Statistical Learning Theory

Lijun Zhang

Nanjing University, China

August 13, 2017

Outline

Outline

Introduction

Nothing is more practical than a good theory!

[Vapnik, 1998]

Introduction

Nothing is more practical than a good theory!

[Vapnik, 1998]

- The Goal
  - Provide a framework for studying the problem of inference
- The Methodology
  - Assumptions of statistical nature about the underlying phenomena
- No Free Lunch
  - If there is no assumption on how the past (i.e. training data) is related to the future (i.e. test data), prediction is impossible.
  - If there is no a priori restriction on the possible phenomena that are expected, it is impossible to generalize and there is thus no better algorithm

Outline

Formalization [Bousquet et al., 2004]

- Notations
  - Input space: $\mathcal{X}$, Output Space $\mathcal{Y} = \{-1, 1\}$
  - Function: $g : \mathcal{X} \mapsto \mathcal{Y}$
  - Risk of g:

$$R(g) = \Pr(g(X) \neq Y) = \mathrm{E}\left[\mathbb{1}_{g(X)\neq Y}\right]$$

  - Regression function:

$$\eta(\mathbf{x}) = \mathrm{E}\left[Y|X = \mathbf{x}\right] = 2\Pr\left[Y = 1|X = \mathbf{x}\right] - 1$$

  - Target function (or Bayes classifier):

$$t(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x}))$$

Formalization [Bousquet et al., 2004]

- Notations
  - Input space: $\mathcal{X}$, Output Space $\mathcal{Y} = \{-1, 1\}$
  - Function: $g : \mathcal{X} \mapsto \mathcal{Y}$
  - Risk of g:

$$R(g) = \Pr(g(X) \neq Y) = \mathrm{E}\left[\mathbb{1}_{g(X)\neq Y}\right]$$

  - Regression function:

$$\eta(\mathbf{x}) = \mathrm{E}\left[Y|X = \mathbf{x}\right] = 2\Pr\left[Y = 1|X = \mathbf{x}\right] - 1$$

  - Target function (or Bayes classifier):

$$t(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x}))$$

Theorem 1

$t(\cdot)$ is the optimal function, i.e.,

$$R(t) = \inf_g R(g), \quad t \in \operatorname*{argmin}_g R(g)$$

Proof of Theorem 1

$$
\begin{aligned}
R(g) &= \Pr(g(X) \neq Y) \\
&= \mathrm{E}\left[\mathbb{1}_{g(X)\neq Y}\right] \\
&= \mathrm{E}_X\left[\mathrm{E}_Y\left[\mathbb{1}_{g(X)\neq Y}|X\right]\right] \\
&= \mathrm{E}_X\left[\Pr(g(X) \neq Y|X)\right]
\end{aligned}
$$

Proof of Theorem 1

$$
\begin{aligned}
R(g) &= \Pr(g(X) \neq Y) \\
&= \mathrm{E}\left[\mathbb{1}_{g(X)\neq Y}\right] \\
&= \mathrm{E}_X\left[\mathrm{E}_Y\left[\mathbb{1}_{g(X)\neq Y}|X\right]\right] \\
&= \mathrm{E}_X\left[\Pr(g(X) \neq Y|X)\right]
\end{aligned}
$$

Then, $t(\cdot)$ minimizes $R(\cdot)$ if

$$
\begin{aligned}
t(\mathbf{x}) &= \underset{y\in\{-1,1\}}{\operatorname{argmin}} \Pr(y \neq Y|X=\mathbf{x}) \\
&= \begin{cases} 1, & \Pr[Y=1|X=\mathbf{x}] \geq \Pr[Y=-1|X=\mathbf{x}] \\ -1, & \Pr[Y=1|X=\mathbf{x}] < \Pr[Y=-1|X=\mathbf{x}] \end{cases} \\
&= \operatorname{sign}\left(2\Pr[Y=1|X=\mathbf{x}]-1\right)
\end{aligned}
$$

LAMDA Learning And Mining from DatA

Empirical Risk

- The Dilemma

Although

$$t(\mathbf{x}) = \text{sign}(\eta(\mathbf{x})) = \text{sign}\left(2\Pr\left[Y = 1 | X = \mathbf{x}\right] - 1\right)$$

is optimal, the distribution is unknown in general.

