Data Mining **Winter semester, 2002**

Lecture 6: November 19, 2002

Lecturer: Amos Fiat *Scribe: Roi Yehoshua and Kobi Lindzen*

6.1 The Vapnik-Chervonenkis Dimension

6.1.1 Preface

In the previous lecture we saw lower bound to the number of examples needed to be drawn in order to create an hypothesis which will meet the PAC model criteria ($m \leq O((\frac{1}{\epsilon}(\log(C) + log(\frac{1}{\delta})))$. Clearly this bound is meaningful only in the case where C is a concept class of finite cardinality, still, in the rectangles example we were able to find a PAC model though the size of C was infinite. In this lecture we will show when can a non-trivial infinite concept classes be learned from a finite sample.

6.1.2 Definition of the VC Dimension

As we defined in previous lectures, let X be a set called the **instance space** ($\Re^n$, $\{0,1\}^n$, etc.). Let C be a concept class over X, and both C and X may be infinite. Let $S \subseteq X$ be a finite set of elements.
Definition: For any concept class C over X, and any $S \subseteq X$,

$$\Pi_C(S) = \{c \cap S : c \in C\}.$$

By setting an arbitrary order for S elements we can say that if $S = \{x_1, x_2, \ldots, x_m\}$ then we can think of $\Pi_C(S)$ as the set of vectors $\Pi_C(S) \subseteq \{0,1\}^m$ defined by

$$\Pi_C(S) = \{(c(x_1), c(x_2), \ldots, c(x_m)) : c \in C\}.$$

Where $c(x_1) = 1$ if $x_1 \in c$ (and otherwise 0). 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. According to the second definition of $\Pi_C(S)$ it is obvious that $\Pi_C(S) \leq 2^m$.
Definition: if $\Pi_C(S) = \{0,1\}^m$ (where $m = |S|$), then we say that S is **shattered** by C. In other words S is shattered if every subset of S is correlated with some concept in C.

6.1.3 Why shattering ?

Assume we have a learning problem with $C = \{c_1, \ldots\}$ and a set S of examples with known labels. If S is shattered by C there is a compatible concept to every assignment of S, thus, we will not be able to predict any data about the correct concept. This leads us to our key definition.

Definition: VCD of C is the maximum size of a set shattered by C.

$$VCD(C) = \max\{|S| : |\Pi_C(S)| = 2^{|S|}\}.$$

If arbitrary large finite set can be shattered by C, then $VCD(C) = \infty$. We can say that the VCD is a measure of the complexity of a class concept family.

6.2 Examples of geometric concepts and VC Dimension

Intervals of the real line

Let X be all the points on the real line, C are all the interval $[p, q]$ (infinite number of concepts) and we would like to find $VCD(C)$. if $S = \{a, b, c\}$ and the labels are as shown in Figure 6.1 then the subset $\{a, c\}$ is not the result of any intersection between S and $c \in C$. $\Rightarrow$ $|\Pi_C(S)| < 2^3$. $\Rightarrow$ S is not shattered by C. $\Rightarrow$ $VCD(C) < 3$ (a, b, c are general points). For this concept class any set of two points can be shattered (see Figure 6.2). Thus the VC dimension for this class is two.

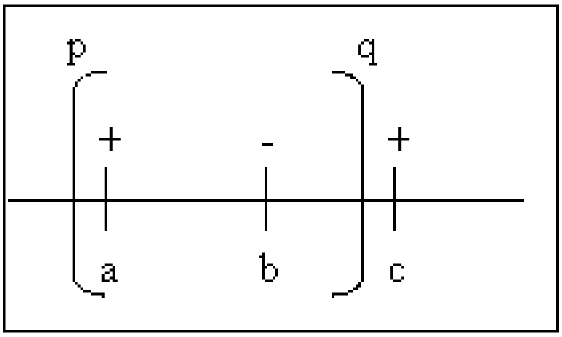

Figure 6.1: Intervals and three points in X. This set can not be shattered by any $c \in C$ due to the subset $\{a, c\}$.

