統計的学習理論

目次

目次

機械学習研究における「再現性の危機」

towards data science DATA SCIENCE MACHINE LEARNING PROGRAMMING VISUALIZATION AI PICKS MORE | CONTRIBU

The Machine Learning Crisis in Scientific Research

Is an experiment still scientific if it is not reproducible?

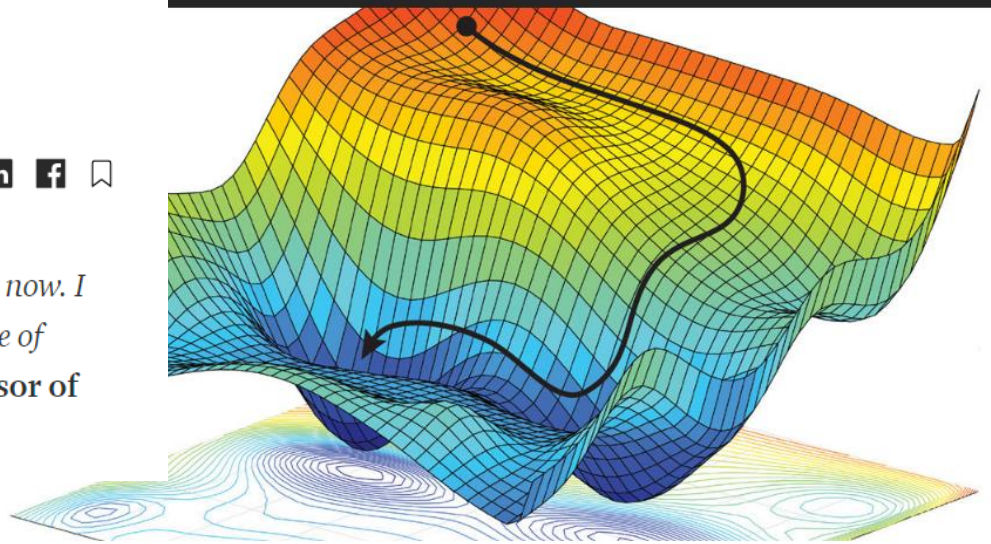
Matthew Stewart, PhD Researcher Follow

Nov 19, 2019 · 10 min read ★

> *"There is general recognition of a reproducibility crisis in science right now. I would venture to argue that a huge part of that does come from the use of machine learning techniques in science."* — **Genevera Allen, Professor of Statistics and Electrical Engineering at Rice University**

https://towardsdatascience.com/the-machine-learning-crisis-in-scientific-research-91e61691ae76

News Careers Journals



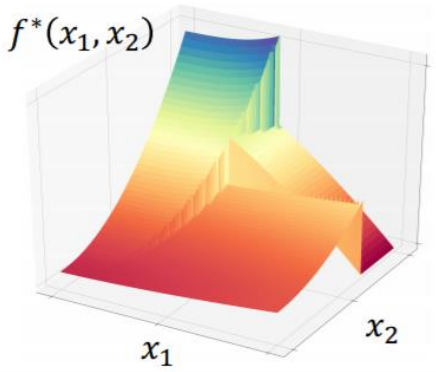
Gradient descent relies on trial and error to optimize an algorithm, aiming for minima in a 3D landscape.
ALEXANDER AMINI, DANIELA RUS. MASSACHUSETTS INSTITUTE OF TECHNOLOGY, ADAPTED BY M. ATAROD/*SCIENCE*

AI researchers allege that machine learning is alchemy

https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy

深層学習の理論研究

- f^*が滑らかでない場合、DNNが他に優越



区分上でのみ滑らかな関数

$f^* = \Sigma_m f_m \otimes 1_{R_m}$

f_m: 滑らかな関数, 1_{R_m}: 区分上の指示関数

近似レートの差別化

(Imaizumi & Fukumizu (2019))

