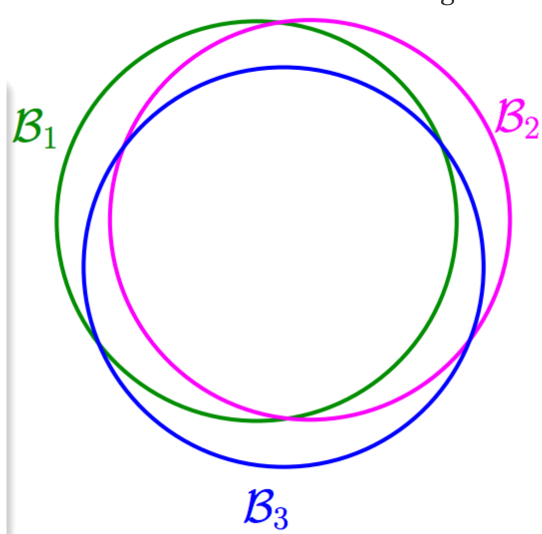
Week 5: Training vs. Testing

Table of Contents

Reducing Number of Hypothesis

1. Finite-bin Hoeffiding's Inequality implies that the probability of encountering bad samples in a multi-hypotheses setting is directly proportional to number of available hypotheses. In addition, perceptron learning algorithm by definition implies an infinite number of hypotheses(lines), or $M = \inf$, which violates the basic assumption of finite-bin Hoeffding's.
2. For similar hypotheses h_1 and h_2, their individual probabilities of encountering bad samples likely overlap
    - Bad sample $\rightarrow E_{out}(h_1) \approx E_{out}(h2)$
    - Similar hypotheses $rightarrow \mathcal{D}, E_{in}(h_1) = E_{in}(h2)$ for most samples
    - Union bound in finite-bin Hoeffding thus *over-estimates* size of error



3. In order to reduce M to a finite (and small) number, hypotheses shall be categorized by their labeling outcome.
    - Use number of categories (**effective number of lines**), in place of number of hypotheses M in finite-bin Hoeffding's

4. **Effective number of lines**: maximum kinds of lines with respect to N inputs $x_1, x_2, \ldots, x_N$
    - Must be $\leq 2N$
        - N points to be labeled {1, -1} → 2^N permutations at most, however some cases are *not linear separable*, thus further reducing effective number of lines provided by perceptron
    - finite 'grouping' of infinitely-many lines $\in \mathcal{H}$

Dichotomies

1. Dichotomies
    - Given points $x_1, x_2, \ldots, x_n$, a **dichotomy** is a hypothesis *limited to* the eyes of these points
    - $\mathcal{H}(x_1, x_2, \ldots, x_n)$ represents all dichotomies implemented by hypothesis set $\mathcal{H}$ on points $x_1, x_2, \ldots, x_n$

	hypotheses $\mathcal{H}$	dichotomies $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$
e.g.	all lines in $\mathbb{R}^2$	{○○○○, ○○○×, ○○××, ...}
size	possibly infinite	upper bounded by 2^N

2. Former definition of dichotomy
    - A **dichotomy** is a partition of a whole (or a set) into two parts (subsets). This couple of parts must be:
        - **Jointly exhaustive**: All contents (or points in the set) must be included in one part/subset *or* the other
        - **Mutually exclusive**: Nothing can belong simultaneously to both parts

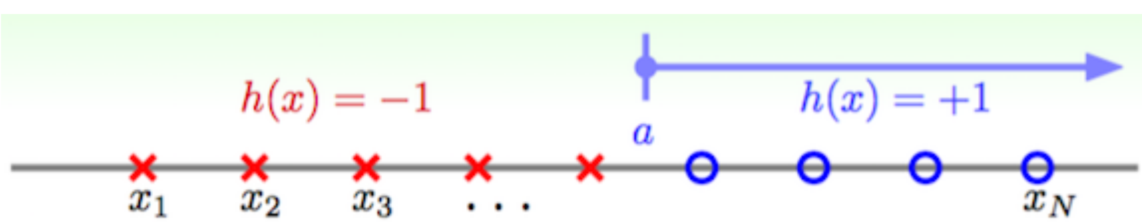
Growth Function

1.**Growth Function**: Remove dichotomy's dependency on inputs $(x_1, x_2, \ldots, x_n)$, by taking **max** of all possible $(x_1, x_2, \ldots, x_n)$

