Notes 22: Rademacher complexity

1. Rademachar complexity

Given training samples $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ with $y_i \in \{+1, -1\}$ and hypothesis class $\mathcal{H}$

Empirical Risk Minimization algorithm

Output $h \in \mathcal{H}$ that minimizes empirical error on S

Generalizes Consistent Hypothesis Algorithm from Notes09
 samples need not be labeled by any $h \in \mathcal{H}$ (e.g. labels y_i may be corrupted, as in RCN)
Can we bound generalization error of this algorithm, similar to the Theorem in Notes13?

Training/empirical error of hypothesis $h : X \to \{+1, -1\}$ on S is

$$\frac{1}{m} \sum_{1 \leqslant i \leqslant m} \mathbb{1}(h(x_i) \neq y_i) = \mathop{\mathbb{E}}_{i \in [m]}[\mathbb{1}(h(x_i) \neq y_i)] = \mathop{\mathbb{E}}_{i \in [m]}\left[\frac{1 - y_i h(x_i)}{2}\right] = \frac{1}{2} - \frac{1}{2} \mathop{\mathbb{E}}_{i \in [m]}[y_i h(x_i)]$$

$\mathbb{E}_{i \in [m]}[y_i h(x_i)]$ can be interpreted as **correlation** between predictions $h(x_i)$ with labels y_i
Correlation is always between -1 and 1 (as the average of m numbers, each being -1 or 1)
Finding hypothesis to minimize training error $\iff$ Finding hypothesis to maximize this correlation
 i.e. $\arg\max_{h \in \mathcal{H}} \mathbb{E}_{i \in [m]}[y_i h(x_i)]$

Now imagine true labels y_i are replaced with **Rademacher random variables** σ_i
 i.e. $\sigma_i = +1$ with probability $1/2$ and $\sigma_i = -1$ with probability $1/2$, independently across i
Fix hypothesis class $\mathcal{H}$ (with $\{+1, -1\}$-valued hypotheses)
Definition **Empirical Rademacher complexity** of $\mathcal{H}$ wrt S is

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\sigma \in \{+1,-1\}^n}\left[\sup_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{i \in [m]}[\sigma_i h(x_i)]\right]$$

sup instead of max to allow infinite $\mathcal{H}$
e.g. $|\mathcal{H}| = 1 \implies \hat{\mathcal{R}}_S(\mathcal{H}) = 0$ (regardless of $h(x_i)$, $\mathbb{E}[\sigma_i] = 0$)
e.g. $\mathcal{H}$ shatters $\{x_1, \dots, x_m\} \iff |\mathcal{H}| = 2^m \implies \hat{\mathcal{R}}_S(\mathcal{H}) = 1$ (can force $\sigma_i h(x_i) = 1$)
In general $0 \leqslant \hat{\mathcal{R}}_S(\mathcal{H}) \leqslant 1$ (exercise)
Intuitively, it measures how well $h \in \mathcal{H}$ correlates with random noise σ_i

---

Above definition can be generalized to real-valued functions $f : X \to \mathbb{R}$ (not just $h : X \to \{+1, -1\}$)
Fix a collection $\mathcal{F}$ of real-valued functions over X
Fix training samples $S = \{x_1, \dots, x_m\}$ over X
Redefinition **Empirical Rademacher complexity** of $\mathcal{F}$ wrt S is

$$\hat{\mathcal{R}}_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma \in \{+1,-1\}^n}\left[\sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{i \in [m]}[\sigma_i f(x_i)]\right]$$

Now fix distribution $\mathcal{D}$ over X
Rademacher complexity = average empirical rademacher complexity over m samples from $\mathcal{D}$

$$\mathcal{R}_m(\mathcal{F}) = \mathop{\mathbb{E}}_{x_1, \dots, x_m \sim \mathcal{D}}\left[\hat{\mathcal{R}}_{\{x_1, \dots, x_m\}}(\mathcal{F})\right]$$

When $\mathcal{F} = \mathcal{H}$, $\mathcal{R}_m(\mathcal{H})$ measures how expressive $\mathcal{H}$ is, much like VC dimension, but in a different way
$\mathcal{R}_m(\mathcal{H})$ depends on the distribution $\mathcal{D}$ while VC dimension is distribution-independent
 Sometimes gives better generalization bounds than VC dimension for certain distributions
$\mathcal{R}_m(\mathcal{F})$ can be defined for any family $\mathcal{F}$ of real-valued functions, not just binary classifiers
e.g. In linear regression where samples $(x, c(x))$ have a dependent variable given by target $c : X \to \mathbb{R}$
 Goal: Find linear hypothesis $h : X \to \mathbb{R}$ minimizing (say) squared loss $\mathbb{E}_{x \sim \mathcal{D}}[(h(x) - c(x))^2]$
 The corresponding $\mathcal{F} = \{(h(x) - c(x))^2 \mid \text{linear } h\}$

---

2. Generalization bound

Notation $\mathbb{E}_{\mathcal{D}}[f] = \mathbb{E}_{x\sim\mathcal{D}}[f(x)]$ for distribution $\mathcal{D}$ over X
Notation (empirical average) $\hat{\mathbb{E}}_S[f] = \mathbb{E}_{i\in[m]}[f(x_i)]$ where $S = \{x_1, \dots, x_m\}$

