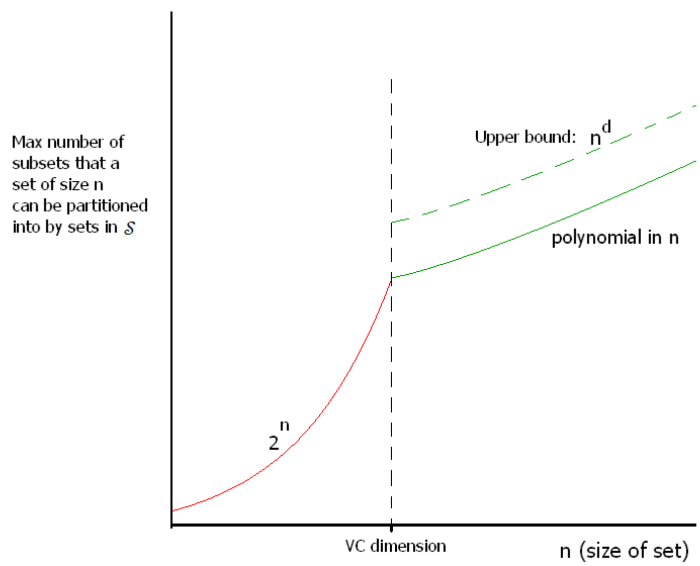
CS 4850: Mathematical Foundations for the Information Age

Lecture #34 - April 13, 2009

Scribes: Scott Rogoff and Andrew Owens

Primal Shatter Function $\Pi_{\mathcal{S}}(n)$

For a set system $\mathcal{S} = (X, \mathcal{S})$ of VC-dimension d, we know that a set A of ≤ d points can be shattered into all 2^n unique subsets by intersection with sets in $\mathcal{S}$. What is the maximum number of unique subsets that can be obtained for a set A of size n > d? The number of possible subsets as a function of n is known as the *Primal Shatter Function* and is denoted by $\Pi_{\mathcal{S}}(n)$. The function will have value 2^n until n = d, at which point it will grow at some slower pace.



Example: Consider the case of rectangles. We can shrink any rectangle until all of its sides contain a point inside, and consider this to be a representative canonical rectangle for any rectangle that contains the same points. A rectangle is defined by at most 4 points that define its edges. Therefore we can choose from at most n points to represent the left side of the rectangle, the right side, etc. Thus there are at most n^4 unique canonical rectangles.

We claim that for a set system of finite VC-dimension, the function representing the number of subsets that can be obtained is polynomial in n. More specifically:

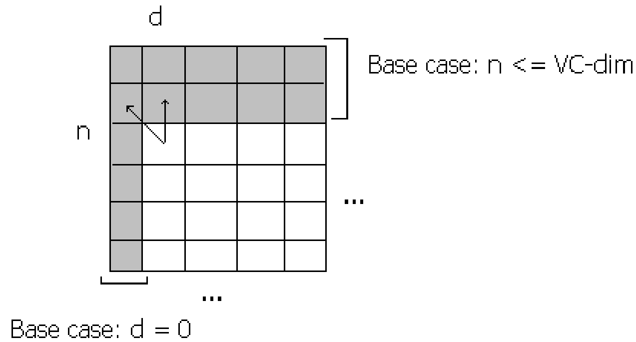
Lemma: For any set system $\mathcal{S} = (X, \mathcal{S})$ with VC-dimension d, an upper bound on $\Pi_{\mathcal{S}}(n)$ is given by

$$\Pi_{\mathcal{S}}(n) \le \sum_{i=0}^{d} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$$

Proof: We will prove this by induction on n, and then for fixed n by induction on d. Each entry of the table (n,d) will depend on the value of (n-1,d) and (n-1,d-1). Thus we need to fill in the upper part of the chart and the left column as base cases.

Base Case for n (n ≤ d): We know that $\Pi_{\mathcal{S}}(n) = 2^n$ because our set system can shatter n points. Further, we have $\sum_{i=0}^{d}\binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$, so this base case holds.

Base Case for d (d = 0): We are attempting to determine the value of $\Pi_{\mathcal{S}}(k)$ where $\mathcal{S}$ has a VC-dimension of 0. What does this mean? For one thing, a set system with VC-dimension 0 must have at most one set. If it had two different sets C and D, then one of those sets must have 1 point x which is not in the other. Then we could pick that point x and the sets C and D would shatter it, meaning the set system at least has VC-dimension 1. So a set system of VC-dimension 0 can at most only shatter the empty set. We have $\Pi_{\mathcal{S}}(n)$ = 1 and $\sum_{i=0}^{0}\binom{n}{i} = \binom{n}{0}$ = 1, so this base case holds.