DNNのレート：

$O(\max\{W^{-\beta/d}, W^{-\alpha/2(d-1)}\})$

他手法(カーネル等)のレート：

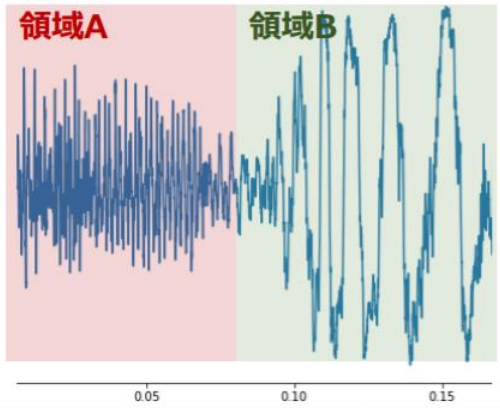
$O(\max\{W^{-\beta/d}, W^{-\alpha/4(d-1)}\})$

(αは区分の境界線の滑らかさ)

区分の境界が複雑な形

➡ DNNが速いレートを達成

- f^*が不均一な滑らかさを持つ場合、DNNが優越



Besov空間の関数

$f^* = \Sigma_j c_j \phi_j + \Sigma_{j,k} c_{j,k} \psi_{j,k}$

$\|c_\cdot\|_p + \left(\Sigma_k 2^{qk(\beta+1/2-1/p)} \|c_{\cdot,k}\|_p^q\right)^{1/q} < \infty$

近似レートの差別化

(Suzuki (2019))

DNNのレート：

$O(W^{-\beta/d})$

他手法(カーネル等)のレート：

$O(W^{-(\beta-(1/p-1/2)_+)/d})$

(pは不均一さの程度)

不均一さがより強い

➡ DNNが速いレートを達成

13

今泉允聡 IBIS2019企画セッション「深層学習の汎化誤差のための近似性能と複雑性解析」
http://ibisml.org/ibis2019/files/2019/11/slide_imaizumi.pdf

