In this lecture we obtain a bound for the Rademacher complexity of a hypothesis class in terms of its VC dimension. Using this we obtain bounds for generalisation error in terms of VC dimension.

1 Upper Bounds on Error

Let $H$ be a class of functions from a domain set $X$ to label set $\{-1, +1\}$. Let $G$ be the associated class of 0-1 loss functions. We start by observing a simple relationship between the respective empirical Rademacher complexities of $H$ and $G$.

**Proposition 1.** Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq X \times \{-1, +1\}$, let $S'$ denote the corresponding set $\{x_1, \ldots, x_m\}$ of unlabelled domain elements. Then $R_S(G) = \frac{1}{2} R_{S'}(H)$.

**Proof.** Notice that for $h \in H$, the corresponding 0-1 loss function $g : X \times \{-1, +1\} \rightarrow \{0, 1\}$ can be written $g(x, y) = \frac{1}{2}(1 - yh(x))$. Now we have

$$R_S(G) = \mathbb{E}_\sigma \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i, y_i) \right]$$

$$= \mathbb{E}_\sigma \left[ \sup_{h \in H} \frac{1}{2m} \sum_{i=1}^{m} \sigma_i (1 - y_ih(x_i)) \right]$$

$$= \mathbb{E}_\sigma \left[ \frac{1}{2m} \sum_{i=1}^{m} \sigma_i \right] + \mathbb{E}_\sigma \left[ \sup_{h \in H} \frac{1}{2m} \sum_{i=1}^{m} -\sigma_i y_ih(x_i) \right]$$

$$= \mathbb{E}_\sigma \left[ \sup_{h \in H} \frac{1}{2m} \sum_{i=1}^{m} \sigma_i y_ih(x_i) \right]$$

$$= \frac{1}{2} R_{S'}(H). \quad (1)$$

Line (1) follows because $\mathbb{E}[\sigma_i] = 0$ and because $-\sigma_i$ and $\sigma_i$ are identically distributed. \qed

The following lemma gives an upper bound for the Rademacher complexity of finite sets of vectors, each of length at most one.

**Lemma 1 (Massart’s Lemma).** Let $A \subseteq \mathbb{R}^m$ be a finite set of vectors with $\|a\| \leq 1$ for all $a \in A$. Then

$$\mathbb{E}_\sigma \left[ \max_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \right] \leq \sqrt{2 \log |A|},$$

where the $\sigma_i$ are independent random variables uniform over $\{-1, +1\}$ and $a_1, \ldots, a_m$ are the components of vector $a$.

**Proof.** We will show that $\mathbb{E}_\sigma \left[ \max_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \right] \leq \frac{\log |A|}{t} + \frac{1}{2}$ for all $t > 0$. From this the lemma follows by choosing $t = \sqrt{2 \log |A|}$. We establish the inequality by the following derivation, valid for
all $t > 0$:

$$
t E\left[ \max_{a \in A} \sum_{i=1}^{m} \sigma_{i} a_{i} \right] = E\left[ \max_{a \in A} \sum_{i=1}^{m} t \sigma_{i} a_{i} \right]
\leq E\left[ \log \sum_{a \in A} \exp \left( \sum_{i=1}^{m} t \sigma_{i} a_{i} \right) \right] \quad (2)
\leq \log \left( E\left[ \sum_{a \in A} \exp \left( \sum_{i=1}^{m} t \sigma_{i} a_{i} \right) \right] \right) \quad (3)
= \log \left( \sum_{a \in A} \left( \exp \left( \frac{1}{2} (ta_{i})^{2} \right) \right) \right) \quad (4)
\leq \log \left( \sum_{a \in A} \exp \left( \frac{1}{2} \|ta\|^{2} \right) \right) \quad (5)
= \log \left( \sum_{a \in A} \exp \left( \frac{1}{2} \|ta\|^{2} \right) \right)
\leq \log \left( |A| e^{\frac{1}{2} \|ta\|^{2}} \right) \leq \log |A| + \frac{t^{2}}{2} \quad (6)
$$

Line (2) use the inequality $\max_{j} x_{j} \leq \log(\sum_{j} \exp(x_{j}))$. Line (3) uses Jensen’s Inequality (see the Probability Handout), Line (4) uses the independence of the $\sigma_{i}$, while Line (5) uses the inequality $e^{x} + e^{-x} \leq e^{x^{2}}$, valid for all $x \in \mathbb{R}$.

**Corollary 1.** Let $H$ be a family of functions on a domain $X$, taking values in $\{-1, +1\}$, with VC dimension $d$. Let $S = \{x_{1}, \ldots, x_{m}\} \subseteq X$, then

$$
R_{S}(H) = O\left( \sqrt{\frac{\log(m/d)}{m/d}} \right)
$$

**Proof.** Write $A = \{ \frac{1}{\sqrt{m}} (h(x_{1}), \ldots, h(x_{m})) : h \in H \}$. Then $A$ is a set of vectors in $\mathbb{R}^{m}$, each of length at most one, and $|A| \leq \left( \frac{cm}{d} \right)^{d}$ by Sauer’s Lemma. Thus

$$
R_{S}(H) = E_{\sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_{i} h(x_{i}) \right]
\leq \frac{1}{\sqrt{m}} \sqrt{2 \log \left( \frac{cm}{d} \right)^{d}} \quad \text{by Massart’s Lemma}
\leq O\left( \sqrt{\frac{\log(m/d)}{m/d}} \right)
$$

**Theorem 1.** Let $H$ be a family of functions on a domain $X$, taking values in $\{-1, +1\}$, with VC dimension $d$. Let $S$ be a sample of size $m$ drawn independently from a fixed distribution $D$ on $X$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$ we have

$$
\text{err}(h) \leq \hat{\text{err}}(h) + O\left( \sqrt{\frac{\log(m/d)}{m/d}} \right) + O\left( \sqrt{\frac{\log 1/\delta}{m}} \right)
$$
Proof. Let $G$ be the family of 0-1 loss functions associated with $H$. From Corollary 1 and Proposition 1 we have that $R_m(G) = O\left(\sqrt{\frac{\log(m/d)}{m/d}}\right)$. The statement of the theorem now follows from Theorem 2 in Lecture 5, noting that if $g$ is the 0-1 loss function corresponding to $h \in H$ then $err(h) = L(g)$ and $\hat{err}(h) = L_S(g)$. 

2 Discussion

Notice the similarity between Theorem 1 and the corresponding bound for finite hypothesis classes, stated at the end of Lecture 3. The main difference is that the $\log |H|$ term there has been replaced by the VC dimension $d$.

We can read Theorem 1 as saying that to achieve a fixed error $\varepsilon$ with fixed confidence $\delta$ the ratio $m/d$ should also be fixed, i.e., the required number of examples scales linearly with the VC dimension of $H$. For example, consider the hypothesis class $\mathcal{H} = \{H_n : n \in \mathbb{N}\}$, with $H_n$ the class of linear threshold functions on $\mathbb{R}^n$. Here the VC dimension of $H_n$, and hence the required number of examples to learn a hypothesis in $H_n$, is linear in $n$. Thus this class is PAC learnable. More generally, a hypothesis class $\mathcal{H} = \{H_n : n \in \mathbb{N}\}$ is PAC learnable if the VC dimension of $H_n$ grows polynomially in $n$ (provided, of course, that one can compute a hypothesis that minimises empirical error on a sample). For an example of non-polynomial growth in the VC dimension, take $H_n$ to be the collection of Boolean functions on $\{0, 1\}^n$ defined by DNF formulas. Here the VC dimension of $H_n$ is $2^n$ and, as we will see, this class is not PAC learnable.