- Empirical Risk

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(\mathbf{x}_i) \neq y_i}$$

where $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d. in $\mathcal{X} \times \mathcal{Y}$

Practical Solutions

- Empirical Risk Minimization

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g)$$

- Structural Risk Minimization

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}_d, d \in \mathbb{N}} R_n(g) + \mathrm{pen}(d, n)$$

- Regularization

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g) + \lambda \|g\|^2$$

Bounds

- g_n is a function learnt from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- $g_n \in \mathcal{G}$
- Error Decomposition

$$R(g_n) - \min_g R(g)$$

$$= \underbrace{R(g_n) - \min_{g \in \mathcal{G}} R(g)}_{\text{Estimation Error}} + \underbrace{\min_{g \in \mathcal{G}} R(g) - \min_g R(g)}_{\text{Approximation Error}}$$

- Risk Bound

$$R(g_n) - \min_{g \in \mathcal{G}} R(g)$$

- Generalization Error Bound

$$R(g_n) - R_n(g_n)$$

LAMDA Learning And Mining from DatA

Empirical Process

- Loss class

$$\mathcal{F} = \left\{ f : (\mathbf{x}, y) \mapsto \mathbb{1}_{g(\mathbf{x}) \neq y} | g \in \mathcal{G} \right\}$$

- Define

$$Pf = \mathrm{E}\left[f(X, y) \right] = R(g), \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i, y_i) = R_n(g)$$

- Empirical Process

$$\left\{ Pf - P_n f \right\}_{f \in \mathcal{F}}$$

- Supremum of Empirical Process

$$\sup_{f \in \mathcal{F}} Pf - P_n f, \quad \sup_{f \in \mathcal{F}} \left| Pf - P_n f \right|$$

LAMDA Learning And Mining from DatA

Empirical Process

- Lipschitz Loss class

$$\mathcal{F} = \{f : (\mathbf{x}, y) \mapsto \ell(y, g(\mathbf{x})) | g \in \mathcal{G}\}$$

  - $|\ell(y, g(\mathbf{x})) - \ell(y, g'(\mathbf{x}))| \leq L|g(\mathbf{x}) - g'(\mathbf{x})|$

- Define

$$Pf = \mathrm{E}\left[f(X, y)\right] = R(g), \quad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i, y_i) = R_n(g)$$

- Empirical Process

$$\{Pf - P_n f\}_{f \in \mathcal{F}}$$

- Supremum of Empirical Process

$$\sup_{f \in \mathcal{F}} Pf - P_n f, \quad \sup_{f \in \mathcal{F}} |Pf - P_n f|$$

Empirical Process

- Lipschitz Loss class

$$\mathcal{F} = \{f : (\mathbf{x}, y) \mapsto \ell(y, g(\mathbf{x})) | g \in \mathcal{G}\}$$

  - $\ell(y, g(\mathbf{x})) = \max(0, 1 - yg(\mathbf{x}))$, $\ell(y, g(\mathbf{x})) = \log(1 + e^{-yg(\mathbf{x})})$

- Define

$$Pf = \mathrm{E}[f(X, y)] = R(g), \quad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i, y_i) = R_n(g)$$

- Empirical Process

$$\{Pf - P_n f\}_{f \in \mathcal{F}}$$

- Supremum of Empirical Process

$$\sup_{f \in \mathcal{F}} Pf - P_n f, \quad \sup_{f \in \mathcal{F}} |Pf - P_n f|$$

LAMDA Learning And Mining from DatA

Outline

Problem

- The key problem is to bound

$$Pf - P_n f$$

for some $f \in \mathcal{F}$.

Theorem 2 (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ be independent random variables. Assume that the X_i are almost surely bounded, that is, $\Pr(X_i \in [a_i, b_i]) = 1,\ 1 \leq i \leq n$. Denote

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, we have

$$\Pr\left(\left|\overline{X} - \mathrm{E}[\overline{X}]\right| \geq t\right) \leq 2 \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right)$$

An Elementary Analysis

Theorem 3

Assume $0 \leq f(\mathbf{x}, y) = \ell(y, g(\mathbf{x})) \leq 1$. *With probability at least* $1 - \delta$, *for any* $f \in \mathcal{F}$ *that is independent from* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$|P_n f - P f| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$$