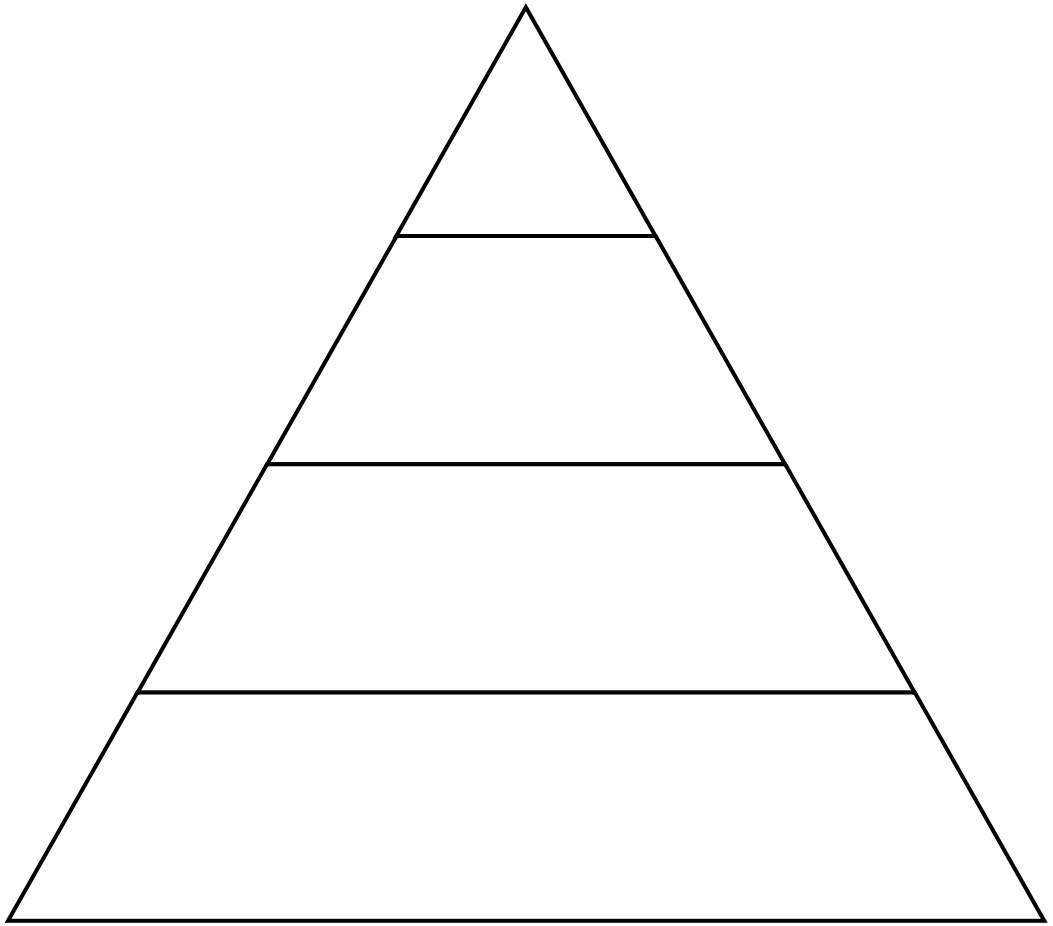
学習理論のモチベーション

- **手法の意味**を理解し正しい使い方ができるようになる
- **手法の正当性**を保証できる（本当に解が求まるか）
- **手法の最適性**をある尺度において保証できる
- **手法の性能向上**や**新手法開発**の指針となる

学習理論のモチベーション

- **手法の意味**を理解し正しい使い方ができるようになる
- **手法の正当性**を保証できる（本当に解が求まるか）
  - →統計的一致性
- **手法の最適性**をある尺度において保証できる
  - →Minimax最適性・許容性
- **手法の性能向上**や**新手法開発**の指針となる

必要な前提知識



今回前提とする部分

統計学

確率論

解析学

(解析学の基礎・ルベーグ積分・関数解析)

数学基礎

(集合論・位相論)

基礎事項の確認

確率論

- 確率変数列の収束
  - 分布収束：　$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$
  - 確率収束：　$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0$
  - 概収束　：　$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1$

  ※概収束⇒確率収束⇒分布収束

  - 確率変数列$\{Z_n\}_{n\in\mathbb{N}}$の確率オーダーが$O_p(r_n)$とは：

$$\lim_{z\to\infty} \lim_{n\to\infty} \sup \Pr\left(\frac{|Z_n|}{r_n} > z\right) = 0$$

基礎事項の確認

確率論

- 大数の法則：$\bar{X}_n \to \mu \quad as \quad n \to \infty$
  - 平均と分散を持つi.i.d.に従う可積分確率変数列の標本平均は平均に、確率収束する（弱法則）/概収束する（強法則）

- 中心極限定理：$\sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \sigma^2) \quad as \quad n \to \infty$
  - 平均と分散を持つi.i.d.に従う可積分確率変数列の標準化された標本平均$(\bar{X} - \mu)/\sigma$に$\sqrt{n}$を乗じたものは標準正規分布に分布収束する

