Learning Threshold Functions: Beyond the VC Dimension

VC Dimension

VC Dimension

- Characterizes the expressiveness of a hypothesis class for binary classification tasks
- Applied to the binary classification setting and cannot be applied to the multiclass setting

VC Dimension In simple terms:

- The maximum number of points that can be shattered in a hypothesis space

Threshold Function:

- A threshold function $h_\tau : R \rightarrow \{0, 1\}, \tau \in R$, is a function where $h_\tau(x) = 1$ if and only if $x \geq \tau$

Fundamental Theorem of PAC Learning:

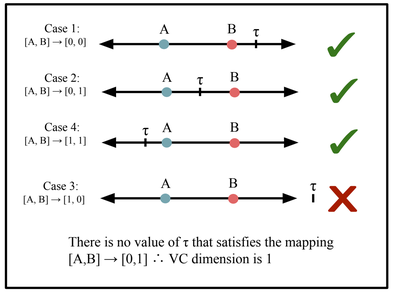
The Fundamental Theorem of PAC learning states that any concept that can be learned from a training set with a small error rate can also be generalized to new, unseen examples with a high level of accuracy for sufficiently large training data. Importantly, the VC dimension characterizes the number of training examples needed to achieve a given level of accuracy in PAC learning. **The VC dimension can be used to make statements about the number of training examples needed to achieve a certain level of accuracy in PAC learning.**

upper bound on the sample complexity $m_H(\epsilon, \delta)$

$$m_{\mathcal{H}}(\epsilon, \delta) \leq c_2 \cdot \frac{\mathrm{VCdim}(\mathcal{H}) \cdot \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

where c_2 is some constant, VCdim(H) is the VC dimension of H, ϵ and δ are the PAC parameters

Visualizing VC Dimension



There is no value of τ that satisfies the mapping [A,B] → [0,1] ∴ VC dimension is 1

Let set S = {A, B} be a set that is shattered by the hypothesis class H if there exists all possible mappings of A and B in {0,1} given a threshold function h:

$R \rightarrow \{0, 1\}, \tau \in R$, is a function where $h_\tau(x) = 1$ if and only if $x \geq \tau$

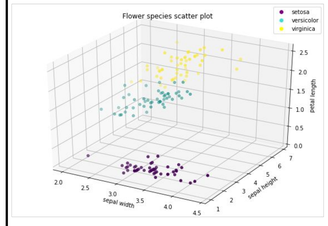
Algorithm

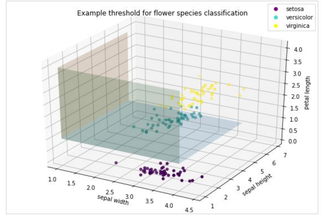
We implemented an algorithm that takes a 1-dimensional dataset and returns a simple classification function that is constrained to a single threshold parameter T. Data points above the threshold are classified as 1 and points below are 0.

The algorithm works by sorting the data, then while iterating through the data, insert a threshold in between the two current data points and counting the number of misclassifications. We then output the threshold function that minimized empirical risk through minimizing misclassifications.

To produce a function that could help us classify Iris species based on the Bezdek Iris dataset, we generalized this algorithm to the multidimensional and multiclass space. We did so by applying a different multidimensional threshold to each class. To identify the proper threshold, we analogously sorted the data by the minimum of its input vector entries. We iteratively selected the threshold that produced the least number of misclassifications, then combine the classification function for each class into a single classification function.

A vector is classified positively for a class if all of the vector's entries are above the threshold T. This is what an example single class threshold function looks like on the Bezdek Iris dataset in a 3-dimensional space with T= 1:





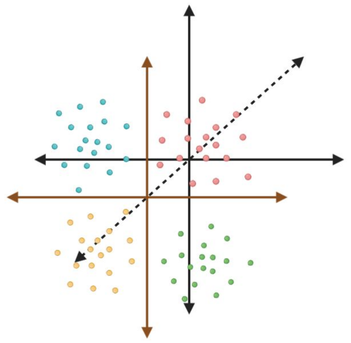
Results:
Training performance: 31.67% all correct, 65.0% somewhat correct
Validation performance: 33.33% all correct, 60.0% somewhat correct

We trained the model on 80% of the dataset and tested its performance on the rest of the data it had yet to see. We saw great results on validation data that were about equal to the model's performance on training data. This is a good sign that the model is not overfitting. The metrics we used to evaluate our model were "all correct %" and "somewhat correct %". All correct classifications were ones where the correct class was classified as positive and all other classes were classified as negative. Somewhat correct classifications were ones where the correct class was classified as positive. Overall the accuracy of the algorithm is quite low, but we think that the results are actually very good given the constraints of a single parameter model in a multi-dimensional space.

