Stability and Generalization in Pattern Recognition

Fateme Sheikholeslami

Dept. of ECE and DTC, University of Minnesota

April, 2016

1. O. Bousquet, and A. Elisseeff. "Stability and generalization," *Journal of Machine Learning Research,* vol. 2, pp. 499-526, Mar 2002.
2. J. Shawe-Taylor, and N. Cristianini, *Kernel methods for pattern analysis,* Cambridge University Press, 2004.

Road map

- Overview on pattern recognition
- Concentration of a fixed function
  - McDiarmid inequality
  - Hoeffding's inequality
- Concentration of a class of functions
  - Capacity and regularization
    - Rademacher complexity
- Stable algorithms
- Generalization bounds
  - Polynomial bounds
  - Exponential bounds
- Stability and generalization of regularized RKHS learning

Pattern recognition

- Choose a function from a class of functions which achieves a certain objective
  - Often interested in $\min_{f \in \mathcal{F}} \mathbb{E}[f(\mathbf{x})]$

- Considerations
  - $\{\mathbf{x}_i\}_{i=1}^N$ is drawn from an unknown pdf
  - $\mathbb{E}[f(\mathbf{x})]$ is approximated by its empirical value $\hat{\mathbb{E}}[f(\mathbf{x})] := \frac{1}{N} \sum_{\mathbf{x}_i \in \mathcal{S}} f(\mathbf{x}_i)$ on a "training" set $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$

- Performance
  - What conclusion can be made about $\mathbb{E}[f(\mathbf{x})]$ based on its empirical measure?

Concentration of a fixed function on a finite dataset

Question 1

How concentrated a fixed function of a finite dataset $h(X_1, ..., X_M) \in \mathbb{R}$ is around its mean?

McDiarmid's Inequality

Let $X_i \in \mathcal{A}$ denote independent random variables, and assume $h : \mathcal{A}^N \to \mathbb{R}$

If $\sup_{x_1,...,x_N,\hat{x}_i \in \mathcal{A}} |h(x_1, ..., x_N) - h(x_1, ..., \hat{x}_i, ..., x_N)| \leq c_i, \quad 0 \leq c_i \leq N$

$$\Longrightarrow \quad \forall \epsilon > 0 \quad \Pr\Big(h(x_1, ..., x_M) - \mathbb{E}[h(x_1, ..., x_N)] \geq \epsilon\Big) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^N c_i^2}\right)$$

Hoeffdings's Inequality

If $X_1, ..., X_N$ are independent r.v. satisfying $X_i \in [a_i, b_i]$, then for the r.v.

$S_N := \sum_{i=1}^N X_i$, we have $\quad \forall \epsilon > 0 \quad \Pr\Big(S_N - \mathbb{E}[S_N] \geq \epsilon\Big) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right)$

Example: concentration of the sum of a finite dataset

- Example: consider $S_N(\mathbf{x}_1, ..., \mathbf{x}_N) = \frac{1}{N}\sum_{i=1}^N x_i = \hat{\mathbb{E}}[X]$ where $x_i \in [a, b]$.

$$|S_N(x_1, ..., x_N) - S_N(x_1, ..., \hat{x}_i, ..., x_N)| \leq (b-a)/N \implies \Pr\Big(|\hat{\mathbb{E}}[X] - \mathbb{E}[X]| \geq \epsilon\Big) \leq 2\exp\left(\frac{-2N\epsilon^2}{(b-a)^2}\right)$$

- Example: consider the center of mass for the sample set $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$

$$\phi_{\mathcal{S}} := \frac{1}{N}\sum_{i=1}^N \phi(\mathbf{x}_i)$$

  - What can be concluded about its concentration?

