Computational Learning Theory **Spring Semester, 1997**

Lecture 7: April 9,1997

Lecturer: Yishay Mansour *Scribe: Vladimir Goldner* [1]

7.1 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 proccess.
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 labeled 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 Ayelet Shachar,Itamar Tsadok& Odelia Dalal

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.

7.2 THE VC-DIMENSION

7.2.1 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:

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

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 leftest '1' and the rightest '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–dimention and show the connection between the VC–dimension and learning. The concept of the VC–dimention, will provide us a substitute to $\ln |C|$, for infinite concept classes.

7.2.2 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.

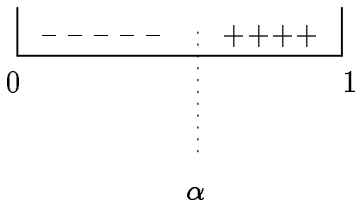
Definition *VCdim (Vapnik-Chervonenkis dimension)* of C is the maximum size of a set shattered by C:

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

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

7.2.3 Some examples of geometric concepts

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



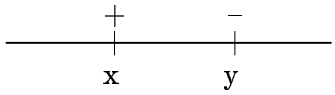
Note that although the number of concepts is uncountable the concept class C_1 is learnable. The reason for C_1's learnability is the structure of the concept class, as in fact VCdim(C)=1:

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 devides the plane. All the positive points are above or on the line, and all the negative points are below the line.
  Formally:

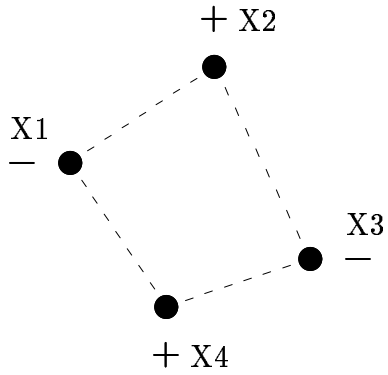
$$C_2 = \{c_w | w = (\alpha_1, \alpha_2, \theta)\},$$

where

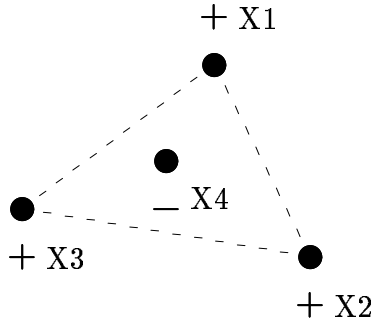
$$c_w(x_1, x_2) = 1 \iff \alpha_1 x_1 + \alpha_2 x_2 \geq \theta \qquad (x_1, x_2) \in R^2.$$

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–dimention 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:

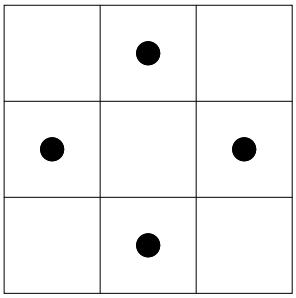
$$< x_1, + >, < x_2, + >, < x_3, + >, < x_4, - >$$

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.

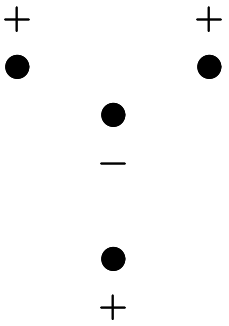
It is enough to have at least one setting for which all the classification are possible, in order to have $VCdim(C_3) \geq 4$, even if there is a setting for which there is a classification which is not possible.

We can show that the dimention is at least 4 :

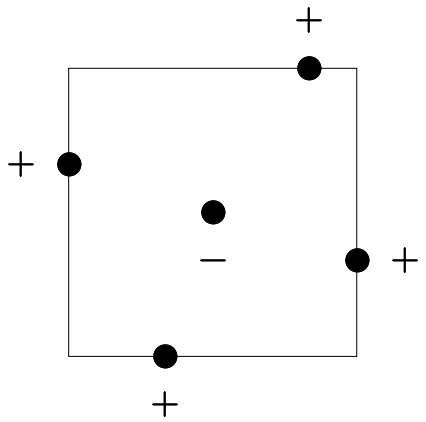


It is obvious from the figure above that four points can be shattered, by chosing the appropriate rectangle.

