '0' the weight will be at most ϵ by probability $1-\delta$).
Frequently, there is a significant structure in C, and we like to formally quantify how this structure helps our learning algorithms. We will introduce the definition of VC–dimension and show the connection between the VC–dimension and learning. The concept of the VC–dimension, will provide us a substitute to $\ln|C|$, for infinite concept classes.

6.2.2 Definitions

We start with few definitions. Assume C is a concept class defined over instance space X. let $S \subseteq X$.

Definition For each class C over X and for any $S \subseteq X$:

$$\Pi_C(S) = \{c \cap S | c \in C\}$$

Equivalently, if $S = \{x_1, \ldots, x_m\}$ then we can think of $\Pi_C(S)$ as the set of vectors $\Pi_C(S) \subseteq \{0,1\}^m$ defined by $\Pi_C(S) = \{< c(x_1), \ldots, c(x_m) >: c \in C\}$.

This is the projection of the concept class C on the input S, namely $\Pi_C(S)$ is all the possible functions that C creates on S. We are interested in how many different functions

C creates on S. In effect we are reducing the concept class C into the concept class $C|S$, where $S = \{x_1, ..., x_m\}$. The concept class $C|S$ is finite with at most 2^m different concepts, thus $|\Pi_C(S)| \leq 2^m$.

Definition A concept class C *shatters* S if $2^{|S|} = |\Pi_C(S)|$.

In other words a class shatters a set of inputs if every possible function on S can be represented by some $c \in C$.

Now we are ready to define the notion of VC–dimension.

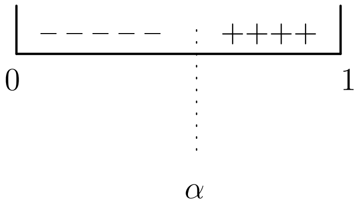
Definition *VCdim (Vapnik-Chervonenkis dimension)* of C is the maximum size of a set shattered by C:

$$VCdim(C) = max\{d : \exists S : |S| = d \ \ and \ \ \Pi_C(S) = \{0,1\}^d\}.$$

$VCdim(C) = \infty$, if such a maximum as above doesn't exist, i.e. there exist sets as large as we want which are shattered by C.

6.2.3 Some examples of geometric concepts

- C_1 - the concepts are c_α for $\alpha \in [0,1]$:



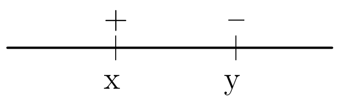
Note that although the number of concepts is uncountable the concept class C_1 is learnable. The reason for C_1's learnability is the structure of the concept class, as in fact VCdim(C)=1:

Let $S = \{\frac{1}{2}\}$, we will show 2 concepts, such that $|\Pi_C(\{\frac{1}{2}\})| = 2$.

$$c_{\frac{3}{4}} \Rightarrow \frac{1}{2} \ is \ '-' \ \ (negative \ \ example)$$

$$c_{\frac{1}{4}} \Rightarrow \frac{1}{2} \ is \ '+' \ \ (positive \ \ example)$$

thus the VC–dimension is at least 1.



We will now show that the dimension is less than 2. For any two points, x, y ($y > x$), the assignment that lets x be '+' and y be '−', is impossible. Thus, $VCdim(C_1) < 2$, and we derive $VCdim(C_1) = 1$.

- C_2 - straight line which divides the plane. All the positive points are above or on the line, and all the negative points are below the line.
  Formally:

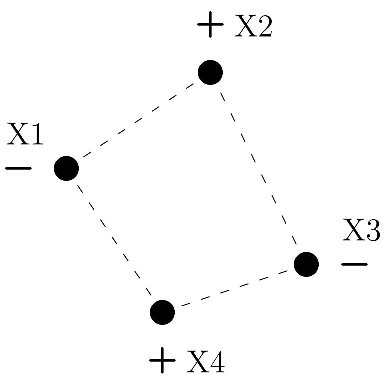
$$C_2 = \{c_w | w = (\alpha_1, \alpha_2, \theta)\},$$

where

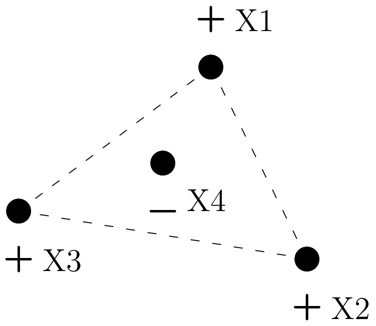
$$c_w(x_1, x_2) = 1 \iff \alpha_1 x_1 + \alpha_2 x_2 \geq \theta \qquad (x_1, x_2) \in R^2.$$

Note that here too, the number of "legal" assignments is distinctively smaller than those possible in principle, implying a highly structured concept space. If we take any 3 points not on the same line, all 3 assignments are possible. This shows that $VCdim(C_2) \geq 3$. However, for any 4 points in the plane there are two different possible structures, for each we show an impossible assignment, and thus the VC–dimension is less than 4.