基礎事項の確認

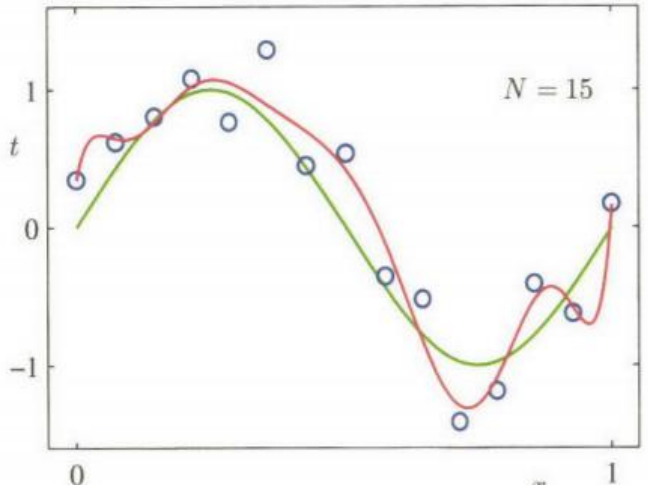
統計学

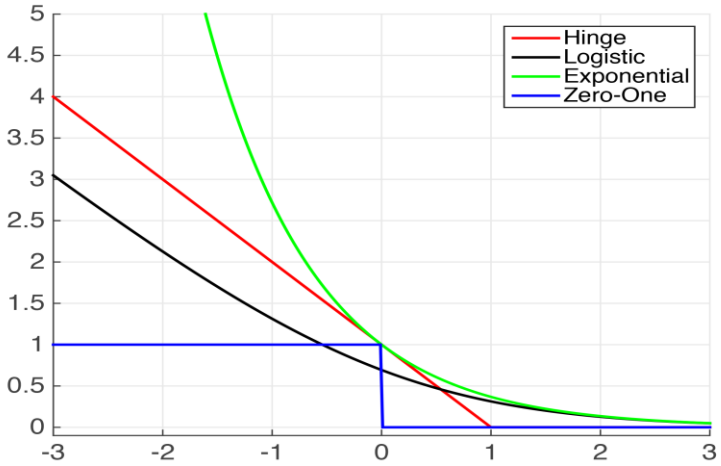
- 推定量の望ましい性質
  - 一致性　：$\lim_{n\to\infty} \Pr(|\theta_n - \theta| < \epsilon) = 1$
  - 不偏性　：$\mathbb{E}[\theta_n] = \theta$
  - 効率性　：不偏性を満たしつつ分散最小
  - 漸近正規性：$n \to \infty$である正規分布に法則収束

目次

用語の定義

- 入力空間　$\mathcal{X}$
- 出力空間　$\mathcal{Y}$　　$X \in \mathcal{X},\ Y \in \mathcal{Y},\ (X,Y) \sim_{i.i.d} D$
- 仮説集合　$\mathcal{H} \ni h : \mathcal{X} \to \mathcal{Y}$
- アルゴリズム　$\mathrm{A} : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$
- サンプル　$S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
- 損失関数　$\ell(h(X), Y)$
  - $\mathbf{1}_{[h(X) \neq Y]}$　0-1 loss (classification)
  - $\left(Y - h(X)\right)^2$　square loss (regression)
  - $\left(1 - Yh(X)\right)_+$　hinge loss
  - $-\log(h(X))$　log loss (density estimation)





問題設定

期待リスク　$R(h) = \mathbb{E}_{(X,Y)\sim D}[\ell(h(X), Y)]$

経験リスク　$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$

これ以降、以下を仮定する

- $\hat{h}$　学習の結果得られた仮説 (e.g. 経験リスク最小化, CV)
- $h^\star = \arg_{\mathrm{h}\in\mathbb{M}} \inf R(h)$　可測関数の中で期待リスクを最小にする仮説
- $h_{\mathcal{H}} = \arg_{\mathrm{h}\in\mathcal{H}} \inf R(h)$　仮説集合の中で期待リスクを最小にする仮説

一般的な事実

- 仮定(classical)
  - $X_1, \ldots, X_n$は$\mathbb{R}$上の確率分布関数Fにおいてi.i.d
  - 経験分布関数 $F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{[X_i \le x]}$
  - 経験過程 $\{Z_n(x) \equiv n^{-1}\big(F_n(x) - F(x)\big) : x \in \mathbb{R}\}$
- **一様大数の法則(Glivenko-Cantelliの定理)**

$$\| F_n - F \|_\infty \equiv \sup_{-\infty < x < \infty} |F_n(x) - F(x)| \to_{a.s.} 0.$$

- **一様中心極限定理(Donskerの定理)**

$$Z_n \rightsquigarrow Z \equiv U(F)$$

ただし$U(F)$は標準Brownian bridge

前頁の図説と汎化ギャップの概収束

目次

目次

数理的に妥当な学習アルゴリズム

- **統計的一致性(statistical consistency)**

任意の分布 D 及び任意の $\varepsilon > 0$ に対して、

$$\lim_{n\to\infty} \Pr_{S\sim D^n}(R(\hat{h}) \leq R(h^\star) + \varepsilon) = 0$$

が成り立つとき、学習アルゴリズム $S \mapsto \hat{h}$ は統計的一致性を持つという。

※ D に制約を課す場合もある(e.g. SVMでは$\mathcal{X}$にコンパクト性を仮定)

