Rademacher Complexity

$\mathcal{H}$: hypothesis set $h: \mathcal{X} \to \mathcal{Y}$ $h \in \mathcal{H}$

We can have various notions of a loss function $L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Given $\mathcal{H}$ we define the losses

$$\mathcal{G} := \{ g: \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \mid g(x,y) = L(h(x), y) \}$$

We are trying to associate a complexity measure with $\mathcal{G}$ that measures how rich $\mathcal{G}$ is.

Idea: If $\mathcal{G}$ is very expressive it can even fit random noise.

Def The Rademacher complexity of $\mathcal{G}$ wrt samples of size m is

$$R_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [$$

McDiarmid's Inequality

Let V be a set and $f: V^m \to \mathbb{R}$ be a f^n s.t.

$\forall i \ \exists c_i > 0 \ \forall x_1, \dots x_m, x_i' \in V$

$$|f(x_1, x_2, \dots, x_i, \dots, x_m) - f(x_1, \dots, x_i', \dots, x_m)| \leq c_i$$

Let $X_1, \dots X_m$ be random variables taking values in V.

Then $\forall \epsilon > 0$ $\quad \mathbb{P}(f(X_1, \dots X_m) \geq \mathbb{E}[f(X_1, \dots X_m)] + \epsilon) \leq e^{-2\epsilon^2/mc^2}$

where $c^2 = \sum c_i^2$.

If $f(x_1, \dots x_m) = \frac{1}{m} \sum x_i$ & $c = 1/m$ $\quad x_i \in [0,1]$

we get the Hoeffding inequality.

We consider G to be a family of fns $Z \to \mathbb{R}$.
Let $\mathcal{D}$ be a distribution over Z & $g \in G$.
We write $L(g) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [g(z)]$

Given a sample S of size m drawn according to $\mathcal{D}$ from Z we write $L_S(g) = \frac{1}{m} \sum_{i=1}^{m} g(z_i)$.

The basic goal: we want to show that L_S & L are not too far apart with high probability for <u>all</u> $g \in G$. Rademacher complexity will give us a quantitative handle on this.

Let $\vec{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_m)$ be a family of ind. random variables s.t. $\forall i \in \{1, \cdots, m\}$ $\sigma_i = +1$ with prob. $1/2$ and -1 with prob. $1/2$.

<u>Def</u> The empirical Rademacher complexity of G with respect to the sample S of size m is

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

If we write $\vec{g}_S = (g(z_1), \cdots, g(z_m))$ then we can write $R_S(G) = \mathop{\mathbb{E}}_{\vec{\sigma}} \left[\sup_{g \in G} \frac{\vec{\sigma} \cdot \vec{g}_S}{m} \right]$

If this is high it means that G can correlate well with the random variables σ_i which means G is very expressive.

Def The Rademacher complexity of G wrt S of size m is

$$R_m(G) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[R_S(G)]$$

Note R_m depends on the distribution.

Thm 1 Let G be a family of functions $Z \to [0,1]$. Let S be a sample of size m drawn iid according to $\mathcal{D}$ from Z. Then $\forall \delta > 0$ with probability at least $1-\delta$, the following holds

$$\forall g \in G \quad L(g) \le L_S(g) + 2R_m(G) + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right)$$

Proof Want a bound on $L(g) - L_S(g)$.

Define a new random variable $\Phi(S)$ by

$$\Phi(S) = \sup_{g\in G}\left(L(g) - L_S(g)\right)$$

Idea: give a bound on $\mathbb{E}[\Phi]$ & use a concentration inequality to argue that Φ is close to $\mathbb{E}[\Phi]$ with high probability.

Write $S = (z_1, \cdots, z_m)$ and consider Φ as a function of $(z_1, \cdots z_m)$.