1.



2.



In case 1 the four points are on the convex-hull of the four points. In this case, one can easily verify that the assignment:

$$< x_1, - >, < x_2, + >, < x_3, - >, < x_4, + >$$

is impossible.
In case 2 there is a point inside the convex-hull of the other three points. In this case, the assignment:

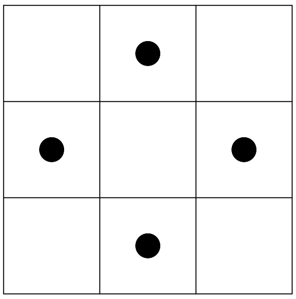
$$< x_1, + >, < x_2, + >, < x_3, + >, < x_4, - >$$

is impossible.
Thus, $VCdim(C_2) = 3$.

- C_3 - Parallel Rectangles. Rectangles for which the edges are parallel to the axis. Positive examples are points inside the rectangle, and negative examples are points outside the rectangle.

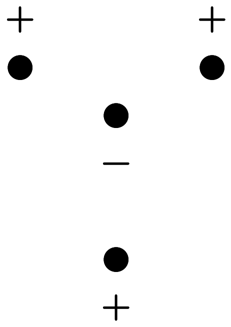
It is enough to have at least one setting for which all the classification are possible, in order to have $VCdim(C_3) \geq 4$, even if there is a setting for which there is a classification which is not possible.

We can show that the dimension is at least 4 :

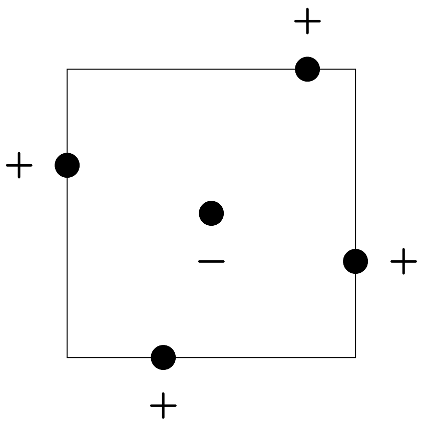


It is obvious from the figure above that four points can be shattered, by choosing the appropriate rectangle.

Example to a setting of four points and a classification which is not possible:

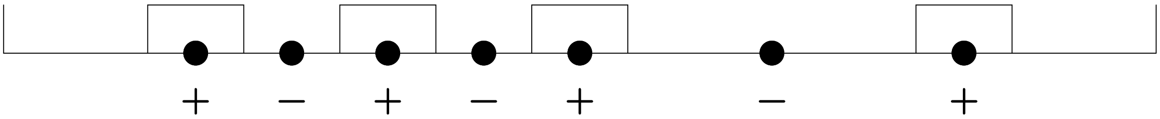


However, for any 5 points in the plane, and for every structure we can draw a rectangle by 4 points and the fifth is inside, so if we take an assignment which assign the external points to be '+' and the internal point to be '-' we get an impossible assignment (see the following figure).



And thus we get $VCdim(C_3) = 4$.

- C_4 - A finite union of intervals.



For any set of points we could cover the positive points by choosing the intervals small enough.

Thus, $VCdim(C_4) = \infty$.

- C_5 - Parity.
$X = \{0,1\}^n$. The concept class is

$$\chi_S(x) = \oplus_{i \in S} x_i$$

where $S \subset \{1, ..., n\}$. The lower bound: $VCdim(C_5) \geq n$
Let $e_i = 0...010...0$ unit vectors, where '1' appears in the i-th place. There are n such vectors. For any bits assignment $b_1, ..., b_n$ for the vectors $e_1, ..., e_n$ we choose the set

$$S = \{i : b_i = 1\}$$

We get

$$\chi_S(e_j) = \begin{cases} 1 & \text{j} \in S \\ 0 & \text{j} \notin S \end{cases}$$

Thus, we conclude $VCdim(C_5) \geq n$.