目次

代表的な確率集中不等式

- **Markovの不等式**
  - 任意の$\mathbb{R}$上可積分な確率変数X, 正の実数ϵに対して、$\mathbb{E}[|X|]$が存在するならば、

$$\Pr(|X| \geq \epsilon) \leq \frac{1}{\epsilon}\mathbb{E}[|X|]$$

- **Chebyshevの不等式**

$$\Pr(|X - \mu| \geq \sigma\epsilon) \leq \frac{1}{\epsilon^2}$$

- **Hoeffdingの不等式**
  - $X_i \in [0,1]$である$\mathbb{R}$上独立な確率変数$X_1, \ldots, X_n$に対して、

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

- **Bernsteinの不等式**
  - 独立な確率変数列が$\mathbb{E}[X_i] = 0, |X_i| \leq \varsigma$を満たすとき、

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2\varsigma\epsilon}{3}}\right)$$

性能評価

- 学習した仮説の性能を評価するとき、以下をバウンドする

$$R(\hat{h}) - R(h^{\star})$$

- 1つの方法は次の評価。$\exists \delta \in (0,1), \forall \epsilon > 0,$

$$\Pr_{s \sim D^n}\left(R(\hat{h}) - R(h^{\star}) < \epsilon\right) > 1 - \delta$$

※$\Pr(\cdot \geq \varepsilon) \leq \delta(\varepsilon)$の形で表される不等式を**集中不等式**と呼ぶ

性能評価(Bias-Variance分解)

2値分類の例を考える。経験誤差最小化(ERM)で学習を行うとき、

$$R(\hat{h}) - R(h^\star)$$
$$= R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\hat{h}) - R(h_{\mathcal{H}}) + R(h_{\mathcal{H}}) - R(h^\star)$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h_{\mathcal{H}}) - R(h_{\mathcal{H}}) + R(h_{\mathcal{H}}) - R(h^\star) \ (\because ERM)$$
$$\leq 2\max_{h\in\mathcal{H}}\left|\hat{R}(h) - R(h)\right| + R(h_{\mathcal{H}}) - R(h^\star)$$

Variance　　Bias

Hoeffdingの不等式より、

$$\Pr\left(2\max_{h\in\mathcal{H}}\left|\hat{R}(h) - R(h)\right| \geq \epsilon\right) \leq \sum_{h\in\mathcal{H}} \Pr\left(\left|\hat{R}(h) - R(h)\right| \geq \frac{\epsilon}{2}\right) = 2|\mathcal{H}|e^{-\frac{n\epsilon^2}{2}} \equiv \delta$$

このとき、確率$1-\delta$以上で$2\max_{h\in\mathcal{H}}\left|\hat{R}(h) - R(h)\right| \leq \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}}$ なので、

$$R(\hat{h}) - R(h^\star) \leq R(h_{\mathcal{H}}) - R(h^\star) + \sqrt{\frac{2}{n}\log\frac{2|\mathcal{H}|}{\delta}}$$

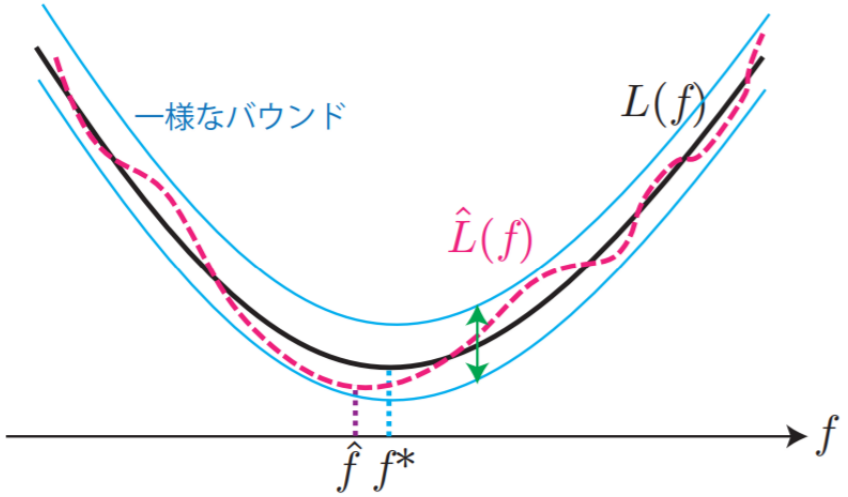
Varianceを押さえられた

性能評価

$$\boldsymbol{R}(\hat{\boldsymbol{h}}) - \boldsymbol{R}(\boldsymbol{h}^{\star})$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^{\star}) - R(h^{\star})$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + O_p\left(\sqrt{\frac{1}{n}}\right) \quad (\because \text{Hoeffding's ineq.})$$