Theorem 1. *Let $\mathcal{F}$ be a family of functions from X to $[0,1]$, and training set $S = \{x^1, \dots, x^m\}$ where x^i are independently drawn from $\mathcal{D}$. With prob $\geqslant 1-\delta$ over S, simultaneously for all $f \in \mathcal{F}$,*

$$\mathbb{E}_{\mathcal{D}}[f] \leqslant \hat{\mathbb{E}}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{2m}}$$

Proof. Bounding $\mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_S[f]$ for all $f \in \mathcal{F}$ $\iff$ bounding $\sup_{f\in\mathcal{F}}(\mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_S[f]) =: \Phi(S)$

Claim 1 With prob $\geqslant 1-\delta$ over S, $\Phi(S) \leqslant \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln 1/\delta}{2m}}$
Proving this Claim requires McDiarmid's inequality, a generalization of Hoeffding

Lemma 2 (McDiarmid)**.** *Suppose $g : X^m \to \mathbb{R}$ satisfies, for any $x_1, \dots, x_m \in X$, $1 \leqslant i \leqslant m$, $x_i' \in X$,*

$$|g(x_1, \dots, x_i, \dots, x_m) - g(x_1, \dots, x_i', \dots, x_m)| \leqslant c_i.$$

Assume random variables $X_1, \dots, X_m$ are independent. Then for any $\varepsilon > 0$,

$$\mathbb{P}[g(X_1, \dots, X_m) \geqslant \mathbb{E}[g(X_1, \dots, X_m)] + \varepsilon] \leqslant \exp\left(-2\varepsilon^2 \Big/ \sum_{1\leqslant i\leqslant m} c_i^2\right)$$

The Claim follows from McDiarmid's with $g = \Phi, c_i = 1/m$ and $\varepsilon = \sqrt{\frac{\ln 1/\delta}{2m}}$
Why does Φ satisfy the required inequalities with $c_i = 1/m$?
Because every $f \in \mathcal{F}$, the function $S \mapsto \mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_S[f]$ satisfies those inequalities with $c_i = 1/m$
And Φ is the supremum over $f \in \mathcal{F}$ of $S \mapsto \mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_S[f]$

Claim 2 $\mathbb{E}_S[\Phi(S)] \leqslant \mathbb{E}_{S,S'}[\sup_{f\in\mathcal{F}}(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])]$ where S' is independent m samples from $\mathcal{D}$
Reason: $\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S[\sup_{f\in\mathcal{F}}(\mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_S[f])] = \mathbb{E}_S[\sup_{f\in\mathcal{F}}(\mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[f]] - \hat{\mathbb{E}}_S[f])]$
where we have used $\mathbb{E}_{\mathcal{D}}[f] = \mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[f]]$ (average of f equals expected empirical average of f)
Next $\mathbb{E}_S[\sup_{f\in\mathcal{F}}(\mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[f]] - \hat{\mathbb{E}}_S[f])] = \mathbb{E}_S[\sup_{f\in\mathcal{F}}(\mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]])]$
because moving $\hat{\mathbb{E}}_S[f]$ inside $\mathbb{E}_{S'}$ does not change its value
Finally $\mathbb{E}_S[\sup_{f\in\mathcal{F}}(\mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]])] \leqslant \mathbb{E}_{S,S'}[\sup_{f\in\mathcal{F}}(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])]$
because the supremum of an expectation is at most the expectation of the supremum

---

Claim 3 $\mathbb{E}_{S,S'}[\sup(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])] = \mathbb{E}_{S,S',\sigma}[\sup_{f\in\mathcal{F}} \mathbb{E}_{i\in[m]}[\sigma_i(f(x_i') - f(x_i))]]$
Note: S' is called the ghost sample and we use ghost sampling technique here
For each pair of elements x_i, x_i' in S, S', swap the two with probability $1/2$, and do nothing otherwise
Call the resulting two sets of samples $T = \{z_1, \dots, z_m\}, T' = \{z_1', \dots, z_m'\}$
Then S, S' and T, T' are identically distributed
Hence $\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]$ is identically distributed as $\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f]$
But $\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f] = \mathbb{E}_{i\in[m]}[f(z_i') - f(z_i)]$ is identically distributed as $\mathbb{E}_{i\in[m]}[\sigma_i(f(x_i') - f(x_i))]$
since $f(z_i') - f(z_i) = f(x_i') - f(x_i)$ if not swapped, and $= f(x_i) - f(x_i')$ if swapped
Generating (T, T') corresponds to generating (S, S', σ), so we take expectation over σ as well

---

Claim 4 $\mathbb{E}_{S,S'}[\sup_{f\in\mathcal{F}} \mathbb{E}_{i\in[m]}[\sigma_i(f(x_i') - f(x_i))]] \leqslant 2\mathcal{R}_m(\mathcal{F})$

$$\mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{i\in[m]}[\sigma_i(f(x_i') - f(x_i))]\right] \leqslant \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{i\in[m]}[\sigma_i f(x_i')] + \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{i\in[m]}[-\sigma_i f(x_i)]\right]$$

$$= \mathop{\mathbb{E}}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}[\sigma_i f(x_i')]\right] + \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}[-\sigma_i f(x_i)]\right] = \mathcal{R}_m(\mathcal{F}) + \mathcal{R}_m(\mathcal{F})$$

Last equality uses the fact that $-\sigma_i$ is identically distributed as σ_i

---

Combining the above Claims, we get the Theorem □