The upper bound: $VCdim(C_5) \leq n$
We present two simple proofs for the upper bound:

1. There are 2^n parity functions. Thus $VCdim(C_5) \leq log_2 2^n = n$.
2. Given $n+1$ vectors, there is a vector that is the linear combination of the others:

$$e_j = e_1 \oplus ... \oplus e_k$$

   Therefore, the values of $e_1, ..., e_k$ fix the value of e_j. So the assignment

$$b_1 = 0, ..., b_k = 0, \; b_j = 1$$

   is impossible.

- C_6 - OR of n literals.
  $X = \{0,1\}^n, \; S \subset \{1, ..., n\}, ; \; \bar{S} \subset \{1, ..., n\}$. The concept class is:

$$C_S(x) = (\cup_{i \in S} x_i) \cup (\cup_{j \in \bar{S}} \bar{x_j})$$

  The lower bound: $VCdim(C_6) \geq n$
  Let $e_i = 0...010...0$ where '1' appears in the i-th place. There are n such vectors. For any bits assignment $b_1, ..., b_n$ for the vectors $e_1, ..., e_n$ we choose the sets

$$S = \{i : b_i = 1\}$$

  so the target concept is

$$C_S(x) = \cup_{i \in S} x_i$$

  Thus, we conclude $VCdim(C_6) \geq n$.

Claim 6.1 *The upper bound: $VCdim(C_6) \leq n$*

Proof: Suppose we have $n+1$ vectors.
In one of the previous lectures we saw algorithm ELIM that maintains a literals list L, that is initialized to the set of all literals, and each assignment of 0 to vector removes all positive literals of the vector from L.
Let assign 0 to the first vector. It removes n literals from L. If some next vector does not remove any literal from L, then all it's literals was already removed, and then we can't assign 1 to such vector, so there is an infeasible assignment.
Else each next vector removes at least one literal.
We will show that there exists an order of vectors, such that the second vector can remove at least **two** literals.

Lemma 6.2 *Given set of 3 (or more) vectors, there exist two of them such that they differ by at least two bits.*

Proof: We can observe that if two vectors differ by only one bit, then the XOR's of bits of each of them are different. Then if two different vectors have the same XOR, then they are different by more than one bit. Given 3 (or more) vectors, there exist two of them with same XOR. The lemma is follows. □

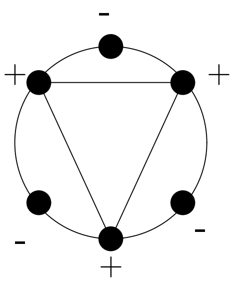
Let a, b two such vectors. We can perform two first ELIM stages on this two vectors. So, b will remove at least two literals from L (in addition to n literals, that was removed by a), because b has two bits, that differ from the appropriate bits of a.
All the next vectors remove from L at least one literal. Then before the assignment of the last $(n+1)$ vector, the list L is empty. So we can't assign 1 to it. Thus,

$$VCdim(C_6) \leq n$$

□

- C_7 - Convex polygons.
  Points inside the convex polygon are positive and outside are negative.
  Again, we have no bound on the number of edges, and we want to show $VCdim(C_7) = \infty$; i.e. for every d there is a set whose size is d that can be shattered by convex polygons.

  Let S be a set of d points on the circle perimeter. We show that for every labelling of the points in S, there exists $c_t \in C$ that is consistent with the labelling. The concept c_t connects the + points. The polygon includes all the positive examples and none of the negative ones. Thus, for any d points on the perimeter of the unit circle, all the 2^d classifications are possible. Therefore, $VCdim(C) = \infty$.



  It is clear that for any d one can choose d points located on a circle for which any 2^d assignments exist. In the figure above $d = 6$.

- C_8 - Hyper Plane.
  Let l_w be a hyper plane which devides R^n into 2 sets of points:
  l_w^+ - points above or on the hyper plane l_w.
  l_w^- - points below the hyper plane l_w.

Formally; for $w = (\alpha_1, ..., \alpha_n, \theta) \in R^{n+1}$

$$l_w = \{y \in R^n | \sum_{i=0}^{n} \alpha_i y_i = \theta\}$$

We define the classification by h_w as,

$$h_w(y) = 1 \iff \sum_{i=0}^{n} \alpha_i y_i \geq \theta$$

We will prove the following bound.

Theorem 6.3

$$VCdim(C_8) = n + 1$$

First we show that there are at least $n + 1$ points that can be shattered.