    Measure of accuracy $g(\mathcal{S}) := \|\boldsymbol{\phi}_{\mathcal{S}} - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})]\|$

- Can we apply McDiarmid's Inequality? $\hat{\mathcal{S}} = \{\mathbf{x}_1, ..., \hat{\mathbf{x}}_i, ..., \mathbf{x}_N\}$

$$|g(\mathcal{S}) - g(\hat{\mathcal{S}})| = \|\boldsymbol{\phi}_{\mathcal{S}} - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\| - \|\boldsymbol{\phi}_{\hat{\mathcal{S}}} - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\| \leq \|\boldsymbol{\phi}_{\mathcal{S}} - \boldsymbol{\phi}_{\hat{\mathcal{S}}}\| = \frac{1}{N}\|\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\hat{\mathbf{x}}_i)\| \leq \frac{2R}{N}$$

$\|\phi(\mathbf{x})\| \leq R$

Yes! $\Longrightarrow$ $$\Pr\Big(g(\mathcal{S}) - \mathbb{E}_{\mathcal{S}}[g(\mathcal{S})] \geq \epsilon\Big) \leq \exp\left(\frac{-2N\epsilon^2}{4R^2}\right)$$

Example: concentration of sample center of mass in feature space

- Furthermore

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}[g(\mathcal{S})] &= \mathbb{E}_{\mathcal{S}}[\|\boldsymbol{\phi}_{\mathcal{S}} - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})]\|] = \mathbb{E}_{\mathcal{S}}[\|\boldsymbol{\phi}_{\mathcal{S}} - \mathbb{E}_{\tilde{\mathcal{S}}}[\boldsymbol{\phi}_{\tilde{\mathcal{S}}}]\|] \\
&= \mathbb{E}_{\mathcal{S}}[\|\mathbb{E}_{\tilde{\mathcal{S}}}[\boldsymbol{\phi}_{\mathcal{S}} - \boldsymbol{\phi}_{\tilde{\mathcal{S}}}]\|] \leq \mathbb{E}_{\mathcal{S}\tilde{\mathcal{S}}}[\|\boldsymbol{\phi}_{\mathcal{S}} - \boldsymbol{\phi}_{\tilde{\mathcal{S}}}\|] \\
&= \mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}\tilde{\mathcal{S}}}\left[\frac{1}{N}\left\|\sum_{i=1}^{N}\sigma_i(\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\tilde{\mathbf{x}}_i))\right\|\right] \qquad \text{where } \sigma_i \in \{\pm 1\} \text{ with equal probability} \\
&= \mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}\tilde{\mathcal{S}}}\left[\frac{1}{N}\left\|\sum_{i=1}^{N}(\sigma_i\boldsymbol{\phi}(\mathbf{x}_i) - \sigma_i\boldsymbol{\phi}(\tilde{\mathbf{x}}_i))\right\|\right] \leq 2\mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}}\left[\frac{1}{N}\left\|\sum_{i=1}^{N}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i)\right\|\right] \\
&= \frac{2}{N}\mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}}\left[\left(\left\langle\sum_{i=1}^{N}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{N}\sigma_j\boldsymbol{\phi}(\mathbf{x}_j)\right\rangle\right)^{1/2}\right] \\
&\leq \frac{2}{N}\left(\mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}}\left[\sum_{i,j=1}^{N}\sigma_i\sigma_j\kappa(\mathbf{x}_i, \mathbf{x}_j)\right]\right)^{1/2} = \frac{2}{N}\left(\mathbb{E}_{\boldsymbol{\sigma}\mathcal{S}}\left[\sum_{i=1}^{N}\kappa(\mathbf{x}_i, \mathbf{x}_j)\right]\right)^{1/2} \leq \frac{2R}{\sqrt{N}}
\end{aligned}$$

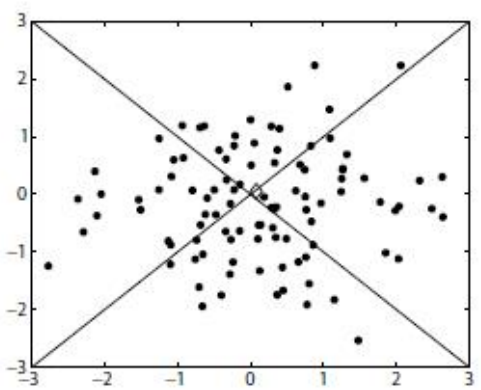
- Previously, we had $\Pr\Big(g(\mathcal{S}) - \mathbb{E}_{\mathcal{S}}[g(\mathcal{S})] \geq \epsilon\Big) \leq \exp\left(\frac{-2N\epsilon^2}{4R^2}\right)$

