generalization error bounds

for binary classification,
a classical PAC bound is:

\[
\forall w \in W \quad \mathbb{E}[L_1(w)] \leq \mathbb{E}[\hat{L}_n(w)] + \frac{1}{\sqrt{n}} \sqrt{d \log d + \log \frac{2}{\delta}}
\]

where \( d \) is VC-dimension

\[d = \mathbb{E} \mathcal{E}_{h,w} : w \in W\]

VC-dimension of \( \mathcal{H} \) is defined as:

\[
\text{max} \, \exists \, m \geq 3 \, \text{a set of}\, m \, \text{pts., s.t.} \, \forall \text{labellings of these points,} \\
\quad \exists w \text{ at some gain the correct label on these points,}
\]

\[
\Rightarrow \text{shattering the set of pts.}
\]

for linear classifiers of \( p \) parameters, VC-dim = \( p + 1 \)

one issue for this bound is that it is true for all distributions \( \Rightarrow \) too loose bound

\( \Rightarrow \) motivates going to data distribution dependent
\[ \hat{R}_{\text{emp}}(\mathcal{H}) \triangleq \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\sigma_i \neq h(x_i)} \right] \]

Bound:

\[ \forall w, \quad L(w) \leq \hat{L}_n(w) + \hat{R}_{\text{emp}}(\mathcal{H}) + \frac{1}{\sqrt{n}} \sqrt{\log{|\mathcal{H}|}} \]

High-level idea to prove bound:

"double sample trick" → use a second sample \( \mathcal{D}_n \)

\[ L(w) = \mathbb{E}_{\mathcal{D}_n} [\hat{L}_n(w)] \]

"symmetrization trick" → bound \( \sup \) of differences between \( L(w) \) and \( \hat{L}_n(w) \)

Union bound as usual + concentration inequality

Structured prediction generalization bounds [Cortes et al. NIPS 2016]
general loss \( l(y, \hat{y}) \) st. \( l(y, \hat{y}) = 0 \) if \( y = \hat{y} \)

Suppose \( s(x, y) = \sum_{c \in C(x, y)} \) a set of cliques of a graph which model factor factor

\[ \forall w \in W \quad L(w) \leq \text{hinge}(w) + 4 \sqrt{2} \frac{\hat{e}_n}{\sqrt{n}} + 3 \frac{\max_i \sigma_i}{\sqrt{n}} \]

where \( \hat{e}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \text{sup}_{w \in W} \hat{s}_{C_i, c, y} \) is only depends \( (x(i), y) \) on \( l(y, \hat{y}) \)

\[ \text{empirical factor graph complexity} \]

\[ \text{set of cliques} \quad (x, y) \]

Thm. 2: if \( s(x, y, c, w) = = \langle w, \Phi_c(x, y, c) \rangle \)

and consider \( W_n \triangleq \{ w : \| w \|_2 \leq \sqrt{L} \} \)

Let \( R = \max_i \| \Phi_c(x, y, c) \|_2 \)

then \( \hat{e}_n \leq \frac{R \sqrt{L}}{\sqrt{n}} \frac{1}{\sqrt{L}} \max_{i, c, y} \sigma_i \)

\[ \text{so want small cliques} \]

back to bound: \( L(w) \leq \text{hinge}(w) + \frac{R \sigma_i}{\sqrt{n}} \)

IFT6132 Page 3
back to bound: \[ L(w) \leq \text{Shage}(w) + \left( \frac{K \| \theta \|_1}{\min_c \text{Shage}(c)} \right) \frac{1}{\||w||^2_2} \]

\[ \min_{w} \text{RHS suggests} \]

\[ \text{sum struct alg. } \quad w_n^* = \text{argmin}_{w} \text{Shage}(w) + \frac{1}{2} \|w\|^2_2 \]

missing link: \[ \min_{w} f(w) \]

\[ \text{If } f \text{ is convex } \Rightarrow \text{for } \lambda > 0 \text{ s.t. } \]

\[ \min_{w} f(w) + \frac{1}{2} \|w\|^2_2 \]