In other words, with probability at least $1 - \delta$, for any $g \in \mathcal{G}$ that is independent from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$|R(g) - R_n(g)| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$$

LAMDA
Learning And Mining from DatA

An Elementary Analysis

Theorem 3

Assume $0 \leq f(\mathbf{x}, y) = \ell(y, g(\mathbf{x})) \leq 1$. *With probability at least* $1 - \delta$, *for any* $f \in \mathcal{F}$ *that is independent from* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$|P_n f - P f| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$$

In other words, with probability at least $1 - \delta$, for any $g \in \mathcal{G}$ that is independent from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$|R(g) - R_n(g)| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$$

- The above result cannot be used for empirical risk minimizer

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g)$$

Remedy

Theorem 4

Assume $0 \leq f(\mathbf{x}, y) = \ell(y, g(\mathbf{x})) \leq 1$. *With probability at least* $1 - \delta$, *for any/all* $f \in \mathcal{F}$

$$|P_n f - Pf| \leq \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{F}|}{\delta}}$$

With probability at least $1 - \delta$,

$$|R(g_n) - R_n(g_n)| \leq \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{F}|}{\delta}}$$

- If $|\mathcal{F}|$ is small, analysis is easy.

Boole's inequality or Union Bound

Formally, for a countable set of events $A_1, A_2, \ldots$, we have

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

An Example

With a probability at least $1 - \delta$,

$$X \leq f(\delta)$$

With a probability at least $1 - \delta$,

$$X \geq -f(\delta)$$

Boole's inequality or Union Bound

Formally, for a countable set of events $A_1, A_2, \ldots$, we have

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

An Example

With a probability at least $1 - \delta$,

$$X \leq f(\delta)$$

With a probability at least $1 - \delta$,

$$X \geq -f(\delta)$$

By the union bound, with a probability at least $1 - 2\delta$, we have

$$|X| \leq f(\delta)$$

By the union bound, with a probability at least $1 - \delta$, we have

$$|X| \leq f(\delta/2)$$

A More Advanced Analysis (I)

- Generalization Error Bound

$$R(g_n) - R_n(g_n) = Pf_n - P_n f_n \leq \sup_{f \in \mathcal{F}} (Pf - P_n f)$$

where $f_n : (\mathbf{x}, y) \mapsto \ell(y, g_n(\mathbf{x}))$

Theorem 5 (McDiarmid's inequality)

Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \mapsto \mathbb{R}$ satisfies

$$\sup_{x_1, \ldots, x_n, x_i' \in A} \left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i$$

for every $1 \leq i \leq n$. Then, for every $t > 0$,

$$P\left\{ f(X_1, \ldots, X_n) - \mathrm{E}\left[f(X_1, \ldots, X_n) \right] \geq t \right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

LAMDA Learning And Mining from DatA

A More Advanced Analysis (II)

- We need to upper bound

$$\sup_{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n),(\mathbf{x}_i',y_i')} \left| \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \right|$$

A More Advanced Analysis (II)

- We need to upper bound

$$\sup_{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n),(\mathbf{x}_i',y_i')} \left| \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \right|$$

For any $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_i', y_i')$,

$$\sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \leq \frac{1}{n} \sup_{f\in\mathcal{F}} \left(f(\mathbf{x}_i, y_i) - f(\mathbf{x}_i', y_i')\right) \leq \frac{1}{n}$$

$$\sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) \leq \frac{1}{n} \sup_{f\in\mathcal{F}} \left(f(\mathbf{x}_i', y_i') - f(\mathbf{x}_i, y_i)\right) \leq \frac{1}{n}$$

A More Advanced Analysis (II)

- We need to upper bound

$$\sup_{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n),(\mathbf{x}_i',y_i')} \left| \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \right|$$

For any $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_i', y_i')$,

$$\sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \leq \frac{1}{n} \sup_{f\in\mathcal{F}} \left(f(\mathbf{x}_i, y_i) - f(\mathbf{x}_i', y_i')\right) \leq \frac{1}{n}$$

$$\sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) \leq \frac{1}{n} \sup_{f\in\mathcal{F}} \left(f(\mathbf{x}_i', y_i') - f(\mathbf{x}_i, y_i)\right) \leq \frac{1}{n}$$