Claim 6.4

$$VCdim(C_8) \geq n + 1$$

Proof: Let $E = \{\overrightarrow{0}, \overrightarrow{e_1}, ..., \overrightarrow{e_n}\}$ of size $n + 1$ in order to show that C_8 shatters it. Any classification of the vectors of E can be viewed as a subset $S \subset E$ of positive classification and each vector in $E \setminus S$ is classified as negative. For each S we define a hyper plane W_s,

$$W_s = (\alpha_1^s, ..., \alpha_n^s, \theta^s),$$

where,

$$\theta^s = \begin{cases} -\frac{1}{2} & 0 \in S \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

and,

$$\alpha_i^s = \begin{cases} 1 & e_i \in S \\ -1 & \text{otherwise.} \end{cases}$$

W_s is the hyper plane which classifies every vector in S as $'+'$ and vectors in $E \setminus S$ as $'-'$. Since

$$h_{w_s}(e_i) = 1 \iff \alpha_i^s \geq \theta^s \iff e_i \in S$$

$$h_{w_s}(0) = 1 \iff 0 > \theta^s \iff 0 \in S$$

□

We showed that there exists a set of size $n+1$ which C_8 shatters, hence $VCdim(C_8) \geq n+1$.
We will now show that $VCdim(C_8) = n+1$.
Before further examination can be done, some general definitions and Radon theorem will be shown.

Definition A subset A is convex if $\forall x, y \in A$ the line connecting x to y is in A.

Formally:
$\forall \lambda$ such as $0 < \lambda < 1 \qquad \lambda x + (1-\lambda)y \in A$

Definition The *Convex Hull* of S is the smallest convex set which contains all the points of S. We denote it as $conv(S)$.

We are now ready to state (Radon Theorem).

Theorem 6.5 *(RADON Theorem) Let E be a set of d + 2 points in R^d. There is a non empty subset S of E such that*

$$conv(S) \cap conv(E \setminus S) \neq \phi$$

Proof: Let:

$$E = \{x_0, ..., x_{d+1}\}$$

where $x_i \in R^d$
Since E contains $d+2$ vectors, we can solve for the following $d+1$ equations and find $(\alpha_0, ..., \alpha_{d+1}) \neq \overrightarrow{0}$, such that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0,$$

and

$$\sum_{i=0}^{d+1} \alpha_i = 0.$$

Thus, we have a set of $d+1$ linear equations over $d+2$ variables $\{\alpha_i\}_{i=0}^{d+1}$. There exist a non-zero vector $< \alpha_0, ..., \alpha_{d+1} >$ satisfying the above equations, because every $d+1$

points (vectors) are linear dependent.
Assume that $\alpha_0, ..., \alpha_p$ are positive, and $\alpha_{p+1}, ..., \alpha_{d+1}$ are negative.
We define:

- $\alpha = \sum_{i=0}^{p} \alpha_i > 0$
- $\beta_i = \frac{\alpha_i}{\alpha} > 0 \qquad 0 \leq i \leq p$
- $\gamma i = \frac{-\alpha_i}{\alpha} > 0 \qquad p+1 \leq i \leq d+1$

We have that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0 \;\Rightarrow\; \sum_{i=0}^{p} \beta_i x_i = \sum_{i=p+1}^{d+1} \gamma_i x_i$$

Notice that $\sum_{i=0}^{p} \beta_i = \sum_{i=p+1}^{d+1} \gamma_i = 1$.
By definition of convexity,

$$\sum_{i=0}^{p} \beta_i x_i \in conv(x_0, ..., x_p),$$

and

$$\sum_{i=p+1}^{d+1} \gamma_i x_i \in conv(x_{p+1}, ..., x_{d+1}).$$

Hence, there is a point that belongs to the intersection of

$$conv(x_0, ..., x_p) \cap conv(x_0, ..., x_p) \neq \phi$$

□

Claim 6.6

$$VCdim(C_8) < n+2$$

Proof: Proof by contradiction. Assume $E = \{x_1, ..., x_{n+2}\}$ could be shattered. By RADON theorem there is a subset S of E such that $conv(S) \cap conv(E \setminus S) \neq \phi$.
Assume we have hyper plane H_w which classifies S as '+' and $E \setminus S$ as '-', and since hyper planes are convex,

$$S \subset l_w^+ \Rightarrow conv(S) \subset l_w^+,$$

and

$$E \setminus S \subset l_w^- \Rightarrow conv(E \setminus S) \subset l_w^-.$$

Combining the two, we have that,

$$conv(S) \cap conv(E \setminus S) \subset l_w^+ \cap l_w^- = \phi,$$

which is a contradiction to the choice of S. Therefore such a hyper plane H_w exists. □

Combining Claim 5.3 and Claim 5.5, we can now conclude that $VCdim(C_8) = n + 1$

6.3 Bibliographic Notes

The presentation of the material of this lecture follows closely [1].

1. An introduction to Computational Learning Theory.