3.16 Summary

Goal: estimate the excess risk $L(\hat{h}) - L(h^*)$, where $\hat{h}$ is the empirical risk minimizer.

By decomposing:

$$L(\hat{h}) - L(h^*) = L(\hat{h}) - \hat{L}(\hat{h}) + \underbrace{\hat{L}(\hat{h}) - \hat{L}(h^*)}_{\leq 0} + \hat{L}(h^*) - L(h^*)$$

$$\leq L(\hat{h}) - \hat{L}(\hat{h}) + \hat{L}(h^*) - L(h^*)$$

We want to upper bound RHS by ε, then it is reasonable to require $\sup_{h \in H} |L(h) - \hat{L}(h)| \leq \frac{\varepsilon}{2}$, where we use uniform convergence

Results:

(i) Reliable, finite hypothesis:

$$P\{L(\hat{h}) > \varepsilon\} = P\{\hat{h} \in \overset{B}{\overset{\shortparallel}{\{h \in H : L(h) > \varepsilon\}}}\} \leq P\{h \in B : \hat{L}(h) = 0\} \leq |H|(1-\varepsilon)^n$$

$$\leq |H| e^{-n\varepsilon} =: \delta \quad \Rightarrow \quad L(\hat{h}) \leq (\log|H| + \log\tfrac{1}{\delta})/n \quad \text{with } p \geq 1-\delta$$

(ii) Finite hypothesis : by Hoeffding's inequality

$$P\{|L(h) - \hat{L}(h)| \geq \tfrac{\varepsilon}{2}\} = 2e^{-\frac{n\varepsilon^2}{2}} =: \frac{\delta}{|H|}, \text{ we have}$$

$$|L(h) - \hat{L}(h)| \leq \sqrt{\frac{2\log\frac{2|H|}{\delta}}{n}} \quad \text{with } p \geq 1-\delta$$

(iii) Rademacher complexity : $R_n(F) := E[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i)]$

Assume $\ell(z,h) \in [0,1]$. $G_n := \sup_{h \in H} L(h) - \hat{L}(h) = g(z_1, \cdots, z_n)$,

then g follows bounded difference condition (with $c_i := \frac{1}{n}$)

By McDiarmid's inequality, $P\{G_n \geq E(G_n) + \varepsilon\} \leq \exp\{-2n\varepsilon^2\}$.

$E(G_n) = \cdots \leq 2R_n(A)$. Set $\varepsilon = \frac{\varepsilon}{2} - E[G_n]$

$$P\{G_n \geq \tfrac{\varepsilon}{2}\} \leq \exp\{-2n(\tfrac{\varepsilon}{2} - E[G_n])^2\} \leq \exp\{-2n(\tfrac{\varepsilon}{2} - 2R_n(A))^2\} := \tfrac{\delta}{2}.$$

$$\Rightarrow \quad L(\hat{h}) - L(h^*) \leq 4R_n(A) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} \quad \text{with } p \geq 1-\delta.$$

(iv) Massart's finite lemma : Assume $\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \leq M^2 \ \forall f \in F$.

let $W_f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)$. $\exp\{tE[\sup_{f \in F} W_f | Z_{1:n}]\} \leq E[\exp\{t \sup_{f \in F} W_f\} | Z_{1:n}]$

$= E[\sup \exp\{t W_f\} | Z_{1:n}] \leq \sum E[\exp\{t W_f\} | Z \ldots]$

$= E[\sup_{f\in F} \exp\{tW_f\} \mid Z_{1:n}] \le \sum_{f\in F} E[\exp\{tW_f\} \mid Z_{1:n}]$

$\sigma_i \le 1 \xRightarrow{\text{Hoeffding's lemma}} \sigma_i$ sub-Gaussian with param. $2^2/4 = 1$.

then W_f is sub-Guassian with param. $\frac{1}{n^2}\sum_{i=1}^n f(Z_i)^2 \le \frac{M^2}{n}$.

