VC DIMENSION

MATHEMATICAL TRIPOS PART II

1. Setup

We work in a classification setting. We have a set $\mathcal{X}$ and want to classify them into labels $\{-1, 1\}$. We do this through a **hypothesis** function $h : \mathcal{X} \to \{-1, 1\}$.

For a given clasification $h(x)$, we define the **loss** as $\ell(h(x), y) = \mathbb{1}[h(x) \neq y]$, where y the the correct label for x. We define the **risk** of a hypothesis to be it's expected loss, $R(h) = \mathbb{E}[\ell(H(X), Y)]$.

2. VC Dimension

The VC dimension of a set of hypothesis functions is a measure of its ability to overfit the training data. A high VC dimension indicates that the set of functions is likely to overfit the training data, while a low VC dimension indicates that the set of functions is less likely to overfit the training data.

Consider $\{x_1, \dots, x_n\} \in \mathcal{X}$. Given a collection of possible hypothesis $\mathcal{H}$, we can count how many different ways we can partition the x_i (into the two labels ± 1). We define[1]

$$\mathcal{H}(x_1, \dots, x_n) = \{S \subseteq \{x_1, \dots, x_n\} \mid \exists h \in \mathcal{H} \text{ where } h(x_i) = 1 \; \forall x_i \in S\}.$$

We usually just care about the size of this set, and we note that we trivially have $|\mathcal{H}(x_1, \dots, x_n)| \leq 2^n$.

Definition (Shattered)

We say that $\{x_1, \dots, x_n\}$ is **shattered** by $\mathcal{H}$ if $|\mathcal{H}(x_1, \dots, x_n)| = 2^n$.

That is, if for any possible labelling of the points in $\{x_1, \dots, x_n\}$, there exists a hypothesis $h \in \mathcal{H}$ that correctly classifies all of the points in S.

Definition (VC Dimension)

The **VC dimension** $\mathrm{VC}(\mathcal{H})$ is the largest integer n such that some $\{x_1, \dots, x_n\}$ is shattered by $\mathcal{H}$, or ∞ if none exists.

Definition (Shattering Coefficient)

We define the **shattering coefficient** $s(\mathcal{H}, n)$ to be

$$s(\mathcal{H}, n) = \max_{S \subset \mathcal{X}, |S| = n} |\mathcal{H}(S)|.$$

Date: May 27, 2023.

[1] Informally, $\mathcal{H}(x_1, \dots, x_n)$ is the set of 'obtainable paritions' of $\{x_1, \dots, x_n\}$ using a hypothesis in $\mathcal{H}$. Also we just suppose $h(x_i) = 1$, but we could take $h(x_i) = -1$.

So equivalently, $\mathrm{VC}(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid s(\mathcal{H}, n) = 2^n\}$.

To show that $\mathrm{VC}(\mathcal{H}) = n$, we must first find an $\{x_1, \ldots, x_n\}$ that is shattered (this is usually easy), and show that no $\{x_1, \ldots, x_{n+1}\}$ can be shattered (this is usually harder).

Example 2.1 (Finding VC Dimension)

Consider the example of $\mathcal{X} = \mathbb{R}$, and the set of hypothesis functions classify points based on whether or not they are in a given interval:

$$\mathcal{H} = \{h_{a,b} \mid h_{a,b}(x) = \mathbb{1}_{[a,b)}(x),\ a, b \in \mathbb{R}\}.$$

Consider n distinct points $x_1 < x_2 < \cdots < x_n$. These divide up the real line into $n+1$ intervals $(-\infty, x_1], (x_1, x_2], \ldots, (x_{n-1}, x_n], (x_n, \infty)$.

If a and a' are in the same interval, and b and b' are in the same interval, then $h_{a,b}(x_i) = h_{a',b'}(x_i)$ for all i. Thus every possible behaviour of $h_{a,b}$ on the x_i is obtained by picking one of the $n+1$ intervals for each of a and b. Thus

$$s(\mathcal{H}, n) \leq (n+1)^2.$$

Now we compute the VC dimension. Clearly any $\{x_1, x_2\}$ can be shattered, but with three points $\{x_1, x_2, x_3\}$ with $x_1 < x_2 < x_3$, we can never have $h(x_1) = h(x_3) = 1$ and $h(x_2) = 0$, and so $\mathrm{VC}(\mathcal{H}) = 2$.

Example 2.2 (VC Dimension for Vector Spaces)

Consider a vector space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, and from this form the class of hypotheses

$$\mathcal{H} = \{h_f \mid f \in \mathcal{F},\ h_f(x) = \mathrm{sgn}(f(x))\},$$

where we take $\mathrm{sgn}(0) = -1$.

We will prove that $\mathrm{VC}(\mathcal{H}) \leq \dim(\mathcal{F})$.

Let $d = \dim \mathcal{F} + 1$, and let $\{x_1, \ldots, x_d\} \subset \mathcal{X}$. We need to show that this cannot be shattered by $\mathcal{H}$. Consider the linear map $L : \mathcal{F} \to \mathbb{R}^d$ given by

$$L(f) = (f(x_1), \ldots, f(x_d)) \in \mathbb{R}^d.$$

The rank of L is at most $\dim \mathcal{F} = d - 1 < d$, therefore there must exist non-zero $\gamma \in \mathbb{R}^d$ orthogonal to everything in the image of $L(\mathcal{F})$, that is,

$$\sum_{i, \gamma_i > 0} \gamma_i f(x_i) + \sum_{i, \gamma_i \leq 0} \gamma_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{F},$$

where (WLOG) at least one component of γ is strictly positive. Let $I_+ = \{i \mid \gamma_i > 0\}$ and $I_- = \{i \mid \gamma_i \leq 0\}$. Then it is not possible to have

$$h(x_i) = 1 \implies f(x_i) > 0 \text{ for all } i \in I_+,$$
$$h(x_i) = -1 \implies f(x_i) \leq 0 \text{ for all } i \in I_-,$$

as otherwise the LHS of our orthogonality equation would be strictly positive. Thus $\{x_1, \ldots, x_d\}$ cannot be shattered, and $\mathrm{VC}(\mathcal{H}) \leq d - 1$, as required.

3. Sauer-Shelah Lemma

Note in our first example we had $s(\mathcal{H}, n) \leq (n+1)^{\mathrm{VC}(\mathcal{H})}$. This result holds more genererally.

Lemma (Sauer-Shelah)

Let $\mathcal{H}$ have finite VC dimension d. Then

$$s(\mathcal{H}, n) \leq (n+1)^d.$$

Proof. Non-examinable. $\square$