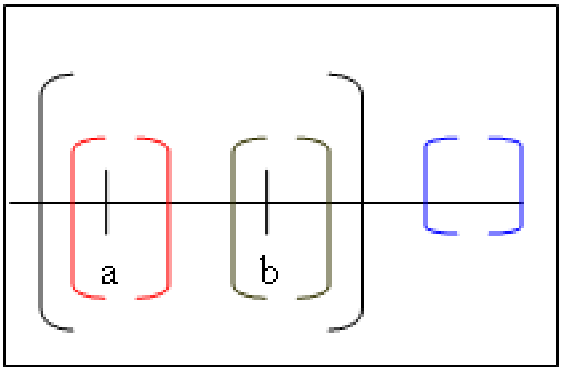

Figure 6.2: Intervals and two points in X. Each subset of $S = \{a, b\}$ can be realized by $c \in C$.

Linear halfspaces in the plane

Let $X = \Re^2$ and $C = \{ax + by \geq c : a, b, c \in \Re\}$ (halfspaces). The size of the concept class set is infinite (as well as the instant space size).
Claim: $VCD(C) = 3$.
In Figure 6.3 we can see that no concept (halfspace) in C can induce the labelling on the four points, still we have to show that **all** sets with size four can not be shattered. If three (out of four) points are collinear it is obvious that they can not be shattered. If they are not collinear one can say that three points creating a triangle where the forth point is either inside or outside this triangle (see Figure 6.4). In both cases, the concepts can not realize all possible labelling, thus, $VCD(C) < 4$. On the other hand, any three points (that are not collinear) can be shattered. Figure 6.5 shows how one labelling out of the possible eight labelling can be realize by a halfspace. All the subset of S can be realized, thus, $VCD(C) \geq 3$. As a result of those two observations we can say now that $VCD(C) = 3$.
In general, for halfspace in $\Re^d$, the VC dimension is $d + 1$.

Axis-aligned rectangles in the plane

Let C be the family of all axis-aligned rectangles in $\Re^2$.
Claim: $VCD(C) = 4$, meaning, one or more sets with size four exist and can be shattered by C. Figure 6.6 shows such set. For every set of five points in the plane , there must be a point that is neither the extreme left, right, top or bottom point of the five. if we label this non-extermal point negative and the remaining fourextermal points positive, no rectangle can realize this labelling since every rectangle that includes the extermal points must also include the internal point (see Figure 6.7). Thus the VC dimension of C is four.

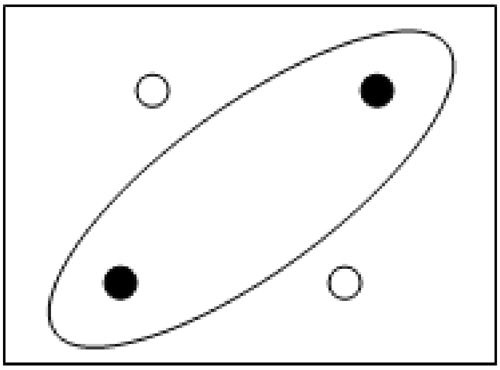

Figure 6.3: Halfspace and four points in X. No halfspace can separate the positive samples (black dots) from the negative samples.

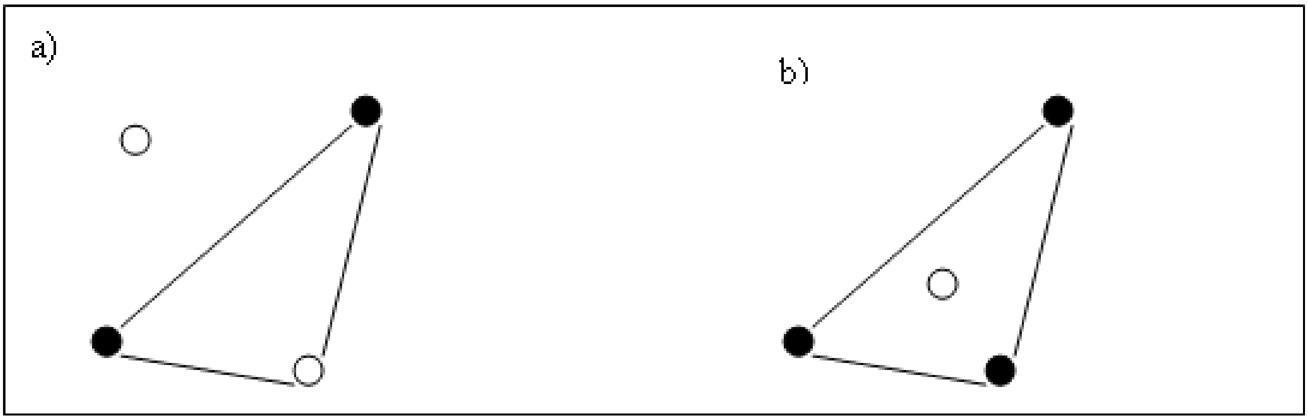

Figure 6.4: General four points in X. Figure a) represents the case where the forth point is outside the triangle, no halfspace can separate the positive samples from the negative samples. Figure b) represents the case where the forth point is inside the triangle, once again, no halfspace can separate this sample from the rest.