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

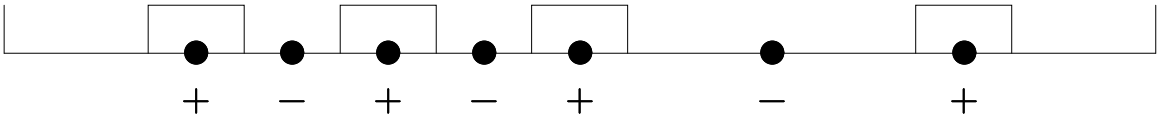


However, for any 5 points in the plane, and for every structure we can draw a rectangle by 4 points and the fifth is inside, so if we take an assignment which assign the external points to be '+' and the internal point to be '-' we get an impossible assignment (see the following figure).



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 chosing 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 7.1 *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 7.2 *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. □

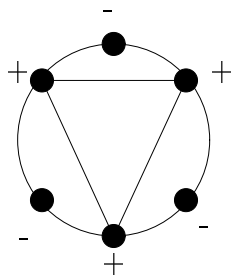
Let a, b two such vectors. We can perform two first ELIM stages on this two vectors. So, b will remove at least two literals from L (in addition to n literals, that was removed by a), because b has two bits, that differ from the appropriate bits of a.
All the next vectors remove from L at least one literal. 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 labeling of the points in S, there exists $c_t \in C$ that is consistent with the labeling. The concept c_t conects 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 7.3

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

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

Claim 7.4

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

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 7.5 *(RADON Theorem) Let E be a set of $d+2$ points in R^n. 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 $d+2$ equations and find $(\alpha_0, ..., \alpha_{d+1}) \neq \overrightarrow{0}$, such that,

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

and

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

There exist $< \alpha_0, ..., \alpha_{d+1} >$ which satisfies the above equations, because every $d+1$ points (vectors) are linear dependant and we can choose the $d+2$ point to satisfy the

last equation.
Without loss of generality, assume that $\alpha_0, ..., \alpha_p$ are positive, and $\alpha_{p+1}, ..., \alpha_{d+1}$ are negative. equation.
We define:

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

Using the first equation, we have that,

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

We know that $\sum_{i=0}^{p} \beta_i = \sum_{i=0}^{p} \frac{\alpha_i}{\alpha} = 1$. Since $\sum_{i=0}^{d+1} \frac{\alpha_i}{\alpha} = 0$ then $\sum_{i=p+1}^{d+1} \gamma_i = 1$.
Since $\sum_{i=0}^{p} \beta_i = 1$ we have,

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

also, since $\sum_{i=p+1}^{d+1} \gamma_i = 1$,

$$\sum_{i=p+1}^{d+1} \gamma_i x_i \in conv(x_0, ..., x_p).$$

Combining the two, with the fact that $\sum_{i=0}^{p} \beta_i x_i = \sum_{i=p+1}^{d+1} \gamma_i x_i$, we have that,

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

□

Claim 7.6

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

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 choise of S. Therefore such a hyper plane H_w exists. □

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

7.3 Lower Bounds

7.3.1 Static Learning Model

Definition The *static learning algorithm* has the following characteristics:

- It asks for a sample of size $m(\epsilon, \delta)$
- It choses it's hypotheses (based on the sample it got).

The main limitation is that the algorithm cannot update the number of examples necessary depending on the sample it receives in step 1.

In the folllowing theorem for the static learning model, we show that if $VCdim = \infty$, then no static algorithm exists. Namely a sample size which depends *only* on ϵ and δ is non-existant for classes of infinite VC–dimention. (If we allow the sampling to depend also on the target concept then the scenario will change; For example, the case of the convex polygon, if we let the sampling size depend on the number of sides.)

