MACHINE LEARNING CS60050

Instructor: Dr. Aritra Hazra
Department of Computer Science Engineering
Indian Institute of Technology, Kharagpur
Scribed by: Arun B; 20CS92R02

Notes for class on 19th February 2021

1 Goal of Learning

If learning is feasible, it is likely that

$$\mathbf{E_{out}(g) \approx E_{in}(g)}$$

If $g \approx f$ then, $E_{out}(g) \approx 0$. This can be achieved if the following condtitions are satisfied:

1. The value of $E_{out}(g)$ must be close to $E_{in}(g)$. i.e., $\mathbf{E_{out}(g) \approx E_{in}(g)}$
2. The value of $E_{in}(g)$ must be very small. i.e., $\mathbf{E_{in}(g) \approx 0}$

Note that, as the model complexity increases, $E_{in}(g)$ reduces while $E_{out}(g) - E_{in}(g)$ increases. Thus, a trade-off must be decided for optimising the learning of the model.

2 Feasibility of Learning

The condition, $\mathbf{E_{out}(g) \approx E_{in}(g)}$ is satisifed, if the probability distribution

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

is satisfied. Here, **M** is the number of non overlapping classifiers and, is ∞. Hence, the feasibility of learning is directly related to **M**.

Improving the value of **M** improves our learning. This can be achieved by considering a finite set of input points instead of the whole input space and counting the number of dichitomies.

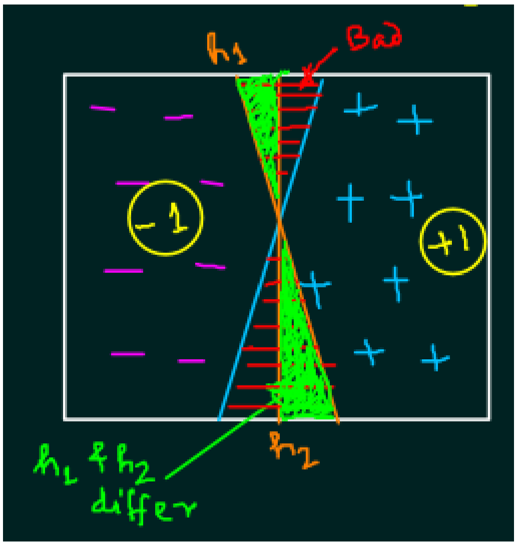


Figure 1: Image showing multiple hypotheses that classify the points correctly

In practice, there are not an infinite non overlapping hypotheses for a given constellation of points. This is shown in Figure 1, where both hypotheses h1 and h2 classify the points correctly. These hypotheses only contribute once in M instead of twice. This results in the reduction of training space from ∞ and offers an option to generalise the training.
In other words, we can reduce the hypothesis

$$\text{from } \mathcal{H}\{\mathbf{X}\} \to \{+\mathbf{1}, -\mathbf{1}\} \text{ to } \mathcal{H}\{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_N}\} \to \{+\mathbf{1}, -\mathbf{1}\}$$

While the number of hypotheses $\mathcal{H}$ can be ∞, the number of dichotomies $\mathcal{H}\{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_N}\}$ is at most 2^N. This result in a new function, $|\mathcal{H}(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_N})| \leq \mathbf{2^N}$.

3 Growth Function

The growth function $m_{\mathcal{H}}(N)$ is defined as the maximum number of dichotomies of a given training space while the points are arranged in the worst possible arrangement.

$$\text{i.e., } m_{\mathcal{H}}(N) = \max_{X_1, X_2, ... X_N \epsilon \mathcal{X}} |\mathcal{H}(X_1, X_2, ..., X_N)|$$

In Figure 2, the constellation in the left is the worst possible arrangement and the constellation in the right is not the worst possible arrangement.