$\hat{h}$に関しては、確率変数列 $\left\{\mathbf{1}_{[\hat{h}(X_i)\neq Y_i]}\right\}_n$ の独立性という条件が満たされないために直接バウンドを与える必要がある

$$R(\hat{h}) - \hat{R}(\hat{h}) \leq \sup_{h\in\mathcal{H}}\left|R(\hat{h}) - \hat{R}(\hat{h})\right| \leq ?$$



無限濃度の仮説集合

- これまでの議論では暗に $|\mathcal{H}| < \infty$を仮定
- しかし仮説集合は連続濃度をもつのがふつう

- 汎化ギャップの一様バウンドを与えることを考える

→直感的に必要そうな変数：データ数・値域・関数の複雑さ

- Rademacher complexity
- VC dimension

→仮説集合の複雑さを示す

Rademacher複雑度

- **経験Rademacher複雑度**
  - $\mathcal{G} \subset \{f : \mathcal{X} \to \mathbb{R}\}$
  - Rademacher変数 $\epsilon_1, \ldots, \epsilon_n$ は$\mathrm{i.i.d}$ で $\Pr(\epsilon_i = 1) = \Pr(\epsilon_i = -1) = \frac{1}{2}$.

$$\hat{\mathfrak{R}}_s(\mathcal{G}) := \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i) \right]$$

- **Rademacher複雑度**

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_{S \sim D} \left[\hat{\mathfrak{R}}_s(\mathcal{G}) \right]$$

- 代表的な性質
  - $\forall c \in \mathbb{R}, \hat{\mathfrak{R}}_s(c\mathcal{G}) = |c| \hat{\mathfrak{R}}_s(\mathcal{G})$
  - $\hat{\mathfrak{R}}_s(\mathcal{G}) = \hat{\mathfrak{R}}_s(conv\, \mathcal{G})$
  - $\hat{\mathfrak{R}}_s(\phi \circ \mathcal{G}) \leq L \hat{\mathfrak{R}}_s(\mathcal{G})$　但しϕはLipschitz連続でLはLipschitz定数 (Talagrand's lemma)

汎化ギャップの一様バウンド

- Rademacher complexityの対称化

$g(X) \in [a, b]$のとき確率$1 - \delta$以上で以下が成立

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(Z)] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right| \leq 2\mathfrak{R}_n(\mathcal{G}) + (b - a) \sqrt{\frac{1}{2\mathrm{n}} \log\left(\frac{1}{\delta}\right)}$$

期待値について以下が成立

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(Z)] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right| \right] \leq 2\mathfrak{R}_n(\mathcal{G})$$

Rademacher複雑度を直接バウンドする

- （線形関数の集合）$\mathcal{G} = \{x \mapsto w^T x \mid w \in \mathbb{R}^d, \|w\| < \Lambda\}$

$$
\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{G}) &= \mathbb{E}_\epsilon \left[\frac{1}{n} \sup_{\|w\| \le \Lambda} w^T \sum_{i=1}^{n} \epsilon_i x_i \right] \\
&= \frac{1}{n} \mathbb{E}_\epsilon \left[\Lambda \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\| \right] \\
&\le \frac{\Lambda}{n} \left(\mathbb{E}_\epsilon \left[\left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|^2 \right] \right)^{\frac{1}{2}} \\
&= \frac{\Lambda}{n} \left(\sum_{i=1}^{n} \|x_i\|^2 \right)^{\frac{1}{2}}
\end{aligned}
$$