Theorem 7.7 *If a concept class C has $VCdim(C) = \infty$ then C is not learnable by any static learning algorithm.*

Proof: The proof is by contradiction. We will set ϵ and δ as follows: $\epsilon < \frac{1}{8}$ and $\delta < \frac{1}{7}$, e.g., $\epsilon = \delta = \frac{1}{10}$. The proof is by contradiction, let $m = m(\frac{1}{10}, \frac{1}{10})$ and let A be a static algorithm that learns C using m examples. Based on our assumption that C has infinite dimension ($VCdim(C) = \infty$), there exist $2m$ points, $z_1, ...z_{2m}$, which C shatters. Let $E = \{z_1, ..., z_{2m}\}$.

We define the distribution D as follows:

$$D(x) = \begin{cases} \frac{1}{2m} & x \in E \\ 0 & \text{otherwise} \end{cases}$$

Thus D is a uniform distribution over E (and 0 elsewhere).
Since E is shattered by C, then any concept is possible. Thus, we shall choose a concept c_t as follows:
For each z_i $Pr[c_t(z_i) = 0] = \frac{1}{2}$. There exists a c_t which fits the random choice, because C shatters E. The learning algorithm A samples m points: $B = \{x_1, ..., x_m\}$ (drawn according to D), clearly $|B| \leq m$.

Let h be the hypothesis which the algorithm A outputs. For every point, $z_i \notin B$, the probability that h makes an error is $\frac{1}{2}$, i.e.

$$Pr[c_t(z_i) \neq h(z_i)] = \frac{1}{2}$$

because for each point not sampled by the algorithm, we chose it's value to be 0 or 1, randomly. Thus it makes no difference what the algorithm chose, the probability of mistake is $\frac{1}{2}$.

Computing the expected error for h gives,

$$E[error(h)] \geq m \cdot \frac{1}{2} \cdot \frac{1}{2m} = \frac{1}{4}$$

- m - at least m points were not seen (the number of points not in B)
- $\frac{1}{2}$ - Probability of error.
- $\frac{1}{2m}$ - Probability of z_i.

We must demonstrate that with a probability $> \frac{1}{10}$ we have an hypothesis, whose error is $> \frac{1}{10}$ (and that will bring to a condtradiction that the error $< \epsilon = \frac{1}{10}$). Let,

$$\alpha = Prob[error(h) \leq \frac{1}{10}] \ .$$

For the algorithm to be correct we need that $\alpha \geq \frac{9}{10}$. We will show that $\alpha \leq \frac{5}{6}$ and reach a contradiction. We bound α as follows,

$$\frac{1}{4} \leq E[error(h)] \leq \alpha \cdot \frac{1}{10} + (1 - \alpha) \cdot 1 = 1 - \frac{9}{10} \cdot \alpha$$

Hence,

$$\alpha \leq \frac{5}{6}$$

Thus we have reached a contradiction to our assumption that $\alpha \geq \frac{9}{10}$.
Therefore if $VCdim(C) = \infty$, there is no integer m such that it is sufficient to sample m examples and gurantee PAC learning with $\epsilon = \frac{1}{10}$ and $\delta = \frac{1}{10}$. □

7.3.2 Quantitative Lower Bounds

In the previous section we showed that if $VCdim(C) = \infty$ then we can not bound the sample size. In fact it was sufficient that $VCdim(C) = 2d$, then a sample of size d is not sufficient.

We improve this bound by showing a dependence also on ϵ.

Theorem 7.8 *If C is a class for which $d = VCdim(C)$, then*

$$m(\epsilon, \delta) = \Omega(\frac{d}{\epsilon})$$

Proof: Let $E = \{z_1, ..., z_d\}$, such that C shatters E. (We know that such a group exists, because $d = VCdim(C)$). Now, differently from the previous proof, we will not use a uniform distribution, but instead, we shall define D as follows:

We add a new point z_0 with probability $1 - 8\epsilon$. So, each z_i $(i > 0)$ has probability $\frac{8\epsilon}{d}$.