- Setting $\delta := \exp\left(\frac{-2N\epsilon^2}{4R^2}\right)$ and after substitution, with probability at least $1-\delta$ we have

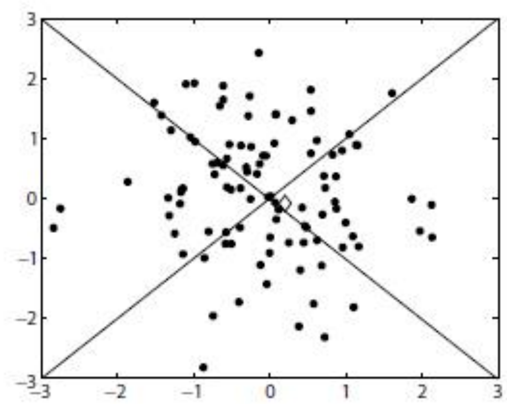
$$g(\mathcal{S}) \leq \frac{R}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right)$$

Example: concentration of sample mean

- Sample mean of random draws of 2-dimensional Gaussian variables



The empirical centre of mass based on a random sample



. The empirical centre of mass based on a second random sample.

A random variable that depends (in a "smooth" way) on the influence of many independent variables (but not too much on any of them) is essentially constant.

Talagrand 1996.

Capacity of a class

- Let us go back to pattern recognition
  - Find the function from a class of functions which achieves a certain objective
    - For instance $\min_{f \in \mathcal{F}} \hat{\mathbb{E}}[f(x,y)]$

Question 2: How concentrated is empirical mean of the sought pattern to its true mean?

- Example
  - Find a function $f \in P_{10}$ that maps creditcard numbers to the card holder's phone number

    $P_{10} :=$ Set of polynomials of degree 10
  - Given 10 training pairs, $\exists f \in P_{10}$ such that perfectly maps the training points!
  - Performance on unseen data? Arbitrarily poor! $\Rightarrow$ overfitting!

❖ Performance of a pattern relies on
(1) concentration of the function value
(2) Richness (capacity) of the class

Rademacher Complexity

- Measures the capacity of a class by its ability to fit random data
- Let $\{\sigma_1, ..., \sigma_N\}$ be independent uniform $\{\pm 1\}$ -valued Rademacher r.v.
- For set $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, define Empirical Rademacher complexity of class $\mathcal{F}$ as

$$\hat{R}_N(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{2}{N}\sum_{i=1}^{N}\sigma_i f(\mathbf{x}_i)\Big| \;\Big|\; \mathbf{x}_1, ..., \mathbf{x}_N\Big]$$

- Rademacher complexity of $\mathcal{F}$

$$R_N(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}[\hat{R}_N(\mathcal{F})] = \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{2}{N}\sum_{i=1}^{N}\sigma_i f(\mathbf{x}_i)\Big|\Big]$$

Rademacher Complexity of kernel-based functions

- Consider the class of linear functions in a kernel defined feature space

$$\mathcal{F} := \{f | f : \mathbf{x} \to \sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}), \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq B^2\}$$

- Consider the class $\mathcal{F}_B := \{f | f : \mathbf{x} \to \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle, \|\mathbf{w}\| \leq B\}$ Regularization!