Short digression: If X is any compact set & $f, g : X \to [0,1]$ are continuous then

$$\left|\sup_{x\in X} f(x) - \sup_{x\in X} g(x)\right| \le \sup_{x\in X} |f(x) - g(x)|.$$

Suppose $\sup f$ is attained at x_0 & $\sup f > \sup g$

$$\sup f(x) - \sup g(x) = f(y) - \sup g(x) \le f(y) - g(y) \le \sup(f-g)$$

if $\sup g > \sup f$ flip the roles of f and g.

Now what happens if we change z_i to z_i'?

$$|\Phi(z_1,\dots z_m) - \Phi(z_1,\dots,z_i',\dots z_m)|$$
$$= \left| \sup_{g\in G}[L(g)-L_S(g)] - \sup_{g\in G}[L(g)-L_{S'}(g)] \right|$$
$$\le \sup_{g\in G} |L(g)-L_S(g)-L(g)+L_{S'}(g)|$$
$$= \sup_{g\in G} \frac{1}{m}|g(z_i)-g(z_i')| \le \frac{1}{m}$$

Thus we can use McDiarmid's inequality with $c = 1/m$. Thus $\forall \varepsilon > 0$

$$P(\Phi(S) \ge \mathbb{E}[\Phi(S)] + \varepsilon) \le e^{-2m\varepsilon^2}$$

Take $\varepsilon = \sqrt{\frac{\log 1/\delta}{2m}}$ $e^{-2m\varepsilon^2} = \delta$

So with probability at least $1-\delta$

$$\Phi(S) \le \mathbb{E}[\Phi(S)] + \varepsilon = \sqrt{\frac{\log 1/\delta}{2m}}$$

Now we need an upper bound on $\mathbb{E}[\Phi(S)]$.

Suppose we have a second sample

$$S' = \{z_1', \dots z_m'\} \text{ also drawn to } \mathcal{D}$$

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\Phi(S)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{g\in G}(L(g)-L_S(g))\right] \quad \text{def of } \Phi$$
$$= \mathbb{E}_S\left[\sup_{g\in G} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}(L_{S'}(g) - L_S(g))\right] \quad L(g) = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}[L_{S'}(g)]$$

For any RV's $\sup_i \mathbb{E}[X_i] \le \mathbb{E}[\sup_i X_i]$

$$\le \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in G}(L_{S'}(g)-L_S(g))\right]$$

$$= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in G}\left\{\frac{1}{m}\sum_{i=1}^m g(z_i')-g(z_i)\right\}\right] \quad \text{def of } L_S$$

Now consider the following sum

$$\sum_{i=1}^{m} \sigma_i \left[g(z_i') - g(z_i) \right]$$

where $\sigma_i = \pm 1$ with equal probabilities.
This is equivalent to taking 2 samples & swapping the values at position i or not according to a fair coin toss

$$\begin{array}{cccccccc} z_1 & z_2 & z_3 & \cdots & \boxed{z_i} & \cdots & z_m \\ z_1' & z_2' & z_3' & \cdots & \boxed{z_i'} & \cdots & z_m' \end{array}$$

↓

If we get H swap else keep.

So the sums will have terms flipped or not with prob $1/2$. This is what σ_i's do.
This swapping process does not change the distribution of S & S' so.

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \left[g(z_i') - g(z_i) \right] \right]$$

$$= \mathop{\mathbb{E}}_{S,S',\vec{\sigma}} \left[\sup_{g \in G} \frac{1}{m} \left[\sum_{i=1}^{m} \sigma_i \left[g(z_i') - g(z_i) \right] \right] \right]$$

$$\leq \mathop{\mathbb{E}}_{S,\vec{\sigma}} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i') \right] + \mathop{\mathbb{E}}_{S',\vec{\sigma}} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \right]$$

$$= 2 R_m(G).$$

Thus $\mathbb{E}[\Phi(S)] \leq 2 R_m(G)$.

$$\Phi(S) \leq 2 R_m(G) + \left[\frac{\log 1/\delta}{2m} \right]^{1/2} \text{ with prob} \geq 1 - \delta$$