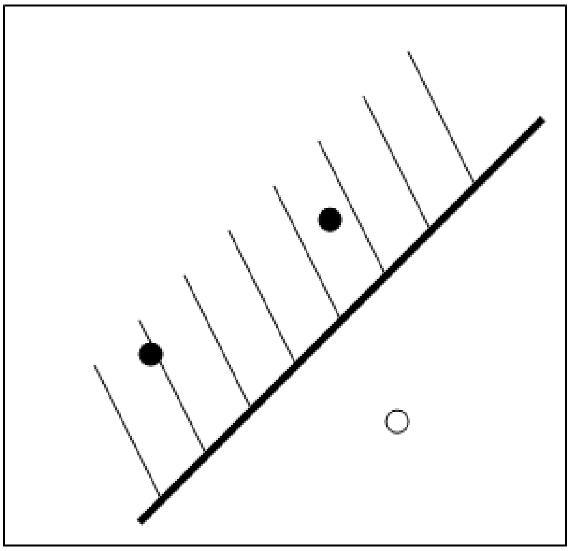

Figure 6.5: Halfspace and three points in X. A concept that realize one optional labelling of this set, it can easily be seen that all combinations can be realized by halfspaces.

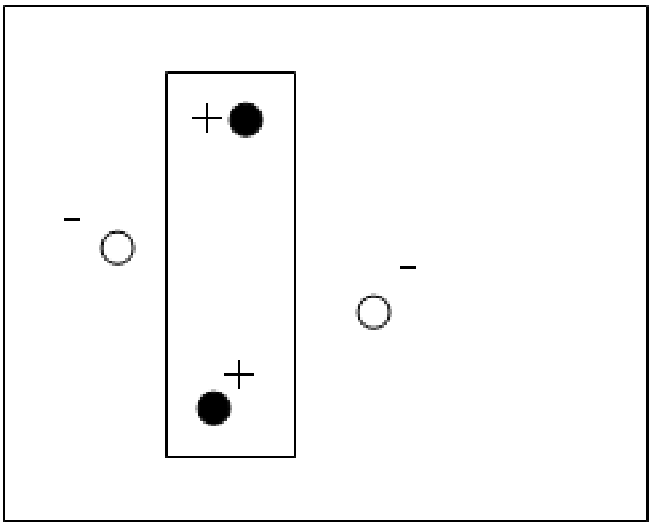

Figure 6.6: Rectangles and four points in X. Each labelling of S can be realized by $c \in C$.

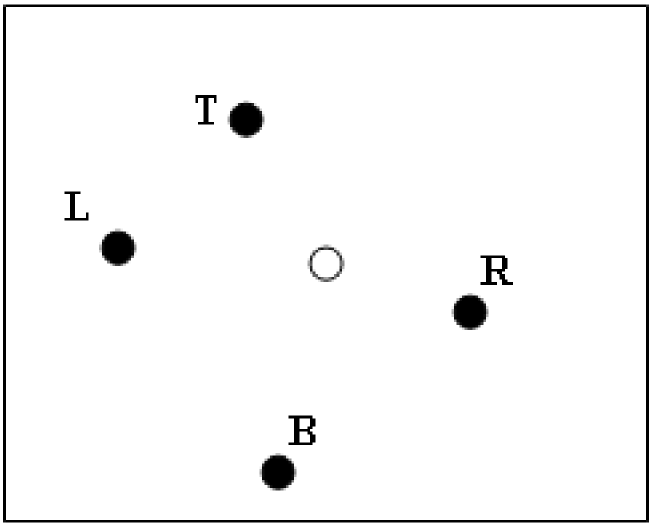

Figure 6.7: Rectangles and five points in X. Unrealizable labelling by axis-aligned rectangles (black dots are positive examples).

6.3 A Polynomial Bound on $|\Pi_C(S)|$

Definition: For any natural number m we define

$$\Pi_C(m) = \max\{|\Pi_C(S)| : |S| = m\}.$$

The function $\Pi_C(m)$ can be thought of as a measure of the complexity of the concept class C: how many behaviors on sets of m points can be realized by C.