If $\kappa : X \times X \to \mathbb{R}$ is a kernel, and $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ is a sample of points, then the empirical Rademacher complexity of the class $\mathcal{F}_B$ satisfies

$$\hat{R}_N(\mathcal{F}_B) \leq \frac{2B}{N} \sqrt{\sum_{i=1}^{N} \kappa(\mathbf{x}_i, \mathbf{x}_i)} = \frac{2B}{N} \sqrt{\operatorname{tr}(\mathbf{K})}$$

Proof:

$$\hat{R}_N(\mathcal{F}_B) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}_B} \left| \frac{2}{N} \sum_{i=1}^{N} \sigma_i f(\mathbf{x}_i) \right| \right] = \frac{2}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\| \leq B} \left| \left\langle \mathbf{w}, \sum_{j=1}^{N} \sigma_j \boldsymbol{\phi}(\mathbf{x}_j) \right\rangle \right| \right]$$

$$\leq \frac{2B}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{\phi}(\mathbf{x}_i) \right\| \right] = \frac{2B}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left(\left\langle \sum_{i=1}^{N} \sigma_i \boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{N} \sigma_j \boldsymbol{\phi}(\mathbf{x}_j) \right\rangle \right)^{1/2} \right]$$

$$= \frac{2B}{N} \left(\mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{i,j=1}^{N} \sigma_i \sigma_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \right] \right)^{1/2} = \frac{2B}{N} \sqrt{\sum_{i=1}^{N} \kappa(\mathbf{x}_i, \mathbf{x}_i)} \qquad . \blacksquare$$

Properties of Rademacher complexity

- Theorem: Let $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_m$ and $\mathcal{G}$ be classes of real functions. Then

  a) If $\mathcal{F} \subseteq \mathcal{G}$ then $\hat{R}_N(\mathcal{F}) \le \hat{R}_N(\mathcal{G})$

  b) $\hat{R}_N(\mathcal{F}) = \hat{R}_N(\mathrm{conv}\mathcal{F})$

  c) For every $c \in \mathbb{R}$, $\hat{R}_N(c\mathcal{F}) = |c|\hat{R}_N(\mathcal{F})$

  d) If $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and satisfies $\mathcal{A}(0) = 0$, then $\hat{R}_N(\mathcal{A}\mathrm{o}\mathcal{F}) \le 2L\hat{R}_N(\mathcal{F})$

  d) For any function $h, \hat{\mathcal{R}}_N(\mathcal{F} + h) \le \hat{\mathcal{R}}_N(\mathcal{F}) + 2\sqrt{\hat{\mathbb{E}}[h^2]/N}$

  e) For any $1 \le q < \infty$, let $\mathcal{L}_{\mathcal{F},h,q} = \{|f - h|^q | f \in \mathcal{F}\}$. If $\|f - h\|_\infty \le 1$

     for every $f \in \mathcal{F}$, then $\hat{R}_N(\mathcal{L}_{\mathcal{F},h,q}) \le 2q\left(\hat{R}_N(\mathcal{F}) + 2\sqrt{\hat{\mathbb{E}}[h^2]/N}\right)$.

  f) $\hat{R}_N(\sum_{i=1}^m \mathcal{F}_i) \le \sum_{i=1}^m \hat{R}_N(\mathcal{F}_i)$

Concentration of a class of functions

Fix $\delta \in (0,1)$, and let $\mathcal{F} := \{f | f : X \to [0,\ 1]\}$. Let $\{\mathbf{x}_i\}_{i=1}^N$ be ind. drawn from distribution $\mathcal{D}$. Then, w.p. at least $1-\delta$ over random draws of sample size N

$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] \leq \hat{\mathbb{E}}[f(\mathbf{x})] + R_N(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2N}} \qquad \forall f \in \mathcal{F}$$

$$\leq \hat{\mathbb{E}}[f(\mathbf{x})] + \hat{R}_N(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2N}}$$

- *Sketch of proof*

For a fixed f: $$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] \leq \hat{\mathbb{E}}_{\mathbf{x}}[f(\mathbf{x})] + \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathbf{x}}h - \hat{\mathbb{E}}h\right)$$

Applying McDiarmid's ineq. on the second term (why?), w.p. at least $1-\delta/2$

$$\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathbf{x}}h - \hat{\mathbb{E}}h\right) \leq \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathbf{x}}h - \hat{\mathbb{E}}h\right)\right] + \sqrt{\frac{\ln(2/\delta)}{2N}} \qquad \delta/2 := \exp\left(-2N\epsilon^2\right)$$