Thus

$$\left| \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \right| \leq \frac{1}{n}$$

which implies

$$\sup_{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n),(\mathbf{x}_i',y_i')} \left| \sup_{f\in\mathcal{F}} \left(Pf - P_n f\right) - \sup_{f\in\mathcal{F}} \left(Pf - P_n' f\right) \right| \leq \frac{1}{n}$$

LAMDA Learning And Mining from DatA

Symmetrization Inequality

Theorem 6

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf-P_nf)\leq \mathrm{E}\left[\sup_{f\in\mathcal{F}}(Pf-P_nf)\right]+\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- We proceed to upper bound $\mathrm{E}\left[\sup_{f\in\mathcal{F}}(Pf-P_nf)\right]$

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}(Pf-P_nf)\right]=\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left(\mathrm{E}\left[P_n'f\right]-P_nf\right)\right]\leq \mathrm{E}\left[\sup_{f\in\mathcal{F}}\left(P_n'f-P_nf\right)\right]$$

$$=\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i',y_i')-f(\mathbf{x}_i,y_i)\right)\right]=\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(f(\mathbf{x}_i',y_i')-f(\mathbf{x}_i,y_i)\right)\right]$$

$$\leq \mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i',y_i')\right]+\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i f(\mathbf{x}_i,y_i)\right]=2\underbrace{\mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i,y_i)\right]}_{:=\mathcal{R}_n(\mathcal{F})}$$

Comparison Inequality [Ledoux and Talagrand, 1991]

Theorem 7

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- How to upper bound $\mathcal{R}_n(\mathcal{F})$

Comparison Inequality [Ledoux and Talagrand, 1991]

Theorem 7

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- How to upper bound $\mathcal{R}_n(\mathcal{F})$

Theorem 8 (Theorem 7 of [Meir and Zhang, 2003])

Let $\{\phi_i\}_{i=1}^n$ *be functions with Lipschitz constant* L_i, *then*

$$\mathrm{E}_\sigma\left[sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i\phi_i(g(\mathbf{x}_i))\right] \leq \mathrm{E}_\sigma\left[sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i L_i g(\mathbf{x}_i)\right]$$

LAMDA Learning And Mining from DatA

Comparison Inequality [Ledoux and Talagrand, 1991]

Theorem 7

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- How to upper bound $\mathcal{R}_n(\mathcal{F})$

Theorem 8 (Theorem 7 of [Meir and Zhang, 2003])

Let $\{\phi_i\}_{i=1}^n$ *be functions with Lipschitz constant* L_i, *then*

$$\mathrm{E}_\sigma\left[sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i\phi_i(g(\mathbf{x}_i))\right] \leq \mathrm{E}_\sigma\left[sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i L_i g(\mathbf{x}_i)\right]$$

$$\mathcal{R}_n(\mathcal{F}) = \mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i, y_i)\right] \leq L\mathrm{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i g(\mathbf{x}_i)\right] = L\mathcal{R}_n(\mathcal{G})$$

LAMDA Learning And Mining from DatA

Bounding Rademacher Complexities

Theorem 9

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 2L\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

Bounding Rademacher Complexities

Theorem 9

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \le 2L\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Suppose

$$\mathcal{G} = \left\{\mathbf{w}^\top\mathbf{x} : \mathbb{R}^d \mapsto \mathbb{R} | \|\mathbf{w}\| \le \gamma\right\}$$

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\le\gamma}\sum_{i=1}^n \sigma_i\langle\mathbf{x}_i,\mathbf{w}\rangle\right]$$

$$\le \mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\le\gamma}\left\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\right\|\|\mathbf{w}\|\right]$$

$$\le \gamma\mathrm{E}\left[\left\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\right\|\right] \le \gamma\sqrt{\mathrm{E}\left[\|\mathbf{x}_i\|^2 + \sum_{u\neq v}\sigma_u\sigma_v\mathbf{x}_u^\top\mathbf{x}_v\right]} \le \gamma D\sqrt{n}.$$

LAMDA Learning And Mining from DatA

Bounding Rademacher Complexities

Theorem 10

With a probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \le \frac{2L\gamma D}{\sqrt{n}} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

- Suppose

$$\mathcal{G} = \left\{\mathbf{w}^\top \mathbf{x} : \mathbb{R}^d \mapsto \mathbb{R} \middle| \|\mathbf{w}\| \le \gamma\right\}$$