Formally,

$$D = \begin{cases} 1 - 8\epsilon & x = z_0 \\ \frac{8\epsilon}{d-1} & x = z_i \ , \ i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Again, we will choose c_t randomally: $Pr[c_t(z_i) = 0] = \frac{1}{2}$. As before, we will ask, how many samples are necessary, under this distribution, in order to receive $\frac{d}{2}$ of the z's. Let's assume that the learner, "knows" only $\frac{d}{2}$ different points of z_i, $i > 0$, then:

$$E[error(h)] \geq \frac{d}{2} \cdot \frac{1}{2} \cdot \frac{8\epsilon}{d} = 2\epsilon$$

Let,

$$\alpha = Prob[error(h) \leq \epsilon] \ .$$

Then:

$$2\epsilon = E[error(h)] \leq \alpha\epsilon + (1 - \alpha) \cdot 1 \ .$$

And we get:

$$\alpha \leq \frac{1 - 2\epsilon}{1 - \epsilon} = 1 - \frac{\epsilon}{1 - \epsilon} \ .$$

So if we choose $\delta < \frac{\epsilon}{1-\epsilon}$, we will fail.

The expectation of examples number needed to reach $\frac{d}{2}$ samples from $\{z_1, ..., z_n\}$ is $\frac{d}{2} \cdot \frac{1}{8\epsilon} = \frac{d}{16\epsilon}$. Thus, in order to examine $\frac{d}{2}$ samples we need $m = \Omega(\frac{d}{\epsilon})$. □

Theorem 7.9 *If C is a class for which $VCdim(C) = d$, and H is a hypothesis class that is not obliged to include the target concept c_t, then*

$$m(\epsilon, \delta) = \Omega(\frac{d}{\epsilon^2})$$

Proof: Suppose, we have a single point $H = \{h_0(x) = 0, h_1(x) = 1\}$, and two distributions:

$$D_0 = \begin{cases} \frac{1}{2} - \gamma & < x, 1 > \\ \frac{1}{2} + \gamma & < x, 0 > \end{cases} \quad D_1 = \begin{cases} \frac{1}{2} + \gamma & < x, 1 > \\ \frac{1}{2} - \gamma & < x, 0 > \end{cases}$$

Let $\epsilon = \frac{\gamma}{2}$. We want to return a hypothesis with $error \leq optimum + \epsilon$. In this case:

- if the distribution is D_0 we must return h_0,
- if the distribution is D_1 we must return h_1.

If we have a sample of size m, then we will choose between h_0 and h_1 by the of majority of the points *(Maximum Likelihood)*.
The question is what is the probability that in m coin flips with deflection $\frac{1}{2}+\gamma$ the majority will be 0.
The median of a binomial distribution of (n, p) is either $\lfloor pn \rfloor$ or $\lceil pn \rceil$ for sufficiently large n. In our case this implies that the median is either $\left\lfloor \frac{m}{2} + \gamma m \right\rfloor$ or $\left\lceil \frac{m}{2} + \gamma m \right\rceil$.
Suppose each x_i is distributed $Prob[x_i = 1] = \frac{1}{2} + \gamma$. Then

$$Prob[\sum x_i \leq \frac{m}{2}] \;\geq\; \frac{1}{2} \;-\; Prob[\frac{m}{2} < \sum x_i \leq \frac{m}{2} + \gamma m]$$

The probability that $\sum x_i = \frac{m}{2} + \gamma m$ is maximal, and approximately equal to

$$\binom{m}{\frac{m}{2} + \gamma m} (\frac{1}{2} + \gamma)^{\frac{m}{2}+\gamma m} (\frac{1}{2} - \gamma)^{\frac{m}{2}-\gamma m}$$

$$\approx \frac{\sqrt{2\pi m}(\frac{m}{e})^m}{\sqrt{2\pi m(\frac{1}{2}+\gamma)}[\frac{m(\frac{1}{2}+\gamma)}{e}]^{m(\frac{1}{2}+\gamma)} \sqrt{2\pi m(\frac{1}{2}-\gamma)}[\frac{m(\frac{1}{2}-\gamma)}{e}]^{m(\frac{1}{2}-\gamma)}} (\frac{1}{2} + \gamma)^{\frac{m}{2}+\gamma m} (\frac{1}{2} - \gamma)^{\frac{m}{2}-\gamma m}$$

$$= \frac{1}{\sqrt{2\pi m(\frac{1}{2}+\gamma)(\frac{1}{2}-\gamma)}} = \frac{c}{\sqrt{m}}$$

where the approximation is using the Sterling formula for factorial:

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$$

Let us compute the probability that $\sum x_i = \frac{m}{2}$,

$$\binom{m}{m/2}(\frac{1}{2}+\gamma)^{\frac{m}{2}}(\frac{1}{2}-\gamma)^{\frac{m}{2}} \approx \frac{\sqrt{2\pi m}(\frac{m}{e})^m}{\sqrt{2\pi\frac{m}{2}}(\frac{m}{2e})^m\sqrt{2\pi\frac{m}{2}}(\frac{m}{2e})^m}(\frac{1}{4}-\gamma^2)^{\frac{m}{2}}$$

$$= \frac{4^{\frac{m}{2}}}{\sqrt{\pi\frac{m}{2}}}(\frac{1}{4}-\gamma^2)^{\frac{m}{2}} = \frac{(1-4\gamma^2)^{\frac{m}{2}}}{\sqrt{\pi\frac{m}{2}}} = \frac{c'}{\sqrt{m}}$$

where we assume that $\gamma^2 << m$. Thus,

$$Prob[\sum x_i \leq \frac{m}{2}] \;\geq\; \frac{1}{2} \;-\; \gamma m \cdot \frac{c'}{\sqrt{m}}$$

If we define $\delta = \frac{1}{4}$, then we get

$$\frac{1}{4} \geq \frac{1}{2} - \gamma m \cdot \frac{c'}{\sqrt{m}}$$

Thus, $c\gamma\sqrt{m} \geq \frac{1}{4}$, and we conclude that the sample size

$$m = \Omega(\frac{1}{\gamma^2}) = \Omega(\frac{1}{\epsilon^2})$$

For a general $VCdim(C) = d$, let $\{x_1, ..., x_d\}$ is a shattered set. Every x_i is chosen by probability $\frac{1}{d}$. Let assign to x_i a value 0 or 1 by probability $\frac{1}{2}+\gamma$ or $\frac{1}{2}-\gamma$ (by D_0 or D_1), in all 2^d distinct distributions, and we don't know which of them was chosen.
The hypothesis class H gives for each x_i either 0 or 1. Given a sample we will assign a value to x_i according to the majority of the classifications we got.
If we will give wrong classifications to $\frac{d}{16}$ points, then we get

$$error = (\frac{1}{2}-\gamma) + \frac{d}{16}\cdot(\frac{1}{d}\cdot 2\gamma) = \frac{1}{2} - \frac{7}{8}\gamma$$

when $\frac{1}{2}-\gamma$ is the optimal error.

Let $\epsilon = \frac{\gamma}{32}$. Then we can't mistake in more than $\frac{d}{16}$ points. Any point that sampled less than $\frac{c}{\gamma^2}$ will get wrong classification by probability $\frac{1}{4}$.

Let d_1 to be the number of points that was sampled less than $\frac{c}{\gamma^2}$ times. If $d_1 < \frac{d}{2}$, then at least $\frac{d}{2}$ points was sampled $\frac{c}{\gamma^2}$ times, that means $m = \Omega(\frac{d}{\gamma^2})$. Otherwise, if $d_1 > \frac{d}{2}$, then by average we will mistake on $\frac{d_1}{4} > \frac{d}{8}$ points. The probability of at least $\frac{d}{16}$ mistakes is at least $\frac{1}{16}$. So, for $\delta = \frac{1}{16}$, we reached to the contradiction, because we saw that we can't mistake in more than $\frac{d}{16}$ points.

Thus, the sample size is at least $\Omega(\frac{d}{\gamma^2}) = \Omega(\frac{d}{\epsilon^2})$.

□

7.4 Bibliographic Notes

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

1. An introduction to Computational Learning Theory.