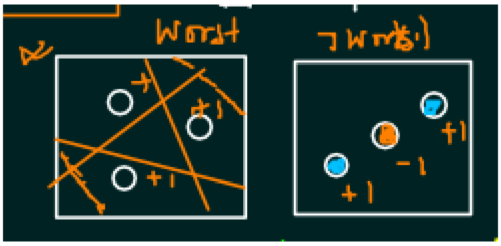


Figure 2: Different constellations possible for three points

This is because, if the points have alternating classes, it will be difficult to find a **single** line to split the points. The number of dichotomies for each type of constellation are given in the following section.

3.1 Growth functions of different constellations

Some types of constellations and their growth functions are:

1. 2D Perceptrons

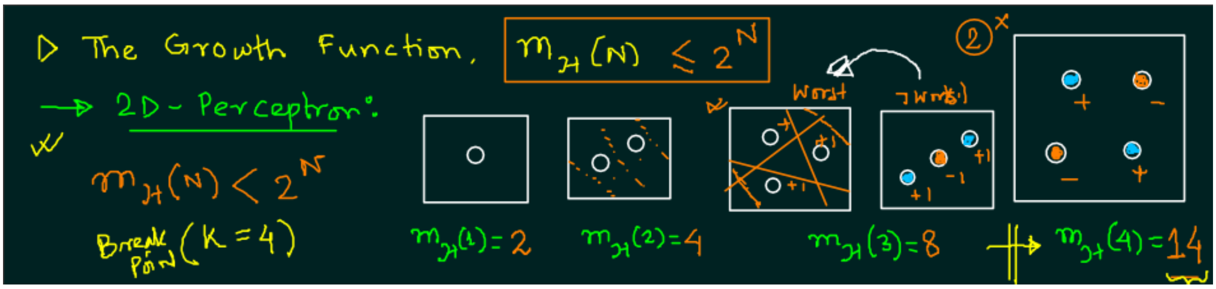


Figure 3: Arrangement of points for 2D Perceptrons

a) Consider the first image with a single point. This point can be classified as a +1 or a -1. So, the number of dichotomies is 2.

b) In the second image with two points, the points may both belong to the +1 class or the -1 class, or they may belong to different classes (both points can be a +1 or a -1). The total number of dichotomies is 4.

c) The third image shows the worst possible way to arrange three points and can be classified as, all points belonging to +1 or -1 (contributes 2), one point belonging to one class and the rest to another (contributes 6) resulting in the total dichotomy being 8

d) The fourth image shows an arrangement that cannot be split if the middle point belongs to a different class.This results in lesser than maximum number of dichotomies so, this arrangement is not considered.

e) The fifth image shows an arrangement of 4 points. All the points may belong to +1 or -1 class (contributes 2), one point belongs to +1 and the rest to -1 or vice versa (contributes 8) and two adjacent points belonging to +1 and the other -1 or vice versa (contributes 4). If the points are arranged in such a manner that opposite points belong to +1 or -1, the points cannot be split using a single line. Hence, the maximum number of dichotomies is 14.