Natarajan Dimension

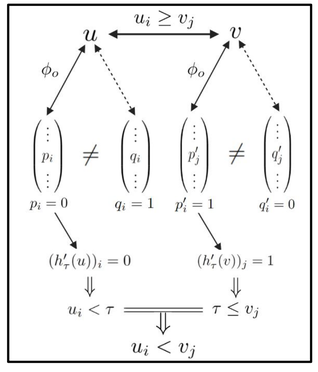
Formally, a set S of vectors in $\mathbb{R}^n$ is shattered by the hypothesis class $\mathcal{H}'$ if the following conditions are satisfied:

1. $\forall \vec{x}_k \in S$, $\exists \vec{x}_{k,A}, \vec{x}_{k,B} \in \{0,1\}^n$, such that $\vec{x}_{k,A} \neq \vec{x}_{k,B}$.
2. $\forall \phi \in \Phi$ where $\Phi = \{\phi \in \mathcal{F} : \phi : \{\vec{x}_k \in S\} \rightarrow \{\vec{x}_{k,A}, \vec{x}_{k,B}\}\}$ $\exists h'_\tau \in \mathcal{H}'$ such that $\forall \vec{x}_k \in S$ $h'_\tau(\vec{x}_k) = \phi(\vec{x}_k)$.



The Natarajan dimension is then given by the cardinality of the largest set $\mathcal{H}'$ can shatter

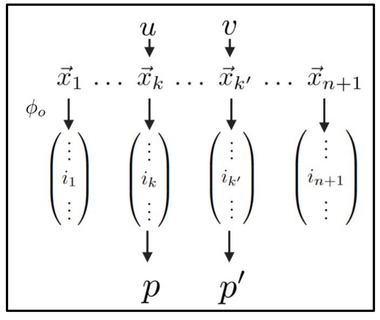
A higher dimensional single parameter threshold function can be visualized in two dimensions as a sliding origin moving along the line $y = x$ and deciding the regions of the feature space according to which quadrant each vector lies within.



We define, $\mathcal{H}' = \{h'_\tau : \tau \in \mathbb{R}\}$, where h'_τ applies the threshold rule element-wise to the input, i.e. for $\vec{x} \in \mathbb{R}$ and $\vec{y} = h'_\tau(\vec{x})$; then $\vec{y}_i = 1 \Leftrightarrow \vec{x}_i \geq \tau, i \in \{1, ...n\}$.

Since $\mathcal{H}$ and $\mathcal{H}'$ both rely on a single parameter and therefore have only one degree of freedom, we attempt a guess that the $VCdim(\mathcal{H}) = Ndim(\mathcal{H}')$. To prove this we need to show that $\mathcal{H}'$ cannot shatter a set S of two vectors. We apply the definition: Take two vectors u and v and their associated vectors p, q and p', q' respectively. We know p, q and p', q' must differ internally in at least one index each i and j. This is a useful metric because we now know that there must be at least one ϕ_o that maps $u_i \rightarrow 1$ and $v_j \rightarrow 0$. Let's say $u_i \geq v_j$ implying $h'_\tau(u)_i$ must greater than or equal to $h'_\tau(v)_j$. Notice, however, this renders h'_τ incapable of reproducing ϕ_o. This appears to have proven only one specific case, but by relabelling we can show that this is actually completely general; for any two vectors you can always find ϕ_o for which no h'_τ can reproduce.

Consider a new hypothesis class, $\hat{\mathcal{H}}$, consisting of more general threshold functions, $\hat{h}_\Gamma : \mathbb{R}^n \rightarrow \{0,1\}^n$, $\Gamma \in \mathbb{R}^n$, where $\hat{h}_\Gamma(\vec{x})_i = 1 \Leftrightarrow \vec{x}_i \geq \Gamma_i$. Now to show the $Ndim(\hat{\mathcal{H}}) \leq n$: Suppose $\hat{\mathcal{H}}$ can shatter a set $S = \{\vec{x}_1, ..., \vec{x}_{n+1}\}$ of $n+1$ vectors in $\mathbb{R}^n$. As before, assume each $\vec{x_k}$ has associated vectors $p_k, q_k \in \{0,1\}^n$ that differ internally by at least one element $i_k \in \{1, .., n\}$. Since there are only n unique choices for i_k and since each of the $n+1$ vectors is mapped to one of those i_k, there must be at least two vectors in S mapped to the same index, i.e. there exists some $k, k' \in \{1, .., n+1\}$ such that $i_k = i_{k'}$. Notice, the components of $\vec{x}_k, \vec{x}_{k'}$ corresponding to the indices $i_k, i_{k'}$ will be compared to the same threshold, $\Gamma_{i_k = i_{k'}}$, when classified by any $\hat{h}_\Gamma \in S$. The rest follows similar to before. Let $i = i_k = i_{k'}$. WLOG assume $(\vec{x_k})_i \leq (\vec{x_{k'}})_i$ and $(p_k)_i = 1, (q_k)_i = 0$. Again, any $\hat{h}_\Gamma$ cannot compute $(\vec{x_k})_i \rightarrow 1$ and $(\vec{x_{k'}})_i \rightarrow 0$ if $(\vec{x_k})_i \leq (\vec{x_{k'}})_i$ since from the definition of $\hat{h}_\Gamma$ this would imply $(\vec{x_k})_i \geq \Gamma_i > (\vec{x'_k})_i$, a contradiction.



Peter Sailer, Yusef Essawi, Nate Kimball