Rademacher複雑度を直接バウンドする

- （再生核ヒルベルト空間$(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$）$\mathcal{G} \subset \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{n}\epsilon_i f(x_i)\right]$$

$$= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{n}\epsilon_i\langle f, k(x_i,\cdot)\rangle_{\mathcal{H}}\right] = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{G}}\left\langle f, \sum_{i=1}^{n}\epsilon_i k(x_i,\cdot)\right\rangle_{\mathcal{H}}\right]$$

$$\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^{n}\epsilon_i k(x_i,\cdot)\right\|_{\mathcal{H}}\right] \leq \frac{1}{n}\left\{\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^{n}\epsilon_i k(x_i,\cdot)\right\|_{\mathcal{H}}^2\right]\right\}^{\frac{1}{2}}$$

$$= \frac{1}{n}\left\{\mathbb{E}_\epsilon\left[\left\|\sum_{i,j}\epsilon_i\epsilon_j k(x_i,x_j)\right\|_{\mathcal{H}}\right]\right\}^{\frac{1}{2}} = \frac{1}{n}\left(\sum_{i=1}^{n}k(x_i,x_i)\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{n}}$$

VC次元との関連

- VC次元(Vapnik-Chervonenkis dimension)

$$VCdim(\mathcal{H}) \coloneqq \max\left\{n \in \mathbb{N}\,;\, \max_{x_1,\dots,x_n \in \mathcal{X}} \Pi_{\mathcal{H}}(x_1,\dots,x_n) = 2^n\right\}$$

$$\Pi_{\mathcal{H}}(x_1,\dots,x_n) \coloneqq \left|\left\{\left(h(x_1),\dots,h(x_n)\right) \in \mathcal{Y}^n\,;\, h \in \mathcal{H}\right\}\right|$$

- Sauer's lemma, Massart's lemmaより、

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{z \in \mathrm{A}} \sum_{i=1}^{n} \epsilon_i z_i\right] \leq \sqrt{\frac{2d}{n}\log\frac{en}{d}}$$

$$|A| = \left|\left\{\left(h(x_1),\dots,h(x_n)\right) \in \{\pm 1\}^n\,;\, h \in \mathcal{H}\right\}\right|$$

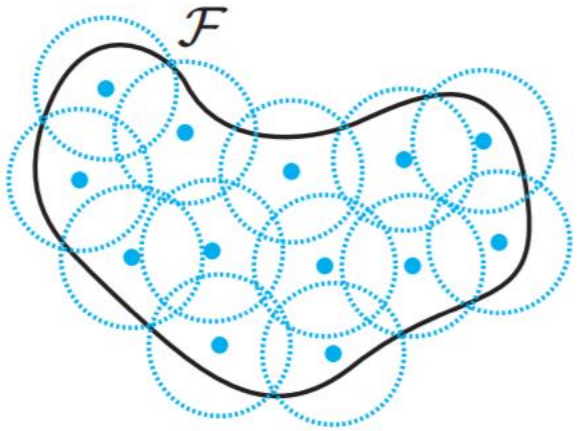
Covering number

- Rademacher complexity（仮説集合の複雑度）をバウンドする

- **Covering number $N(\mathcal{G}, \epsilon, d)$**
  - ノルム d で定まる半径 ϵ の円で $\mathcal{G}$ を覆うために必要な円の数
  - $\log N(\mathcal{G}, \epsilon, d)$ を $\mathcal{G}$ のエントロピーと呼ぶ

- **Dudley積分(chaining)**
  - $\|g\|_n^2 := \frac{1}{n}\sum_{i=1}^n g(x_i)^2$ とすると、

$$\mathfrak{R}_n(\mathcal{G}) \leq \frac{C}{\sqrt{n}} \mathbb{E}_{D^n}\left[\int_0^\infty \sqrt{\log N(\mathcal{G}, \epsilon, \|\cdot\|_n)}\, d\epsilon\right]$$



ERM再訪

$$\boldsymbol{R}(\hat{\boldsymbol{h}}) - \boldsymbol{R}(\boldsymbol{h}^\star)$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^\star) - R(h^\star)$$
$$\leq R(\hat{h}) - \hat{R}(\hat{h}) + O_p\left(\sqrt{\frac{1}{n}}\right) \ (\because \text{Hoeffding's ineq.})$$