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\le\gamma} \sum_{i=1}^{n} \sigma_i \langle \mathbf{x}_i, \mathbf{w}\rangle\right]$$

$$\le \mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\le\gamma} \left\|\sum_{i=1}^{n} \sigma_i \mathbf{x}_i\right\| \|\mathbf{w}\|\right]$$

$$\le \gamma \mathrm{E}\left[\left\|\sum_{i=1}^{n} \sigma_i \mathbf{x}_i\right\|\right] \le \gamma \sqrt{\mathrm{E}\left[\|\mathbf{x}_i\|^2 + \sum_{u\neq v} \sigma_u \sigma_v \mathbf{x}_u^\top \mathbf{x}_v\right]} \le \gamma D \sqrt{n}.$$

LAMDA Learning And Mining from DatA

Risk Bound

- Define

$$g_* \in \operatorname*{argmin}_{g \in \mathcal{G}} R(g)$$

- Risk Bound

$$\begin{aligned}
& R(g_n) - \min_{g \in \mathcal{G}} R(g) \\
=& R(g_n) - R(g_*) \\
\leq& R(g_n) - R(g_*) + R_n(g_*) - R_n(g_n) \\
=& R(g_n) - R_n(g_n) + R_n(g_*) - R(g_*) \\
=& \underbrace{\sup_{f \in \mathcal{F}} (Pf - P_n f)}_{\text{Theorem 10}} + \underbrace{R_n(g_*) - R(g_*)}_{\text{Theorem 3}}
\end{aligned}$$

Risk Bound

- Define

$$g_* \in \operatorname*{argmin}_{g \in \mathcal{G}} R(g)$$

- Risk Bound

$$\begin{aligned}
& R(g_n) - \min_{g \in \mathcal{G}} R(g) \\
=& R(g_n) - R(g_*) \\
\leq& R(g_n) - R(g_*) + R_n(g_*) - R_n(g_n) \\
=& R(g_n) - R_n(g_n) + R_n(g_*) - R(g_*) \\
=& \underbrace{\sup_{f \in \mathcal{F}} (Pf - P_n f)}_{\text{Theorem 10}} + \underbrace{R_n(g_*) - R(g_*)}_{\text{Theorem 3}} \\
\leq& 2 \sup_{g \in \mathcal{G}} |R(g) - R_n(g)| = \underbrace{2 \sup_{f \in \mathcal{F}} |Pf - P_n f|}_{\text{Theorem 10}}
\end{aligned}$$

LAMDA Learning And Mining from DatA

Outline

- Faster Rates
  - Local Rademacher Complexities [Bartlett et al., 2005]
  - Smoothness [Srebro et al., 2010]
  - Strong Convexity [Sridharan et al., 2009]
- Recover Error Bound
  - Suppose the observation $(\mathbf{x}, y)$ is generated according to some model

$$\Pr[Y = y | X = \mathbf{x}] = \frac{1}{1 + e^{-y\mathbf{w}_*^\top \mathbf{x}}}$$

  - Upper bound the difference between $\mathbf{w}_n$ an $\mathbf{w}_*$
  - Compressive sensing, matrix completion
- Active Learning
  - The learner can select the training data actively
  - Establishing theory is extremely difficult

Reference I


Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005).
Local rademacher complexities.
The Annals of Statistics, 33(4):1497–1537.


Bousquet, O., Boucheron, S., and Lugosi, G. (2004).
Introduction to statistical learning theory.
In *Advanced Lectures on Machine Learning*, pages 169–207.


Ledoux, M. and Talagrand, M. (1991).
Probability in Banach Spaces: Isoperimetry and Processes.
Springer.


Meir, R. and Zhang, T. (2003).
Generalization error bounds for bayesian mixture algorithms.
Journal of Machine Learning Research, 4:839–860.


Srebro, N., Sridharan, K., and Tewari, A. (2010).
Optimistic rates for learning with a smooth loss.
ArXiv e-prints, arXiv:1009.3896.


Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009).
Fast rates for regularized objectives.
In *Advances in Neural Information Processing Systems 21*, pages 1545–1552

LAMDA
Learning And Mining from DatA

Reference II


Vapnik, V. N. (1998).
Statistical Learning Theory.
Wiley-Interscience.