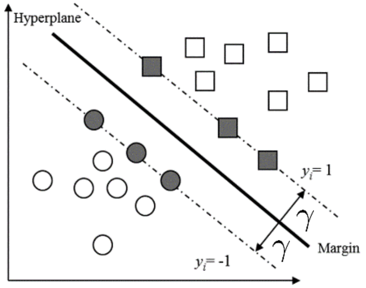
$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] \leq \hat{\mathbb{E}}[f(\mathbf{x})] + \mathbb{E}_{\mathcal{S}}\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathbf{x}}h - \hat{\mathbb{E}}h\right)\right] + \sqrt{\frac{\ln(2/\delta)}{2N}}$$

Can be shown: $$\leq R_N(\mathcal{F}) \leq \hat{R}_N(\mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2N}}$$

w.p. at least $1-\delta/2$. ■

Concentration of kernel-based SVM classifier

- Given a function $g(\mathbf{x})$, a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, a desired margin γ
- Define slack variable



$$\xi_i := (\gamma - y_i g(\mathbf{x}_i))_+ = \begin{cases} \gamma - y_i g(\mathbf{x}_i) & \gamma - y_i g(\mathbf{x}_i) > 0 \\ 0 & \text{otherwise} \end{cases}$$

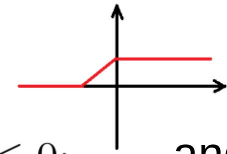
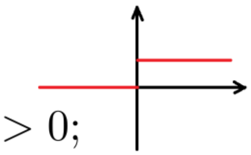
Theorem

Fix $\gamma > 0$, and let $\mathcal{F} := \{f | f(\mathbf{x}, y) = -y g(\mathbf{x}),\ g(\mathbf{x}) = \langle \boldsymbol{\phi}(\mathbf{x}), \mathbf{w} \rangle,\ \|\mathbf{w}\|_{\mathcal{H}} \leq 1\}$.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be ind. drawn from distribution $\mathcal{D}$. Then, w.p. at least $1 - \delta$ over ind. draws of sample size N we have

$$\mathbb{P}\Big(y \neq \operatorname{sign}\big(g(\mathbf{x})\big)\Big) \leq \frac{1}{N\gamma} \sum_{i=1}^{N} \xi_i + \frac{4}{N\gamma}\sqrt{\operatorname{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2N}}$$

Sketch of proof

- Define $\mathcal{P}(a) = \begin{cases} 1, & \text{if} \quad a > 0; \\ 1 + a/\gamma, & \text{if} \quad -\gamma \le a \le 0; \\ 0, & \text{otherwise} \end{cases}$  and $\mathcal{H}(a) = \begin{cases} 1, & \text{if} \quad a > 0; \\ 0, & \text{otherwise} \end{cases}$ 

- Since $\mathcal{P}(a)$ dominates $\mathcal{H}(a)$, we have

$$\mathbb{E}_{\mathbf{x}}[\mathcal{H}(f(\mathbf{x}, y)) - 1] \le \mathbb{E}_{\mathbf{x}}[\mathcal{P}(f(\mathbf{x}, y)) - 1]$$
$$\le \hat{\mathbb{E}}_{\mathbf{x}}[\mathcal{P}(f(\mathbf{x}, y)) - 1] + \hat{R}_{\ell}((\mathcal{P} - 1) o \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2N}}$$

- $(\mathcal{P} - 1)(a)$ is Lipschitz continuous with $L = 1/\gamma$ and $(\mathcal{P} - 1)(0) = 0$

$$\mathbb{P}\Big(y \ne \text{sign}\big(g(\mathbf{x})\big)\Big) = \mathbb{E}_{\mathbf{x}}[\mathcal{H}(f(\mathbf{x}, y))] \le \underbrace{\hat{\mathbb{E}}_{\mathbf{x}}[\mathcal{P}(f(\mathbf{x}, y))]}_{\le \xi_i/\gamma} + 2\underbrace{\hat{R}_{\ell}(\mathcal{F})}_{\hat{R}_{\ell}(\mathcal{G}) \le \frac{2}{N}\sqrt{\text{tr}(\mathbf{K})}}/\gamma + 3\sqrt{\frac{\ln(2/\delta)}{2N}}$$