$$x_H(N) = \max_{x_1, x_2, \ldots, x_N \in \chi} |x1, x2, \ldots, x_N|$$

- Finite, **upper-bounded by** 2^N

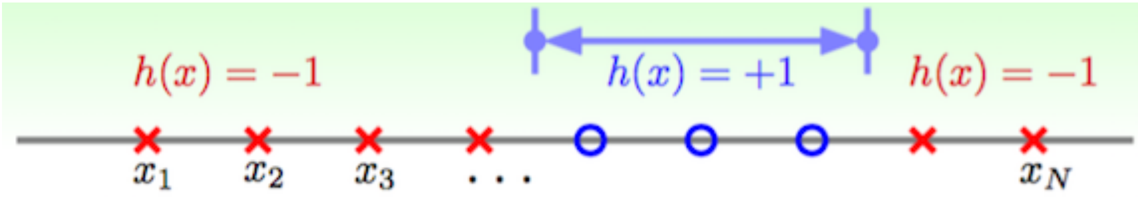
Growth Function of Positive Rays



1. Properties:
    - $\chi = \mathbb{R}$ (1-D)
    - Hypothesis set $\mathbb{H}$ contains h, where *each* $h(x) = sign(x - a)$ *for threshold a*
2. Growth function:
    - One dichotomy for $a \in$ each spot in range (x_n, x_{n+1})

- Growth function given N inputs: $m_{\mathcal{H}}(N) = N + 1$

Growth Function of Positive Interval



1. Properties:
   - $\chi = \mathbb{R}$ (1-D)
   - Hypothesis set $\mathbb{H}$ contains h, where $h(x) = +1$ *iff* $x \in [l, r)$, -1 *otherwise*
2. Growth function:

one dichotomy for each 'interval kind'

$$m_{\mathcal{H}}(N) = \underbrace{\binom{N+1}{2}}_{\text{interval ends in } N+1 \text{ spots}} + \underbrace{1}_{\text{all } \times}$$

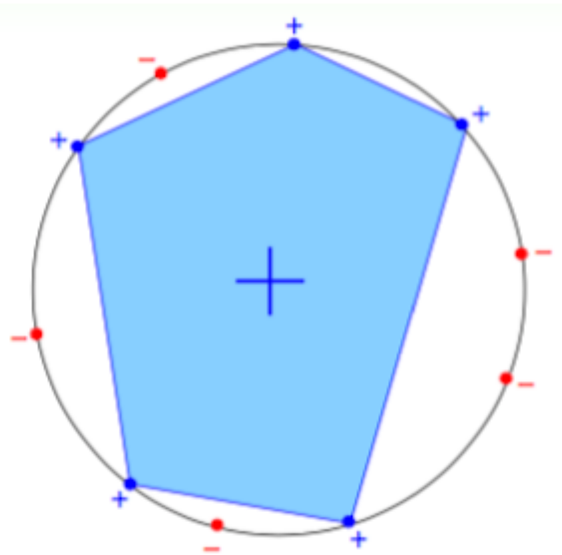
$$= \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

Growth Function of Convex Set

1. Properties:
   - $\chi = \mathbb{R}^2$ (2-D)
   - Hypothesis set $\mathbb{H}$ contains h, where $h(x) = +1$ *iff x in a convex pre-defined convex region, -1*

otherwise

2. Growth function:
   - Imagine the set of inputs $x_1, x_2, \ldots, x_N$ placed on a circle
   - **Every** dichotomy can be implemented by $\mathcal{H}$ using a convex region **slightly extended** from *contour of positive inputs.*
   - The N inputs are **shattered** by $\mathcal{H}$
   - Growth function given N inputs: $m_\mathcal{H}(H) = 2^N$

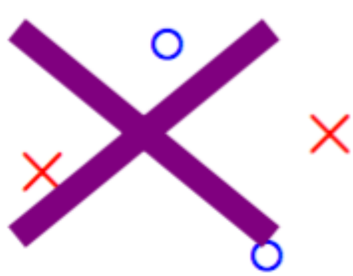


Growth Function of 2-D Perceptron

- By definiton, $m_\mathcal{H}(N) \leq 2^N$
- If input points form convex set $m_\mathcal{H}(N) = 2^N$ as seen above
- Otherwise, $$m_\mathcal{H}(N)*iscase − by − caseand*m_\mathcal{H}(N) < 2^N$ is some cases

Break Point

1. Recall the definition of *shatter* from above (every dichotomy can connect all **positive inputs** to form **convex** region). For 2-D perceptrons, it is possible to shatter up to 3 points, but not 4 points



2. If no k inputs can be shattered by $\mathcal{H}$, then k is a **break point** for hypothesis set $\mathcal{H}$

   - 2D perceptron hypothesis set *mathcalH* with break point k is guarantteed to have growth function $m_\mathcal{H}(k) < 2^k$
     - Vice versa, growth function for hypothesis set without break point would be $m_\mathcal{H}(k) = 2^k$
   - If k is a break point, $k + 1, k + 2, \ldots$ are also break points, but we are mostly interested in the *minimum break point k*

- positive rays: $m_{\mathcal{H}}(N) = N + 1 = O(N)$
  break point at 2
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 = O(N^2)$
  break point at 3
- convex sets: $m_{\mathcal{H}}(N) = 2^N$
  no break point
- 2D perceptrons: $m_{\mathcal{H}}(N) < 2^N$ **in some cases**
  break point at 4