Lecture #7: VC Dimension, Fundamental Theorem

Instructor: Aditya Bhaskara *Scribe: Yuhan Su*

CS 5966/6966: Theory of Machine Learning

February 1st, 2022

Abstract

This lecture is mainly talking about Vapnik–Chervonenkis(VC) Dimension. The definition is if there exists a set of n points that can be shattered by the classifier and there is no set of n+1 points that can be shattered by the classifier, then the VC dimension of the classifier is n.

In other words, VC dimension is a method of the capacity of several functions(complexity, expressive power, richness, or flexibility) that can be learned by an algorithm which is statistical binary classification. Besides, the definition of it is as the cardinality of the largest set of points that the algorithm can shatter, which means the algorithm can learn a perfect classifier for any labeling of at least one configuration of those data points.

1 Introduction

- VC dimension of a set-family Let H be a set family (a set of sets) and C a set. Their intersection is defined as the following set family:

(1) $H \cap C := h \cup C \mid h \in H.$

- We say that a set C is shattered by H if $H \cap C$ contains all the subsets of C, i.e.:

(2) $\mid H \cap C \mid = 2^{(|C|)}.$

Here are some examples:

1. f is a constant classifier. Its VC dimension is 0 since it cannot shatter even a single point. In general, the VC dimension of a finite classification model, which can return at most $2^{(d)}2^{(d)}$ different classifiers.

2. f is a single-parametric threshold classifier on real numbers; i.e, for a certain threshold θ ,
the classifier $f\theta$ returns 1 if the input number is larger than θ and 0 otherwise. The VC dimension of f is 1 because:
(a) It can shatter a single point.
(b) It cannot shatter any set of two points. For every set of two numbers, if the smaller is labeled 1, then the larger must also be labeled 1, so not all labelings are possible.

2 Sauer's Lemma

Let C be a concept class over an instance space X, i.e. a set of functions functions from X to 0, 1 (where both C and X may be infinite). For any S $\subseteq$ X, let's denote by C (S) the set of all labelings or dichotomies on S that are induced or realized by C, i.e. if S = x1, $\cdots$, xm, then C (S) $\subseteq$ 0, 1 m and

C (S) = (c (x1), $\cdots$, c (xm)) ; c $\in$ C .

Also, for any natural number m, we consider C [m] to be the maximum number of ways to split m points using concepts in C, that is

$C[m] = max|\ C(S)\ |;|\ S\ |= m, S \subseteq X$.

Definition 1
$If|C(S)| = 2|S|$ then S is shattered by C.

Definition 2
The Vapnik-Chervonenkis dimension of C, denoted as V Cdim(C), is the cardinality of the largest set S shattered by C. If arbitrarily large finite sets can be shattered by C, then V Cdim(C) = ∞. Note 1 In order to show that the VC dimension of a class is at least d we must simply find some shattered set of size d. In order to show that the VC dimension is at most d we must show that no set of size d + 1 is shattered.

Examples

1. Let C be the concept class of thresholds on the real number line. Clearly samples of size 1 can be shattered by this class. However, no sample of size 2 can be shattered since it is impossible to choose threshold such that x1 is labeled positive and x2 is labeled negative. Hence the V Cdim(C) = 1.

2. Let C be the concept class intervals on the real line. Here a sample of size 2 is shattered, but no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and outer points are positive. Hence, VCdim(C) = 2.

3. Let C be the concept class of k non-intersecting intervals on the real line. A sample of size 2k shatters (just treat each pair of points as a separate case of example 2) but no sample of size 2k + 1 shatters, since if the sample points are alternated positive/negative, starting with a positive point, the positive points can't be covered by only k intervals. Hence VCdim(C) = 2k.

3 Bounds

1. The VC dimension of the dual set-family of $\mathcal{F}$ is strictly less than $2^{(\mathrm{vc}(\mathcal{F})+1)}$), and this is best possible.

2. The VC dimension of a finite set-family H is at most $\log_2 |H|$.This is because $|H \cap C| \leq |H|$ by definition.

3. Given a set-family H, define H_s as a set-family that contains all intersections of s elements of H. Then:

$VCDim(H_s) \leq VCDim(H)$

4. Given a set-family H and an element $h_0 \in H$, define $H \Delta h_0 := \{h \Delta h_0 \mid h \in H\}$ where Δ denotes symmetric set difference. Then:

$VCDim(H \delta h_0 = VCDim(H)$.

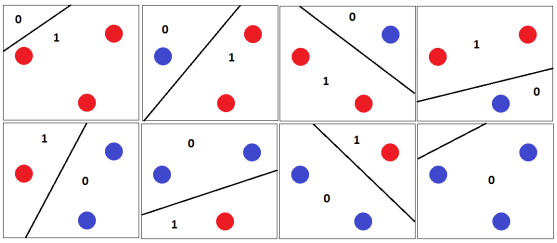

Fig1. 3 points can be classified by H correctly with separating hyper plane.

4 VC DIMENSION OF A NEURAL NETWORK

A neural network is described by a directed acyclic graph G(V,E), where:

1. V is the set of nodes. Each node is a simple computation cell.

2. E is the set of edges, Each edge has a weight.

3. The input to the network is represented by the sources of the graph – the nodes with no incoming edges.

4. The output of the network is represented by the sinks of the graph – the nodes with no outgoing edges.

5. Each intermediate node gets as input a weighted sum of the outputs of the nodes at its incoming edges, where the weights are the weights on the edges.

6. Each intermediate node outputs a certain increasing function of its input, such as the sign function or the sigmoid function. This function is called the

activation function.

The VC dimension of a neural network is bounded as follows:

1. If the activation function is the sign function and the weights are general, then the VC dimension is at most $O(|E| \cdot \log(|E|))O(|E| \cdot \log(|E|))$.

2. If the activation function is the sigmoid function and the weights are general, then the VC dimension is at least $\Omega(|E|^2)\Omega(|E|^2)$ and at most $O(|E|^2 \cdot |V|^2)O(|E|^2 \cdot |V|^2)$.

3. If the weights come from a finite family (e.g. the weights are real numbers that can be represented by at most 32 bits in a computer), then, for both activation functions, the VC dimension is at most $O(|E|)O(|E|)$.

5 Generalizations

The VC dimension is defined for spaces of binary functions (functions to 0,1). Several generalizations have been suggested for spaces of non-binary functions.

1. For multi-valued functions (functions to 0,...,n), the Natarajan dimension can be used.

2. For real-valued functions (e.g. functions to a real interval, [0,1]), Pollard's pseudo-dimension can be used.

3. The Rademacher complexity provides similar bounds to the VC, and can sometimes provide more insight than VC dimension calculations into such statistical methods such as those using kernels.