$\blacksquare$

Algorithms

- Assume data is given as $\mathcal{S} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$ where $\mathbf{z}_i := (\mathbf{x}_i, y_i)$
- Loss functions are usually of interest $f(\mathbf{z}) = \ell(g, \mathbf{z})$ where, e.g. $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} \subset \mathcal{G}$
- Define Algorithm $A : \mathcal{Z}^N \rightarrow \mathcal{G}$ that maps dataset $\mathcal{S}$ into a function $A_{\mathcal{S}} \subset \mathcal{G} : \mathcal{X} \rightarrow \mathcal{Y}$

  e.g., $A_{\mathcal{S}} = \arg\min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \ell(g, \mathbf{z}_i) + \lambda \|g\|_{\mathcal{H}}^2$

- Define "Risk" functions

$$R(A, \mathcal{S}) := \mathbb{E}_{\mathbf{z}}[\ell(A_{\mathcal{S}}, \mathbf{z})]$$

$$R_{emp}(A, \mathcal{S}) := \sum_{i=1}^{N} \ell(A_{\mathcal{S}}, \mathbf{z}_i)/N$$

$$R_{loo}(A, \mathcal{S}) := \sum_{i=1}^{N} \ell(A_{\mathcal{S}^{\backslash i}}, \mathbf{z}_i)/N$$

- So far, we have studied

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f] - \hat{\mathbb{E}}[f] \right| > \epsilon \right)$$

- Different approaches study

$$\mathbb{P}\left(\left| R(A, \mathcal{S}) - R_{emp}(A, \mathcal{S}) \right| > \epsilon \right)$$

while assuming a notion of "stability" for the algorithm A.

Algorithm stability

D1) Algorithm A has *pointwise hypothesis stability* β_N if

$$\forall i \in \{1, ..., N\},\ \mathbb{E}_{\mathcal{S}}\Big[\left|\ell(A_{\mathcal{S}}, \mathbf{z}_i) - \ell(A_{\mathcal{S}^{\setminus i}}, \mathbf{z}_i)\right|\Big] \leq \beta_N$$

D2) Algorithm A has *hypothesis stability* β_N if

$$\forall i \in \{1, ..., N\},\ \mathbb{E}_{\mathcal{S},\mathbf{z}}\Big[\left|\ell(A_{\mathcal{S}}, \mathbf{z}) - \ell(A_{\mathcal{S}^{\setminus i}}, \mathbf{z})\right|\Big] \leq \beta_N$$

D3) Algorithm A has *uniform stability* β_N if

$$\forall \mathcal{S} \in \mathcal{Z}^N,\ \forall i \in \{1, ..., N\},\ \max_{\mathbf{z} \in \mathrm{supp}(\mathcal{D})} \left|\ell(A_{\mathcal{S}}, \mathbf{z}) - \ell(A_{\mathcal{S}^{\setminus i}}, \mathbf{z})\right| \leq \beta_N$$

- Algorithm A is considered *stable* if β_N decreases as $1/N$.

Polynomial bounds with hypothesis stability

Theorem

For Algorithm A with hypothesis stability β_1 and pointwise stability β_2 w.r.t. a loss function $0 \leq \ell(A_{\mathcal{S}}, \mathbf{z}) \leq M$, w.p. at least $1-\delta$ we have

$$R(A, \mathcal{S}) \leq R_{emp}(A, \mathcal{S}) + \sqrt{\frac{M^2 + 12MN\beta_2}{2N\delta}}$$

and

$$R(A, \mathcal{S}) \leq R_{loo}(A, \mathcal{S}) + \sqrt{\frac{M^2 + 6MN\beta_1}{2N\delta}}$$