(\text{use Lagrangian duality})

gives same solution as (1)

\[ \text{sidestep: penalized formulation has sensitive to choice of } \lambda \]

\[ \text{vs. constrained formulation} \]

\[ \text{can think of sum struct as minimizing upper bound on generalization error} \]

\[ \text{properties:} \]

\[ \text{minimizes upper bound, hope that minimizes } L(w) \]

\[ \text{but no general guarantees?} \]

\[ \text{can evaluate bound to get guarantees} \]

\[ \text{not consistent} \]
next: consistency

Consistency & calibration for $f^*$:

Need to relate $L(w)$ to $L^*(w)$ via “calibration function” $[\text{Steinwart}]

relationship is usually very complicated

$\Rightarrow$ current results hold mainly at \underline{non-parametric setting} (no # of parameters)

all functions $h: A \rightarrow Y$ are considered $\Rightarrow$ this encapsulates the dependence on $x$

in the analysis

i.e. we suppose that $S(x,y,w)$ can be arbitrary for any $x$

(i.e. $w$ is $\Theta$-dim)

$\Rightarrow$ could use a \underline{universal kernel}

$h(x,w) \in \mathcal{H}_w$

\underline{R.KHS}

Motivation: generalize $<w, \phi(x)>$ to higher dim. space

$\Phi(A)$ + kernel trick $<\phi(x), \phi(x')> = k(x, x')$

\[ \Phi: A \rightarrow \mathbb{R}^3 \]

\[ \Phi(x) = \left( \frac{x_1^2}{3}, \frac{x_2}{3}, -x_3 \right) \]
\[ \text{polynomial kernel } \phi \left( \langle x, x' \rangle + 1 \right)^d = \bar{k}(x, x') \]

\[ \text{equivalent to mapping data to space of dimension exponential in } d \]

\[ \text{even have } \infty \text{-dim., e.g. } k(x, x') = \exp \left( -\frac{\|x - x'\|^2}{2} \right) \text{ (RBF kernel)} \]

**RKHS** (Reproducing Kernel Hilbert Space)

\[ \Phi : \mathcal{X} \to \mathcal{H} \text{ s.t. } \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = k(x, x') \text{ (uniquely defined by RKHS)} \]

let \( \tilde{\mathcal{H}} = \text{span} \{ k(x, \cdot) : x \in \mathcal{X} \} \)

e.g. \( f \in \tilde{\mathcal{H}} \Rightarrow f = \sum_{i} \alpha_i k(x_i, \cdot) \) for some unique \( \{ x_i, \alpha_i \}_{i=1}^{n} \)

\( \sum_{i=1}^{n} \alpha_i^2 \)

\[ \| f \|_{\tilde{\mathcal{H}}} \leq \sqrt{\sum_{i} \alpha_i^2} \]

\[ \text{then RKHS } \mathcal{H} \text{ is } \text{completion} \left( \tilde{\mathcal{H}} \right) \text{ using } \| \cdot \|_\mathcal{H} \text{ as you norm} \]
then RKHS $H$ is $= \text{completion}(\tilde{H})$ using $\| \cdot \|_H$ as you norm

i.e. add all limit points of $\tilde{H}$- Cauchy sequences to get $H$

you could think of $f(x) = \sum_{i=1}^{\infty} c_i k(x, x_i)$

"reproducing" property of $H$: for $f \in H$

$< f, k(x, x_i) >_H = f(x)$

nice property of RKHS, func. evaluation is $c_{\mathbf{r}}$

mapping $E_x: \mathcal{H} \rightarrow \mathbb{R}$

$E_x(f) = f(x)$

$|f(x) - g(x)| = |< f - g, k(x, \cdot) >|_H$

$\leq \| f - g \|_H \cdot \| k(x, \cdot) \|_H$ i.e. $E_x$ is Lipschitz $c_{\mathbf{r}}$

this property

important to do statistics