$\Rightarrow \exp\{tW_f\} \le \exp\{\frac{t^2M^2}{2n}\}$, then we get

$$\exp\{t\hat{R}_n(F)\} \le |F| \exp\{\frac{t^2M^2}{2n}\} \Rightarrow \hat{R}_n(F) \le \frac{\log|F|}{t} + \frac{tM^2}{2n} \quad \forall t>0$$

minimizing RHS, $\hat{R}_n(F) \le \sqrt{\frac{2M^2\log|F|}{n}}$

(v) Shattering coefficient / VC dimension:

$$S(F,n) := \sup_{z_1,\cdots,z_n} |\{[f(z_1),\cdots,f(z_n)] : f\in F\}|$$

$$VC(H) := \sup\{n : S(H,n) = 2^n\}.$$

(vi) L_2 norm constrained: $F = \{z \mapsto w\cdot z : \|w\|_2 \le B_2\}$. $E[\|Z\|_2^2] \le C_2$.

$$R_n(F) = E[\sup_{\|w\|\le B_2} \frac{1}{n}\sum_{i=1}^n \sigma_i w\cdot Z_i] \le \frac{1}{n}E[\sup_{\|w\|\le B_2} \|w\|_2 \|\sum_{i=1}^n \sigma_i Z_i\|_2]$$

$$\le \frac{B_2}{n} E[\|\sum_{i=1}^n \sigma_i Z_i\|_2] \le \frac{B_2}{n}\sqrt{E[\|\sum_{i=1}^n \sigma_i Z_i\|_2^2} = \frac{B_2}{n}\sqrt{E[\sum_{i=1}^n \|Z_i\|^2]}$$

$$\le \frac{B_2C_2}{\sqrt{n}}.$$

L_1 norm constrained: $\|Z\|_\infty \le C_\infty$ then let $W = \bigcup_{j=1}^d \{B_1e_j, -B_1e_j\}$

$F = \{z \mapsto w\cdot z : \|w\|_1^2 = B_1^2\}$ is convex hull of W.

Since $w\cdot Z_i \le \|w\|_1 \|Z_i\|_\infty \le B_1C_\infty$. By Massart's finite lemma:

$$R_n(F) = R_n(W) \le B_1C_\infty\sqrt{\frac{2\log|W|}{n}} = B_1C_\infty\sqrt{\frac{2\log 2d}{n}}$$

(vii) Simple discretization:

$$R_n(F) = E[\sup_{f\in F}\langle\sigma, f\rangle] \le E[\sup_{g\in C}\langle\sigma, g\rangle + \varepsilon] \le \sqrt{\frac{2\log N(\varepsilon, F, L_2(P_n))}{n}} + \varepsilon$$

Chaining:

$$R_n(F) \le \int_0^{C_0}\sqrt{\frac{\log N(\varepsilon, F, L_2(P_n))}{n}}\, d\varepsilon.$$ C_0 is the coastest resolution,

where $L_2(P_n)$ is the L_2 distance w.r.t. the empirical distrib.

over n data: $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. Let $\rho = L_2(P_n)$, then

$$\rho(f,f') = \left(\frac{1}{n}\sum_{i=1}^n (f(z_i) - f'(z_i))^2\right)^{\frac{1}{2}}$$

Techniques:

(i) Hoeffding's inequality: by Hoeffding's lemma, it's obvious.

(ii) McDiarmid's inequality: by sub-Guassian martingale lemma and constructing $L_i \le Z_i - Z_{i-1} \le U_i$. $U_i - L_i \le c_i$, we proved the ineq.

Others:

(i) Algorithm stability. For an algorithm A:

uniform stability β: if $|\ell(z_0, A(S)) - \ell(z_0, A(S^i))| \le \beta \quad \forall z_0, S, S^i$.

Generalization under uniform stability: by McDiarmid's inequality. for $\forall$ A with β, if $|\ell(z,h)| \le M$, then with prob. $\ge 1-\delta$

$$L(A(S)) \le \hat{L}(A(S)) + \beta + (\beta n + M)\sqrt{\frac{2\log(1/\delta)}{n}}$$

(ii) PAC-Bayesial bounds:

When prior $P(h)$ and Posterior $Q_S(h)$ are given

Occam bound: if H countable, $\ell(z,h) \in [0,1]$, with $p \ge 1-\delta$

$$\forall h \in H: L(h) \le \hat{L}(h) + \sqrt{\frac{\log(1/p(h)) + \log(1/\delta)}{2n}}$$

The difference with finite hypothesis is when we consider union bound.

PAC-Bayesian theorem: with $p \ge 1-\delta$

$$E_{h\sim Q_S}[L(h)] \le E_{h\sim Q_S}[\hat{L}(h)] + \sqrt{\frac{KL(Q_S \| P) + \log(4n/\delta)}{2n-1}}$$