Cor Suppose we draw a sample of size m according to $\mathcal{D}$. Then for any $\delta > 0$ with prob. at least $1-\delta$ the following holds

$$\forall g \in G \qquad L(g) \le L_S(g) + 2R_S(G) + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right)$$

Proof Consider $R_S(G) := \mathbb{E}_\sigma\left[\sup_g \frac{1}{m}\sum_{i=1}^m g(z_i)\right]$ as a function of $(z_1, \cdots z_m) \in S$. Changing one of the z_i to z_i' will change R_S by $1/m$ at most. From McDiarmid's inequality with $c = 1/m$ & $\varepsilon = \sqrt{\frac{\log 1/\delta}{m}}$ we have with prob at least $1-\delta$

$$R_S(G) \le R_m(G) + \sqrt{\frac{\log 1/\delta}{m}}$$

Now by the union bound the probability that this is violated *or* the inequality in the theorem is violated is 2δ. Thus they *both* hold with probability at least $1-2\delta$. Replacing δ by $\delta/2$ & absorbing constants in the $O(\cdot)$ notation we get the result.

— x —

Now we will relate Rademacher complexity & VC dimension.

Let H be a class of functions $X \to \{+1, -1\}$. Let G be the associated class of 0-1 loss fns.

Prop 1 Given a sample $S = \{(x_1, y_1), \cdots (x_m, y_m)\} \in X \times \{+1, -1\}$ Let S' denote the set $\{x_1, \cdots x_m\}$ with labels removed from S. Then

$$R_S(G) = \frac{1}{2} R_{S'}(H)$$

Proof Fix $h \in H$ and let g be the corresponding loss function: $X \times \{+1,-1\} \to \{0,1\}$.
We can write $g(x,y) = \frac{1}{2}(1 - y\,h(x))$

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

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

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

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

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

$$= \tfrac{1}{2} R_S(H).$$

LEMMA Let $A \subseteq \mathbb{R}^m$ be a finite set of vectors with $\|\vec{a}\| \le 1$ for all $\vec{a} \in A$. Then

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

where σ_i are iid RV in $\{+1,-1\}$ uniform.

PROOF Homework → MASSART'S LEMMA

Cor 1 Let H be a family of functions on X taking values in $\{+1,-1\}$ with VC dim $= d$. Let $S = \{x_1, \dots x_m\} \subseteq X$. Then

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

Proof Write $A = \left\{ \frac{1}{\sqrt{m}} (h(x_1), \dots, h(x_m)) \mid h \in H \right\}$

Then A is a family of vectors in $\mathbb{R}^m$ each of length at most 1 & $|A| \leq \left(\frac{em}{d}\right)^d$ [SAUER'S LEMMA]

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

$$= \frac{1}{\sqrt{m}} \mathbb{E}_{\vec{\sigma}} \left[\sup_{\vec{a} \in A} \sum \vec{\sigma} \cdot \vec{a} \right]$$

$$\leq \frac{1}{\sqrt{m}} \sqrt{2 \log \left(\frac{em}{d}\right)^d} \quad \text{[MASSART'S LEMMA]}$$

$$= O\left(\sqrt{\frac{\log(m/d)}{(m/d)}} \right). \quad \blacksquare$$

Putting all these together we get

Thm 2 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 i.i.d from a fixed distribution $\mathcal{D}$. Let $err(h) = \mathbb{P}_{x \sim \mathcal{D}} [h(x) \neq y]$ & $\hat{err}(h)$ be the sample error. Then

$$err(h) \leq \hat{err}(h) + O\left(\sqrt{\frac{\log m/d}{m/d}} \right) + O\left(\sqrt{\frac{\log 1/\delta}{m}} \right)$$

If we use g as the 0-1 loss f^n of h then $err(h) = L(g)$ & $\hat{err}(h) = L_S(g)$ & the result follows from Cor 1, Prop 1 & Thm 1.