By the definition of the VC dimension of C:

$$VCD(C) = \max\{|S| : |\Pi_C(S)| = 2^{|S|}\}.$$

Thus we can clearly see that:

$$\begin{array}{ll} \Pi_C(m) = 2^m & m \leq VCD(C) \\ \Pi_C(m) < 2^m & m > VCD(C) \end{array}$$

In this section we will prove a surprising result, that despite the fact that we might naively expect $\Pi_C(m)$ to grow as an exponential function of m, m grows exponentially until it reaches $VCD(C)$ and then it is bounded by a polynomial in m of degree that equals $VCD(C)$, i.e.:

$$\Pi_C(m) \leq m^{VCD(C)} \quad m > VCD(C)$$

This result suggests that to obtain an algorithm for PAC learning C we need to take a number of examples on the order of the VC dimension of C. The intuition to this is that if the number of examples is $m \leq VCD(C)$ then there are many intersections between concepts in C and the set of examples S, and thus it could be difficult to distinguish between the concepts. However, if $m > VCD(C)$ then the examples will be consistent with only a small part of the concepts in C, and thus it will be easier to distinguish between them and find out which is the true concept.

Definition: We define the function $\Phi_d(m)$ recursively by:

$$\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1)$$

with the initial conditions:

$$\Phi_d(0) = \Phi_0(m) = 1 \ \ \forall d, \forall m.$$

Solving this recursion (not detailed here) gives:

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} = \begin{cases} 2^m & m \leq d \\ O(m^d) & m > d \end{cases}$$

That is, the function grows exponentially with m until m reaches d and then it grows polynomially with m (for fixed d).

Lemma: If $VCD(C) = d$, then for any m, $\Pi_C(m) \leq \Phi_d(m)$.

Proof: The proof is by induction on both d and m. For the base cases, the lemma is easily established when $d = 0$ and m is arbitrary, and when $m = 0$ and d is arbitrary. We assume for induction that for all $m' \leq m$ and $d' \leq d$, with at least one of the inequalities is strict, we have $\Pi_C(m') \leq \Phi_{d'}(m')$. We now show that this inductive assumption establishes the desired statement for d and m.

Let S be any set of size m, and we will show that $|\Pi_C(S)| \leq \Phi_d(m)$. Let $x \in S$ be a point in S. Since $S - \{x\}$ is a set of size $m-1$ we have by induction assumption

$$|\Pi_C(S - \{x\})| \leq \Phi_d(m-1). \tag{6.1}$$

The difference between $\Pi_C(S)$ and $\Pi_C(S - \{x\})$ is that pairs of distinct sets in $\Pi_C(S)$ that differ only on their labelling of x are collapsed to a single representative in $\Pi_C(S - \{x\})$. Thus let us define:

$$C' = \{c \in \Pi_C(S) : x \notin c, c \cup \{x\} \in \Pi_C(S)\}.$$

Then $|C'|$ counts the number of pairs of sets in $\Pi_C(S)$ that merge into the same set in $\Pi_C(S - \{x\})$. Now we notice that $C' = \Pi_{C'}(S - \{x\})$, since by the definition of C' each set in C' is a subset of $S - \{x\}$, and intersecting a subset with its superset gives the same subset. Therefore we can write the equality:

$$|\Pi_C(S)| = |\Pi_C(S - \{x\})| + |\Pi_{C'}(S - \{x\})|. \tag{6.2}$$

We now show that $VCD(C') \leq d - 1$. This is derived from the observation that if $S' \subseteq S - \{x\}$ is shattered by C', then $S' \cup \{x\}$ is shattered by C. Now since $VCD(C) = d$, we must have $|S'| \leq d-1$, because there is not any set of size bigger than d that is shattered by C. Hence $VCD(C') \leq d - 1$. Now by the induction assumption we have

$$|\Pi_{C'}(S - \{x\})| \leq \Phi_{d-1}(m-1). \tag{6.3}$$

¿From 6.1, 6.2 and 6.3 we get:

$$|\Pi_C(S)| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m).$$

which concludes the proof of the lemma. □

Thus we have shown the polynomial bound $\Pi_C(m) = O(m^{VCD(C)})$ for $m > VCD(C)$.

6.4 ϵ-Nets

Let us now fix the true concept $c \in C$, and define the relative error to c as

$$\Delta(c) = \{c\Delta c' : c' \in C\}.$$

where $c\Delta c'$ is the symmetric difference between the sets c and c', i.e. $c\Delta c' = c \cup c' - c \cap c'$. Getting examples in the set $c\Delta c'$ can help us distinguish between the concepts c and c'.

