CSE 513T Theory of Artificial Intelligence & Machine Learning February 10, 2015

Lecture 9

Lecturer: Brendan Juba *Scribe: Golnoosh Dehghanpoor, Samuel Shields*

1 Overview

1. Definition of VC-Dimension
2. Examples of VC-Dimension
3. No Free Lunch Theorem.

2 VC dimension

"Cliff's Notes" version of VC-dimension: VCD is the parameter of a class of representations that tightly characterizes the number of examples needed to learn.

Definition - Dichotomies: For any class of Boolean representations C, over an instance space X, and any S $\subset$ X, are different labelings of S that can be realised by C:

$\pi_C \left\{x^1, ..., x^m\right\} = \left\{C(x^1), ..., C(x^m) : c \in C\right\}$

an equivalent definition is:

if c $\in$ C with $\{x \in X : C(x) = 1\}$
$\pi_C(s) = \{s \cap c : c \in C\}$
if $\pi_C(s) = (0,1)^m$, (that is all possible labelings of s are realizable by c), we say that S is shattered by C.

Definition - The Vapnink-Chervonenkis (VC) Dimension of a representation class C, VCD(C), Is the size of the largest set S that can be shattered by C. If sets of all finite sizes are shattered, then we write $VCD(C) = \infty$

3 VC dimension examples

3.1 Arbitrary Function

Let C be the class of all concepts on Q. (functions such that $c : Q \to \{0,1\}$) Then the VC dimension of this class is infinite, since it can shatter all sets of finite sizes.

3.2 Intervals on the Real Number Line "$\Re$"

In this example C is an interval on the real line. any member of this class is represented by two parameters: $\{a, b\} \in \Re$, which represents an interval between a and b. To determine the VC dimension we need to find the smallest set of size s that C can't label all of it's members (VCD would be s-1).

For a set of size s, its members are s dots on the real line. We say a set is shattered if we can classify all different combinations of positive and negative labelings of these dots with an interval on the line. To show this process, one can start with a small set and work upwards:
In the case of set size 1, one can use an interval to represent that member as a 0 or a 1 trivially.

For set size 2, one can arrange an interval over a, b that includes neither, both, or simply a or b.
For set size 3, one cannot represent alternating elements (for example: 1, 0, 1) due to convexity: For any $a < b < c$;if$a \in I, c \in I$ by convexity of I, every $b \in [a, c]$ is also in I. Because of this, b must also get label 1, so, any such$\{a, b, c\}$is not shattered. Thus, the VC dimension of this class representation is the largest shattered set, 2.

3.3 Linear Half Space

The members of C in this example are half spaces on the real plane and can be represented by 3 parameters: $\{a, b, c\} \in \Re$ s.t. $ax + by > c$.
The case of 1 and 2 points can be easily shattered (half space either includes, excludes, or splits the points).
For set size three (a triangle, three points) one can trivially solve the case where all members are the same. The last case is separating any given point label from the other two. Because the three points are not colinear, we can always draw a line that separates one point from the other 2; thus the case of 3 dots is shattered. (Note: If the three points were collinear, we wouldn't be able to shatter this case - see example 2. However, because we can find at least one position in which this classification works for all possible cases, we still say that the set has been shattered at this size)
To confirm whether or not we can shatter a set size of 4, it is useful to examine multiple cases:

1. 3 or more Colinear Points:

○ ○ ○ ○

○
○ ○ ○

For the same reasons of convexity listed previously, we cannot classify this set, as any half space that includes an interior point necessarily includes at least one other external point, preventing all possible labels.

2. Box (not axis-aligned):

○ ○
○ ○

If one wants a labeling with the points of one diagonal as 1 and the the other points as zero, one cannot shatter this with a linear half space, as any cut that includes one pair of diagonal points necessarily includes one of the other diagonal points.

3. Three external points with one interior point

○
○
○ ○

If we choose a label for the interior point that is different then the external points we can not classify it with the representation due to the convex nature of half spaces: any half space that includes the outer 3 points necessarily includes the interior point.

