CS60020: FADML

Lecture Scribe Notes

By: Siddhant Buriuly (24BM6JP51)

27th March 2025

Idea behind Theory of Generalization

You want a small E_{out}, but all you have access to during training is E_{in}.
So, the key question is:
Can we guarantee that $E_{\text{in}} \approx E_{\text{out}}$?

When you train a model using a finite dataset, you're fitting it to specific examples. **Generalization theory** tries to connect that *training performance (in-sample error)* to the *true performance on all data (out-of-sample error)*.

Feasibility of Learning

The condition $E_{\text{out}}(h) \approx E_{\text{in}}(h)$ is satisfied if the following bound holds:

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

Here, M is the number of non-overlapping hypotheses, often infinite. Therefore, the **feasibility of learning** is directly related to the complexity of the hypothesis set.

Can we replace M ***with*** $m_{\mathcal{H}}(N)$ ***?***

To make learning feasible, we reduce the hypothesis space from the infinite input space:

$$\mathcal{H}\{X\} \to \{+1, -1\} \quad \text{to} \quad \mathcal{H}\{x_1, x_2, \ldots, x_N\} \to \{+1, -1\}$$

This allows us to count the number of possible **dichotomies** instead of infinite hypotheses.

Dichotomy

A **dichotomy** is a way of labeling a set of N input points with $+1$ or -1 using a hypothesis from a hypothesis set $\mathcal{H}$. Each hypothesis assigns a label to every point, and the pattern of these labels forms one dichotomy.

If a hypothesis set can realize all 2^N possible labelings on N points, we say it **shatters** that set. The number of dichotomies that $\mathcal{H}$ can produce over N points is denoted as the **growth function** $m_{\mathcal{H}}(N)$. When $m_{\mathcal{H}}(N) < 2^N$, the hypothesis set's capacity is limited — and this limit connects to the concept of **VC dimension**.

Growth Function

The **growth function** $m_{\mathcal{H}}(N)$ is defined as the maximum number of dichotomies realizable by the hypothesis set $\mathcal{H}$ on any N input points arranged in the worst possible configuration:

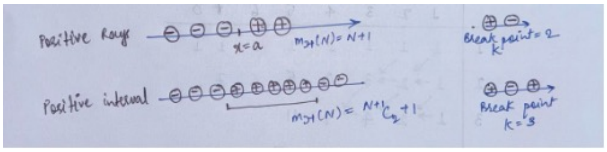
$$m_{\mathcal{H}}(N) = \max_{x_1,\ldots,x_N \in X} |\mathcal{H}(x_1, \ldots, x_N)|$$

Growth Function Examples

1. Positive Rays

Points are arranged along a 1D line. The classifier labels all points to the left of a threshold as -1, and to the right as $+1$. For N points:

$$m_{\mathcal{H}}(N) = N + 1$$

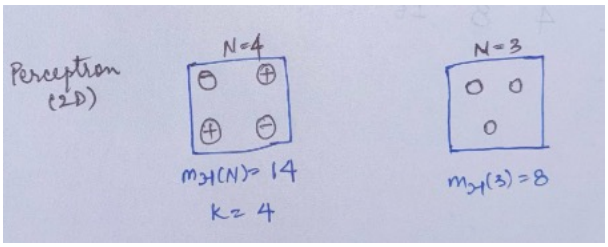


2. Positive Intervals

A contiguous interval is labeled $+1$ and the rest -1. Choosing two regions for interval endpoints gives $\binom{N+1}{2}$ dichotomies, plus one if both fall in the same region:

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1$$

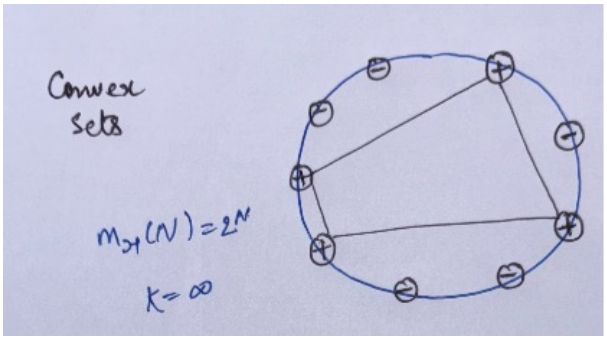
3. 2D Perceptrons



Points are separated by a line in 2D space:

- 1 point: 2 dichotomies
- 2 points: 4 dichotomies
- 3 points: 8 dichotomies
- 4 points: 14 dichotomies

4. Convex Sets



Convex sets always have 2^N dichotomies. Therefore, their break point is $k = \infty$.

Break Point

A **break point** k is the smallest number such that $\mathcal{H}$ cannot shatter every set of k points:

$$m_{\mathcal{H}}(k) < 2^k$$

Examples:

- Positive Rays: $k = 2$
- Positive Intervals: $k = 3$
- 2D Perceptrons: $k = 4$
- Convex Sets: $k = \infty$