Claim: $VCD(C) = VCD(\Delta(c))$.

Proof: For any set S we can define the following mapping between the elements of $\Pi_C(S)$ and $\Pi_{\Delta(c)}(S)$:

$$c' \in \Pi_C(S) \to c'\Delta(c \cap S) \in \Pi_{\Delta(c)}(S).$$

First we will show that this mapping is well defined, i.e. that if $c' \in \Pi_C(S)$ then $c'\Delta(c \cap S) \in \Pi_{\Delta(c)}(S)$. If $c' \in \Pi_C(S)$, then $c' \subseteq S$, and thus $c' = c' \cap S$. Therefore we can write:

$$c'\Delta(c \cap S) = (c' \cap S)\Delta(c \cap S) = (c'\Delta c) \cap S$$

where the last equality can be easily proven for every three sets c', c and S. Now we can see that $(c'\Delta c) \cap S = z \cap S$ where $z \in \Delta(c)$, hence $c'\Delta(c \cap S) \in \Pi_{\Delta(c)}(S)$.

We now show that this mapping is bijective. Each element in $\Pi_{\Delta(c)}(S)$ has the form $(c'\Delta c) \cap S$. As before, we can write $(c'\Delta c) \cap S = (c' \cap S)\Delta(c \cap S)$. So we can map this element to $c' \cap S \in \Pi_C(S)$, and this mapping is exactly the inverse of the mapping defined above.

Since we have found a bijective mapping between $\Pi_C(S)$ and $\Pi_{\Delta(c)}(S)$, we can conclude that $|\Pi_C(S)| = |\Pi_{\Delta(c)}(S)|$. Since this holds for any set S, $VCD(C) = VCD(\Delta(c))$ follows. $\square$

We may further refine the definition of $\Delta(c)$ to consider only those error regions with weight at least ϵ under the fixed distribution D. Those regions are most interesting, since failing to recognize such a region means failing in distinguishing between the true concept c and a concept h which is at least ϵ-different from c (in terms of distribution D). Thus, let

$$\Delta_\epsilon(c) = \{r \in \Delta(c) : Prob_{x \in D}[x \in r] \geq \epsilon\}.$$

We can now make the following definition:

Definition: A set S is an ϵ-net for $\Delta(c)$ if for every region r in $\Delta_\epsilon(c)$ we have $r \cap S \neq \phi$.

An ϵ-net for $\Delta(c)$ is thus a set that hits all the ϵ-heavy regions of $\Delta(c)$. The important property of ϵ-nets is that if the sample S drawn by the learning algorithm forms an ϵ-net for $\Delta(c)$, and the learning algorithm outputs a hypothesis $h \in C$ that is consistent with S, then we are guaranteed that this hypothesis must have error less than ϵ.

As an example, suppose $X = [0, 1]$ and D is the uniform density on X. Suppose that the concept class C consists of all closed intervals $[a, b]$ where $0 \leq a < b \leq 1$, as well as the empty set ϕ, and the target concept is $c = \phi$. The symmetric difference between any concept $h \in C$ and c is h itself, thus $\Delta(c) = C$. For any interval I under the uniform density, $Pr_{x \in D}[x \in I]$ is just the length of I. Thus $\Delta_\epsilon(c)$ consists of all the the intervals whose length is equal or greater than ϵ. Therefore the set of points $S = \{i\epsilon | i = 1, 2, ..., 1/\epsilon\}$ is an ϵ-net for $\Delta(c)$.

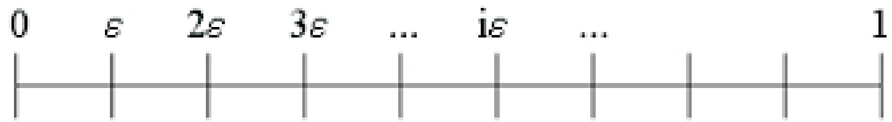

Figure 6.8: An ϵ-net in the interval $[0, 1]$.

We can easily see that each interval in $[0, 1]$ which has length equal or greater than ϵ is hit by a point in S, hence S is an ϵ-net for $\Delta(c)$.

6.5 References

1. Michael J.Kearns, Umesh V.Vazirani, "An introduction to Computational Learning Theory", MIT Press, 1997.