Because there is no configuration of 4 points that can be shattered, $VCD(\text{Halfspaces in}\Re^2) = 3$

3.4 Boxes in $\Re^2$

members of this class are (axis aligned) boxes in the real plane, defined by 4 parameters $a, b, c, d \in R$, These parameters define the minimum and maximum x and y coordinates of the corners of box. This representation class is more powerful than the halfspace (as opposed to an infinite area defined by a line, we can think of finite boxes).
Due to the above property, It can shatter the sets that the previous example could shatter - so it has VC dimension of at least 3.
For the sets of size 4 we can have the following arrangement of dots:



We see that the diagonal case can now be classified using boxes and other labelings can also be handled because the box does not extend infinitely. So sets of size 4 can be shattered by this class representation.

In case of the sets of size 5, one can take any set of 5 points and select the leftmost, rightmost, topmost, and bottommost points. The fifth point (for now we'll call it 'a') will be guaranteed to appear within the selected four points because $L_x \leq a_x \leq R_x$ and $B_y \leq a_y \leq T_y$ (where L,R,B,T represent left, right, bottom, and top points). Due to this convexity, we cannot classify any set of five, and the $VCD(Boxes\text{in}\Re^2) = 4$.

4 No Free Lunch Theorem

Theorem 1 *No free lunch theorem: Let C be concept class having VC dimension d. Then any algorithm that w.p. $> 1/7$ produces a representation hypothesis h, that has $Pr_{x\in D}[h(x) == C(x)] \geq 1 - \epsilon$, must use $\Omega(d/\epsilon)$ examples from the unknown D.*

Observation 2 *As power of the concept class increases, we need more examples for learning. If power is infinite we should just draw examples from all possible points and store them. This provides no predictive power as it is not possible to draw enough examples to reduce error. This means universal learning is impossible, we must fix some C of bounded VCD in order to collect enough data.*

Proof *We first show a $\Omega(d)$ lower bound for some constant $\epsilon \in (0, 1)$. Then we'll extend it to the $\Omega(d/\epsilon)$ bound. Because the VC dimension of the concept class C is d, a set of d points exists that is shattered by C.*

Observation 3 *Intuition: because all labeling combinations of these d points are representable by C, knowing about each subset of them doesn't provide any information about the state of the other points in the set - they are independent.*

There is a set of functions in C such that for every labeling $b^- \in (0, 1)^d$ some C_{b^-} in the set assigns each x^i label b^i.
Consider the following experiment: Fix any PAC-learning algorithm and let's choose $b^i \in (0, 1)^d$ uniformly at random, so each possible labeling occurs with probability $1/2^d$. This probability distribution is the same as tossing d independent fair coins. Because these coin tosses are independent, knowing the outcome of any d/2 of the coins' outcomes does not change the distribution of the other d/2. So if the algorithm produces a hypothesis h from seeing the $d/2$ examples, any labelings that were not seen in

training can be thought of as completely independent from the observed labelings and so it can disagree with prediction of h with probability $1/2$ *(because h is a hypothesis of the examples observed).*

Observation 4 *If we choose D to be uniform on* $(x^1, ...x^d)$*, then any algorithm that uses* $d/2$ *examples, produces h with expected error greater then* $d/2 * 1/2 * 1/d = 1/4$*. (*$d/2$ *is the number of samples the algorithm sees,* $1/2$ *is the probability of that the label disagrees and* $1/d$ *is the probability of example under D)*

So we define p as the probability that the algorithm produces h that labels random c_{b^-} *with error* $< 1/8$*. Then we'll have:*
Probability that error is less than $1/8*$ *the error (which is less than* $1/8$*)+ Probability that error is greater than* $1/8*$*the error (which is between* $1/8$ *and 1)* $< p * 1/8 + (1-p) * 1 = 1 - 7 * p/8$*.*
The above quantity is the expected error of h and because of the previous observation it's greater than $1/4$*. This leads to:*
$p < 6/7 => (1-p) > 1/7$
$p = Pr[err(h) < 1/8] < 6/7 => Pr[err(h) > 1/8] > 1/7$

Observation 5 *The above (probability of failure of algotirhm* $> 1/7$*) happened for a random choice of labels. So, there exists a deterministic choice which gives probability of failure on* $\epsilon = 1/8$ *equal to* $1/7$ *if it uses* $d/2 = \Omega(d)$ *of examples.*