Inductive Step: To bound $\Pi_{\mathcal{S}}(n)$ for the set system $\mathcal{S} = (X, \mathcal{S})$, we can remove one element x from X and consider $\Pi_{\mathcal{S}_1}(n-1)$ for the system

$$\mathcal{S}_1 = (X - \{x\}, \mathcal{S})$$

Let S be some element of $\mathcal{S}$ that does not include x (we will write S ∪ {x} if we want to include it). There are two cases to consider:

Case 1 – exactly 1 of S and S ∪ {x} is in $\mathcal{S}$: Here, we will have exactly 1 of S ∩ A or S ∪ {x} ∩ A as one of our partitions. We can identify either one of these with the set S ∩ A_1 in $\mathcal{S}_1$, where A_1 is the set A without the point x. This says that $\Pi_{\mathcal{S}}(n)$ is equal to $\Pi_{\mathcal{S}_1}(n-1)$. Since $\Pi_{\mathcal{S}_1}(n-1) \leq \sum_{i=0}^{d}\binom{n-1}{i}$ by the inductive hypothesis, and $\sum_{i=0}^{d}\binom{n-1}{i} \leq \sum_{i=0}^{d}\binom{n}{i}$, it must be that $\Pi_{\mathcal{S}}(n) \leq \sum_{i=0}^{d}\binom{n}{i}$ as desired.

Case 2 – both S and S ∪ {x} are in $\mathcal{S}$: In this case, S and S ∪ {x} define distinct subsets in the set system $\mathcal{S}$, but both define the same subset in $\mathcal{S}_1$. This tells us that $\Pi_{\mathcal{S}_1}(n-1)$ and $\Pi_{\mathcal{S}}(n)$ differ by the cardinality of the set $\{X - \{x\} \cap S \mid both\ S\ and\ S \cup \{x\}\ are\ in\ \mathcal{S}\}$. We will define the set

$$\mathcal{S}_2 = (X - \{x\}, \{S \mid S\ and\ S \cup \{x\}\ are\ in\ \mathcal{S}\})$$

Then we have the following recurrence relation:

$$\Pi_{\mathcal{S}}(n) = \Pi_{\mathcal{S}_1}(n-1) + \Pi_{\mathcal{S}_2}(n-1)$$

We know bounds on the latter two terms by the inductive hypothesis, so we just have to add them together.

Claim: $\mathcal{S}_1$ has VC-dimension ≤ d.

To see this, suppose that the VC-dimension is > d. Then there exists some set A, |A| > d that can be shattered by $\mathcal{S}$, which is a contradiction.

Claim: $\mathcal{S}_2$ has VC-dimension ≤ d-1.

To see this, note that if A - {x} is shattered in $\mathcal{S}_2$ then A is shattered in $\mathcal{S}$. If $\mathcal{S}_2$ had VC-dimension > d-1, then $\mathcal{S}$ would have VC-dimension > d, a contradiction.

By the inductive hypothesis, $\Pi_{\mathcal{S}_1}$(n-1) $\leq \sum_{i=0}^{d} \binom{n-1}{i}$ and $\Pi_{\mathcal{S}_2}$(n-1) ≤ $\sum_{i=0}^{d-1} \binom{n-1}{i}$.

Thus

$$\Pi_{\mathcal{S}}(\text{n}) = \Pi_{\mathcal{S}_1}(\text{n-1}) + \Pi_{\mathcal{S}_2}(\text{n-1}) \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i}$$

$$= \left[\binom{n-1}{0} + \binom{n-1}{1} + \cdots + \binom{n-1}{\text{d}-1}\right] + \left[\binom{n-1}{0} + \binom{n-1}{1} \ldots + \binom{n-1}{\text{d}}\right]$$

$$= \binom{n-1}{0} + \left[\binom{n-1}{0} + \binom{n-1}{1}\right] + \ldots \left[\binom{n-1}{\text{d}-1} + \binom{n-1}{\text{d}}\right]$$

$$= \sum_{i=0}^{d} \binom{n-1}{i} + \binom{n-1}{i-1} = \sum_{i=0}^{d} \binom{n}{i}$$

We use the fact $\binom{n-1}{i} + \binom{n-1}{i-1} = \binom{n}{i}$. Combinatorial proof: to choose a subset T of size i from a set S of size n, pick an element x. If x is in T, then there are $\binom{n-1}{i-1}$ ways to choose the remaining i - 1 elements. If x is not in T, then there are $\binom{n-1}{i}$ ways to choose the i elements from S - {x}.