Exponential bounds with uniform stability

- Consider a regression task

Theorem

For Algorithm A with uniform stability β w.r.t. a loss function $0 \leq \ell(A_{\mathcal{S}}, \mathbf{z}) \leq M$,

w.p. at least $1-\delta$ we have

$$R(A,\mathcal{S}) \leq R_{emp}(A,\mathcal{S}) + 2\beta + (4N\beta + M)\sqrt{\frac{\ln(1/\delta)}{2N}}$$

and

$$R(A,\mathcal{S}) \leq R_{loo}(A,\mathcal{S}) + \beta + (4N\beta + M)\sqrt{\frac{\ln(1/\delta)}{2N}}$$

- The bound is tight if β scales as $1/N$.
- Specialized bounds for classification task is also available.

Question

Are commonly-used learning algorithms stable?

Uniform stability of regularized RKHS learning

- Consider the class of linear functions in a kernel defined feature space $\mathcal{G}$

Definition: Loss function $\ell(g, \mathbf{z})$ on $\mathcal{G} \times \mathcal{Y}$ is σ-admissible w.r.t. $\mathcal{G}$ if the associated cost $\ell(g, \mathbf{z}) = c(g(\mathbf{x}), y)$ is convex w.r.t. its first argument, and

$$\forall y_1, y_2 \in \mathcal{D}, \forall y' \in \mathcal{Y}, |c(y_1, y') - c(y_2, y')| \leq \sigma |y_1 - y_2|$$

where $\mathcal{D} = \{y | \exists g \in \mathcal{G}, \exists \mathbf{x} \in \mathcal{X} : g(\mathbf{x}) = y\}$.

Theorem

Assume for given kernel $\kappa(\mathbf{x}, \mathbf{x}) \leq \kappa^2 < \infty$, and let loss $\ell(g, \mathbf{z})$ be σ-admissible w.r.t. $\mathcal{G}$. Then the learning algorithm A defined by

$$A_{\mathcal{S}} = \arg\min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \ell(g, \mathbf{z}_i) + \lambda \|g\|_{\mathcal{H}}^2$$

has uniform stability $\beta \leq \dfrac{\sigma^2 \kappa^2}{2\lambda N}$.

Examples on regularized RKHS learning

- Regularized RKHS learning $A_{\mathcal{S}} = \arg\min_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^{N} \ell(g,\mathbf{z}_i) + \lambda\|g\|_{\mathcal{H}}^2$

- Ex1) Bounded SVM regression
  - $\ell(g,\mathbf{z}) = |g(\mathbf{x}) - y|_\epsilon = \begin{cases} 0 & \text{if } |g(\mathbf{x}) - y| \le \epsilon \\ |g(\mathbf{x}) - y| - \epsilon & \text{otherwise} \end{cases}$ and $\mathcal{Y} = [0, B]$.
  - $\ell(g,\mathbf{z})$ is 1-admissible $\Rightarrow \beta \le \frac{\kappa^2}{2\lambda N} \Rightarrow R \le R_{emp} + \frac{\kappa^2}{\lambda N} + \left(\frac{2\kappa^2}{\lambda} + \kappa\sqrt{\frac{B}{\lambda}}\right)\sqrt{\frac{\ln(1/\delta)}{2N}}$

- Ex2) Regularized least squares
  - $\ell(g,\mathbf{z}) = (g(\mathbf{x}) - y)^2$ and $\mathcal{Y} = [0, B]$.
  - $\ell(g,\mathbf{z})$ is $2B$-admissible $\Rightarrow \beta \le \frac{2B^2\kappa^2}{\lambda N} \Rightarrow R \le R_{emp} + \frac{4\kappa^2\beta^2}{\lambda N} + \left(\frac{8\kappa^2 B^2}{\lambda} + 2B\right)\sqrt{\frac{\ln(1/\delta)}{2N}}$

Summary

- Concentration of a fixed function
  - McDiarmid inequality
  - Hoeffding's inequality

- Concentration of a class of functions
  - Capacity and regularization
    - Rademachar complexity

- Algorithm stability

- Generalization bounds
  - Polynomial bounds
  - Exponential bounds

- Regularized RKHS learning

Thank You!