$\hat{h}$に関しては、確率変数列 $\mathbf{1}_{[\hat{h}(X_i)\neq Y_i]}$ の独立性という条件が満たされないために直接バウンドを与える必要がある

$$R(\hat{h}) - \hat{R}(\hat{h}) \leq \sup_{h\in\mathcal{H}}\left(R(h) - \hat{R}(h)\right)$$
$$\leq \mathfrak{R}_n\left(\ell(\mathcal{H})\right) + \sqrt{\frac{\epsilon}{n}} \ (\text{with prob. } 1 - e^{-\epsilon})$$
$$\leq \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\epsilon}{t}} \ (\text{prop. of Rademacher comp.: 1-Lipschitz } \ell)$$
$$\leq \frac{1}{\sqrt{n}}\mathbb{E}_{D^n}\left[\int_0^\infty \sqrt{\log N(\mathcal{G}, \epsilon, \|\cdot\|_n)}\, d\epsilon\right] + \sqrt{\frac{\epsilon}{t}} \ (\text{chaining}).$$

仮説集合が単純なほどタイトなバウンド

目次

最適性の規準

- $\bar{R}_\theta(\check{h}) \coloneqq \mathbb{E}_{D^n \sim \hat{D}_\theta}\left[\mathbb{E}_{(X,Y)\sim\hat{D}_\theta}[\ell(Y, \check{h}(X))]\right]$

- **許容性(admissibility)**　常に性能を改善する方法が他にない

$$\bar{R}_\theta(\check{h}) \leq \bar{R}_\theta(\hat{h})(\forall\theta \in \Theta)\text{かつ}$$
$$\text{ある } \theta' \in \Theta \text{ で } \bar{R}_{\theta'}(\check{h}) \leq \bar{R}_{\theta'}(\hat{h}) \text{ なる } \check{h} \text{ が存在しない}$$

- **Minimax最適性**　一番不得意な場面でのリスクが最小

$$\max_{\theta\in\Theta} \bar{R}_\theta(\hat{h}) = \min_{\check{f}} \max_{\theta\in\Theta} \bar{R}_\theta(\check{h})$$

その他のトピック

- Large sample theory
- High dimensional case
- PAC Bayes
- Fast learning rate
  - Local Rademacher complexity

参考

- 金森敬文『統計的学習理論』（講談社MLPシリーズ）
- 鈴木大慈「統計的学習理論チュートリアル：基礎から応用まで」(ibis 2012)
  - http://ibisml.org/archive/ibis2012/ibis2012-suzuki.pdf
- 鈴木大慈「統計的学習理論概説」(日本応用数理学会論文誌 Vol.23, 2013)
  - https://www.jstage.jst.go.jp/article/jsiamt/23/3/23_KJ00008829130/_pdf/-char/ja
- 鈴木大慈「ノンパラメトリックバウンドについて」
  - http://ibis.t.u-tokyo.ac.jp/suzuki/misc/nonparabound.pdf
- 鈴木大慈「スパース推定における確率集中不等式」(数理解析研究所講究録 第1908巻 2014年 39-48)
  - https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/223168/1/1908-03.pdf
- 早川知志・鈴木大慈「スパースなパラメータ空間における深層ニューラルネットワークのミニマックス最適性および優位性について」
  - https://satoshi-hayakawa.github.io/pdfs/JJSM2019.pdf

参考

- M. Mohri, el. al. “Foundations of Machine Learning 2nd edition” (MIT Press)
- A. van der Vaart, J. A. Wellner “Weak Convergence and Empirical Processes: with applications to statistics”(Springer series in stats.)
- S. Boucheron, et al. “Concentration inequalities: A non-asymptotic theory of independence”
  - https://www.hse.ru/data/2016/11/24/1113029206/Concentration%20inequalities.pdf
- Shawe-Taylor and Rivasplata: Statistical Learning Theory - a Hitchhiker's Guide (NeurIPS 2018)
  - https://www.youtube.com/watch?v=m8PLzDmW-TY
- K. Kakade “Symmetrization and Rademacher Averages” Stat 928: Statistical Learning Theory, lecture 11
  - http://stat.wharton.upenn.edu/~skakade/courses/stat928/lectures/lecture11.pdf
- V. Kuznetsov, et al. “Theory and Algorithms for Forecasting Non-Stationary Time Series” (NeurIPS 2016)
  - https://cs.nyu.edu/~mohri/talks/NIPSTutorial2016.pdf