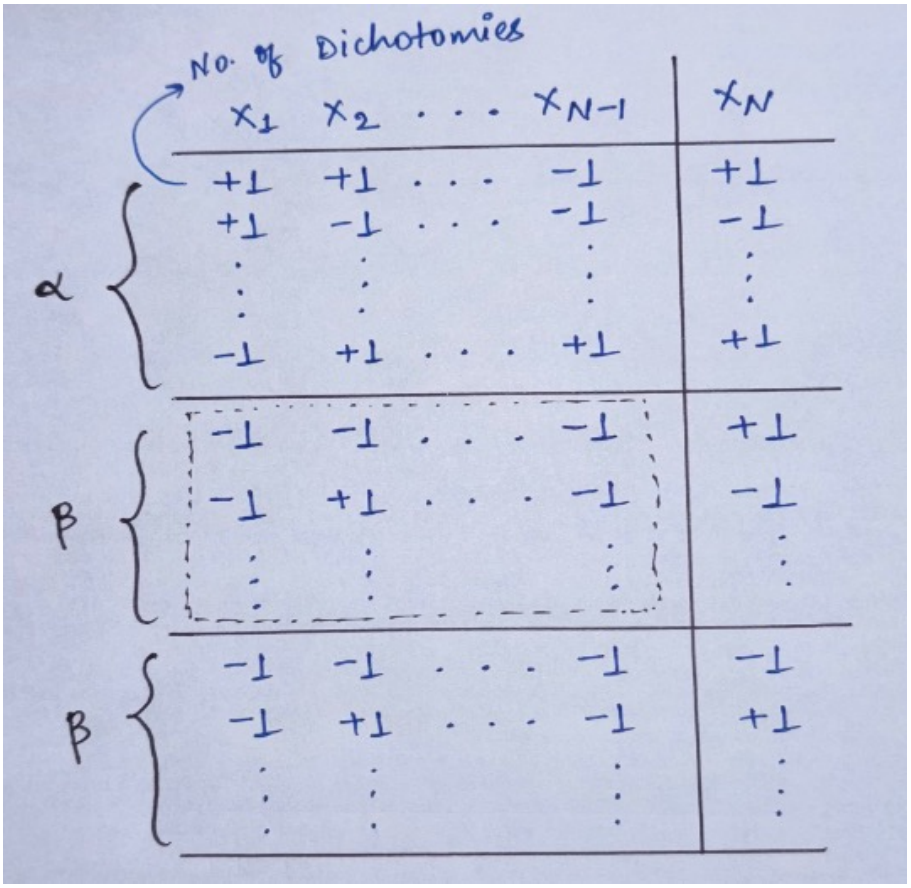
$m_{\mathcal{H}}(N)$ is a Polynomial Bounded

We want to replace M in Hoeffding's inequality with $m_{\mathcal{H}}(N)$, but only if it's bounded by a **polynomial**. It can be proved polynomial if:

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

Let $B(N, k)$ be the maximum number of dichotomies 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.

Let's consider the example:



α is the number of rows in first set and β is the no. of rows in second set, which has the same dichotomies on the first $N-1$ points as the third set except that the Nth column is reverse signed.

Therefore, the total number of dichotomies of this dataset is given by:

$$B(N,k) = \alpha + 2\beta$$

Consider the first $N-1$ columns and the first α and β rows. These are shattered with 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. So :

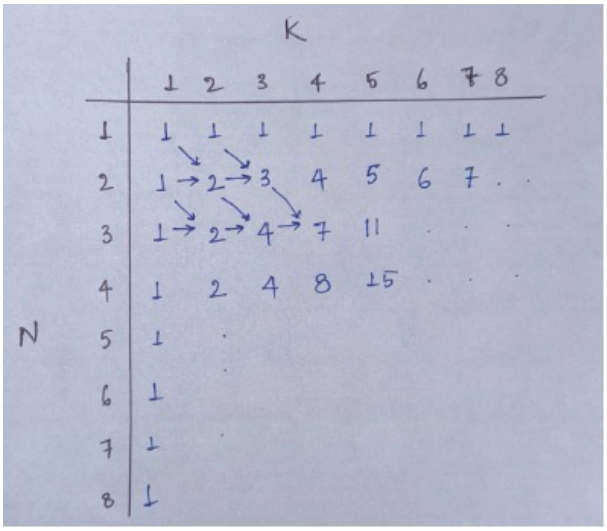
$$\alpha + \beta \leq B(N-1,k)$$

Considering only β rows and first $N-1$ columns, they have a break point of $(k-1)$. So :

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

Combining:

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



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

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

This relation can be proved by math induction.

Example for the polynomial growth function:

2D Perceptrons: $k = 3$

$$B(N,k) \leq \sum_{i=0}^{2} \binom{N}{i} = 1 + N + \frac{N(N-1)}{2}$$

Theory of generalization

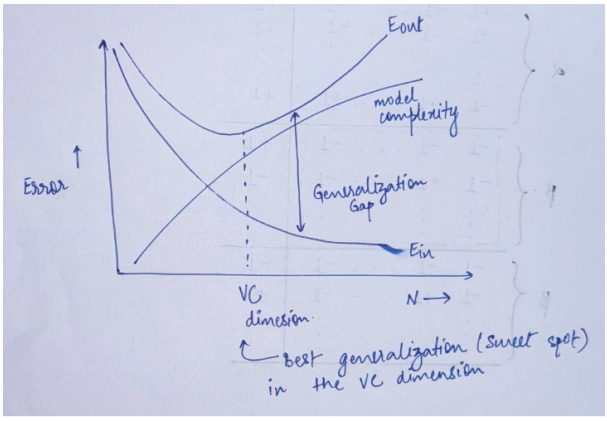
From Hoeffding Inequality

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

To VC Inequality

$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right] \leq 4m_{\mathcal{H}}(2N)e^{-1/8\epsilon^2 N}$$

(There is a long elaborated proof for this)



$$d_{\text{VC}} = k - 1$$

which tells the most number of points that can be shattered. Higher VC dimension means higher complexity of model, which means higher generalization.