2. Positive Rays and Intervals

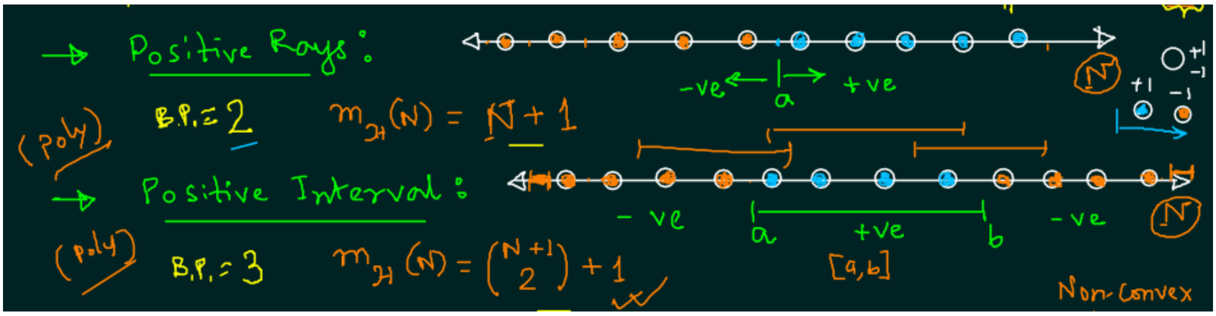


Figure 4: Possible dichotomies for positive rays and intervals

a) Positive rays are a dataset arranged on a line where, any point to the left of a classifier always belongs to -1 and all points to the right always belong to +1. For positive rays, the classifier can exist between all of the N points (contributes N-1) and additionally, all the points may belong to +1 or -1 class (contributes 2). The total dichotomy is N+1.

b) A positive interval is a dataset similiar positive intervals where, any point within an interval may belong to either +1 class or the -1 class (contributes $\binom{N+1}{2}$). Additionally, the points may not belong to either class (contributes 1). Thus the total dichotmy is $\binom{N+1}{2} + 1$.

3. Convex sets

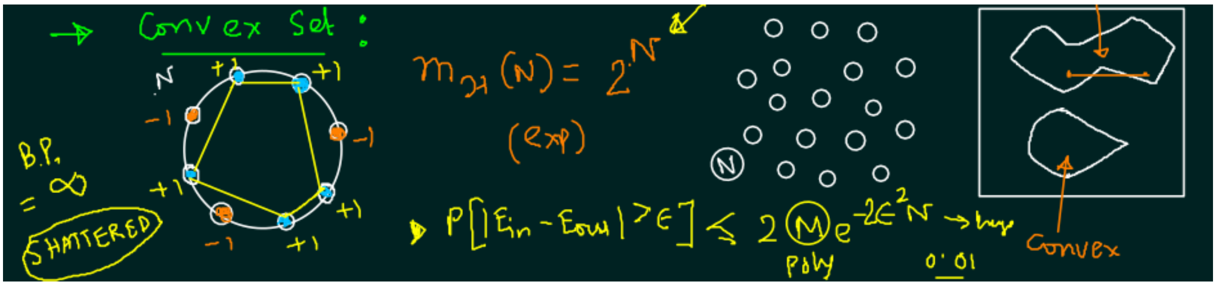

Convex Set :
N
+1
+1
-1
-1
+1
+1
-1
+1
B.P. = ∞
SHATTERED
$m_{\mathcal{H}}(N) = 2^N$
(exp)
$P[|E_{in} - E_{out}| > \epsilon] \leq 2 M e^{-2\epsilon^2 N}$
poly
0.01
N
Convex

Figure 5: Possible dichotomies for a convex set

a) If there exists a set with all points along the boundary of the set, such a set is called a convex set if there exists a straight line from all points to each other without crossing the boundary of the set. In Figure 5, the set on the top is not convex as the line has to cross the boundary.

b) The points are classified as all points which are connected by the lines belonging to one class and the rest to another. Thus, the dichotomy of such a set is always 2^N.

3.2 Break Point

If no dataset of size k can be shattered by $\mathcal{H}$ then, k is a break point for $\mathcal{H}$. Any dataset bigger than k cannor be shattered either. Thus, $m_{\mathcal{H}}(k) < 2^k$.

In other words, k is the minimum size of the dataset for which, $m_{\mathcal{H}}(k)$ is not equal to 2^k. Some example of break points are explained below:

1 Consider 2D perceptrons. From the previous section, we know that the perceptrons cannot classify the points for all arrangements of the number of points is 4 (and by extension, greater than 4). Therefore, the break point for the 2D perceptrons is 4.

2 For positive rays, if the dataset has a single point, we can classify it as belonging to either +1 class or -1 class (dichotomy is 2, which is equal to 2^1). However, for the dataset with more than 1 point, the number of dichotomies reduce to (k+1), which is smaller than 2^k. Hence the break point is 2.

3 In case of positive intervals, we can see that the expression derived in the previous sections is equal to 2^k while $k \epsilon \{1, 2\}$. Therefore, the break point is 3.

4 Convex sets always have dichotomy of 2^k. Therefore, their break point is ∞.

4 Proof that $\mathrm{m}_{\mathcal{H}}$(N) is a polynomial

The growth function, $m_{\mathcal{H}}(N)$ can be used to replace M if it can be proved that it is a polynomial function.

$m_{\mathcal{H}}(N)$ can be proved to be a polynomial if the following condition is satisfied:

$$m_{\mathcal{H}}(N) \leq \text{some quantity} \leq \text{some quantity} \leq \text{a polynomial}.$$

Let B(N,k) be the maximum number of dihotomies with N points and break point k. Since, $m_{\mathcal{H}}(N) = B(N, k)$, we can prove that $m_{\mathcal{H}}(N)$ is linear if B(N,k) is linear. Consider the example shown in Figure 6:

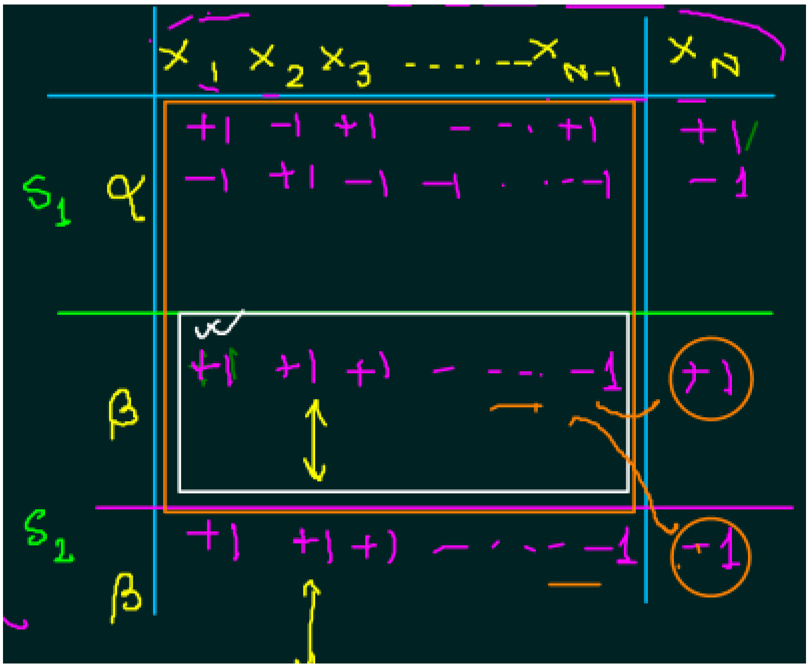


Figure 6: A sample dataset considered for proving B(N,k) is linear

There are N columns in the data labelled as $X_1, X_2, ...X_N$. Let X_N be the output class for this dataset. The subset S_1 has all the rows where X_N has either a +1 or a -1. Let the number of rows in S_1

be α.

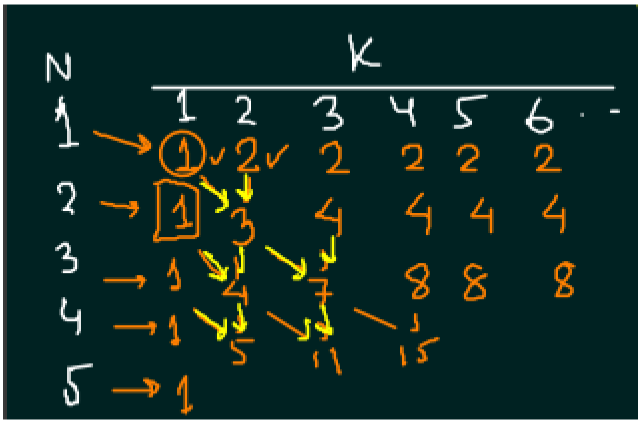
Let the subset S_2 have all the rows having only a +1 (subset S_2^+) or the rows having only a -1 (subset S_2^-). Let the number of rows in S_2^+ and S_2^- be β for both these subsets. Therefore, the total number of rows in subset S_2 is 2β. Therefore, the total number of dichotomies of this dataset is given by, $B(N,k) = \alpha + 2\beta$.

Let's consider the first N-1 columns (i.e., all columns excluding N^{th} column) and all rows of S_1 and S_2^+. These points are shattered by the break point k as we have just removed a unit from the data and the points still map to either +1 or -1 in X_N. This implies, $\alpha + \beta \leq B(N-1,k)$.

Consider only the S_2^+ rows and first N-1 columns. As the data map to only a +1, they have a break point of (k-1). Therefore, $\beta \leq B(N-1,k-1)$.

Now, we can replace the $\alpha + 2\beta$ term in B(N,k) by,

$$B(N,k) \leq (\alpha + \beta) + \beta \leq B(N-1,k) + B(N-1,k-1)$$



N	1	2	3	4	5	6 . -
1	1	2	2	2	2	2
2	1	3	4	4	4	4
3	1	4	7	8	8	8
4	1	5	11	15		
5	1					

Figure 7: Proof that $B(N,k) \leq B(N-1,k) + B(N-1,k-1)$

In the Figure 7, consider B(2,2). The value is 3 which, is equal to the sum of B(1,2)(2) and B(1,1)(1). Note that this distribution follows the Pascal triangle. This proves that $m_{\mathcal{H}}(N)$ is less than, or equal to some quantity.

To prove that $m_{\mathcal{H}}(N)$ is less than, or equal to a polynomial we assume,

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

where, $\sum_{i=0}^{k-1} \binom{N}{i}$ is the binomial distribution. This relation is proved by Mathematical Induction in the following section. Note that the third line is similiar to $B(N,k) \leq B(N-1,k) + B(N-1,k-1)$.

4.1 Proof by Mathematical Induction

$$
\begin{aligned}
\sum_{i=0}^{k-1}\binom{N}{i} &= \sum_{i=0}^{k-1}\binom{N-1}{i}+\sum_{i=0}^{k-2}\binom{N-1}{i} \\
&= 1+\sum_{i=1}^{k-1}\binom{N-1}{i}+\sum_{i=1}^{k-2}\binom{N-1}{i} \\
&= 1+\sum_{i=1}^{k-1}\binom{N-1}{i}+\sum_{i=1}^{k-1}\binom{N-1}{i-1} \\
&= 1+\sum_{i=1}^{k-1}\left[\binom{N-1}{i}+\binom{N-1}{i-1}\right] \\
&= 1+\sum_{i=1}^{k-1}\binom{N}{i} \\
&= \sum_{i=0}^{k-1}\binom{N}{i}
\end{aligned}
\tag{1}
$$

4.2 Examples of $\mathbf{m}_{\mathcal{H}}(\mathbf{N})$ in polynomial form

The polynomial form of the growth functions of examples discussed in previous sections are as follows:

1. **Positive rays:** the break point is 2.

$$\rightarrow m_{\mathcal{H}}(N) \leq \sum_{i=0}^{2-1}\binom{N}{i} = \sum_{i=0}^{1}\binom{N}{i} = 1+N$$

2. **Positive intervals:** the break point is 3.

$$\rightarrow m_{\mathcal{H}}(N) \leq \sum_{i=0}^{3-1}\binom{N}{i} = \sum_{i=0}^{2}\binom{N}{i} = 1+\frac{1}{2}N+\frac{1}{2}N^2$$

3. **2D Perceptrons:** the break point is 4.

$$\rightarrow m_{\mathcal{H}}(N) \leq \sum_{i=0}^{4-1}\binom{N}{i} = \sum_{i=0}^{3}\binom{N}{i} = 1+\frac{5}{6}N+\frac{1}{6}N^3$$

From these examples, we can conclude that the growth function is a polynomial function of **'N'** and is a suitable candidate to replace **'M'**. The proof that $m_{\mathcal{H}}(N)$ is a good replacement of M will be discussed in the next class.