9 Statistical Learning Theory

얼마나 많은 Data가 있어야 어떤 수준의 예측 정확도를 얻을 수 있는지 수학적으로 분석해보자.

Empirical Risk를 잘 줄이는 모델이 true risk도 잘 줄이나? (항상 참은 아님) → 그럼 언제? or 어떤 조건 있을때?

9.1 Introduction

9.1.1 Statistical Learning Theory & No-Free-Lunch Theorems.

- iid sample $(x_1,y_1),\dots,(x_n,y_n) \in \mathbb{R}^d \times \{0,1\}$ 을 사용해 binary classification을 학습시켜보자.

→ goal: Risk $R(h) = E[\mathbb{1}\{h(x) \neq y\}]$를 최소화하는 classifier h 찾기. ↘ if noise = non-zero, error ≠ 0.

- suppose y is Bernoulli distributed with mean $\mu(x)$ ~~1, 0~~ often ≈ 0.5

⇒ class. error $R(h)$을 최소화하는 classifier는 Bayes. Classf. 그런데 (x,y) 결합 분포 모름 → Bayes 계산 불가.

$h^*(x) = \mathbb{1}\{\mu(x) > 1/2\}$ $\mu(x)$ 모름

solution) $\hat{R}(h)$ Empirical Risk Minimization을 사용해보자.

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(x_i) \neq y_i\}$$

근데 이걸 최소화하는 함수는 단순히 training data 만 외우는 overfitting model 일 수도 있다.

= No-Free-Lunch; 보편적으로 좋은 학습 방법은 X. 그럼 어떻게? 1) generative 2) discriminative

9.1.2 Generative vs. Discriminative

- Generative: $P(X,Y)$ or $P(X|Y)$, $P(Y)$를 직접 모델링 e.g) Gaussian Naives Bayes
- Discriminative: $P(y|x)$ or 결정 함수만 모델링 e.g) Logistic Regression, SVM.
- Risk Analysis: $R(\hat{h}) - R(h^*) = \underbrace{R(\hat{h}) - \inf_h R(h)}_{\text{estimation error}} + \underbrace{\inf_h R(h) - R(h^*)}_{\text{approximation error}}$

estimation error: data가 유한해서 생기는 오차
approximation error: 함수 class H가 충분히 넓지 않아 생기는 오차

* Statistical Learning Theory 에서 Empirical Risk Minimization (ERM)의 일반화의 오류를 분석함. 우리는 $h \in H$ 중에서 training data에 대해 Empirical Risk를 최소화하는 $\hat{h}$을 선택함. 궁극적 목표는 True Risk가 작은 모델을 찾는 것

- $R(h)$: true risk (population risk) : $R(h) = E[\text{loss}(h(x),y)]$ 전체 data 분포에 대한 avg. loss
- $\hat{R}(h)$: empirical risk : $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(h(x),y)$ traing data에 대한 avg. loss
- $\hat{h}$: ERM 에서 선택된 가설 : $\hat{h} = \text{argmin}\, \hat{R}(h)$
- h_H: H에서 true risk를 최소화하는 opt. 가설 : $h_H = \text{argmin}\, R(h)$
- $R(\hat{h})$: 선택된 모델 $\hat{h}$의 true risk
- $\hat{R}(\hat{h})$: 선택된 모델 $\hat{h}$의 empirical risk
- $R(h_H)$: H 내의 최적 가설 h_H의 true risk
- $\hat{R}(h_H)$: H 내의 최적 가설 h_H의 empirical risk

$$R(\hat{h}) - R(h_H) = \underbrace{R(\hat{h}) - \hat{R}(\hat{h})}_{\text{일반화 오류 generation gap}} + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h_H)}_{\text{emp. risk 최소화 항상 } \le 0} + \underbrace{\hat{R}(h_H) - R(h_H)}_{\text{일반화 오류 generation gap}} \le |R(\hat{h}) - \hat{R}(\hat{h})| + |\hat{R}(h_H) - R(h_H)|$$

(∵ $\hat{h}$는 리스크를 최소화하는 함수)

→ $R(\hat{h}) - R(h_H) \le |R(\hat{h}) - \hat{R}(\hat{h})| + |\hat{R}(h_H) - R(h_H)|$

→ $|R(\hat{h}) - R(h_H)| \le \sup_{h \in H} |\hat{R}(h) - R(h)|$ ∵ $\hat{h} \in H$ 이므로 어떤 $h \in H$에 대해서도 generation gap은 supremum 보다 작다.

9.2 Risk Minimization & Generation

- $R(f) = \underbrace{R(f) - R(f_H)}_{\text{estimation error}} + \underbrace{R(f_H) - R(f^*)}_{\text{approximation error}} + R(f^*)$

$f_H = \arg\min_{f \in H} R(f)$, $f^* = \arg\min_f R(f)$

- complex model ⇒ approx error ⇓ , estima. error ⇑
- simple model ⇒ approx error ⇑ , estima. error ⇑

* trade-off 설명

- Class H : 우리가 모델을 고르는 후보군 집합.

e.g) Linear Regression $H = \{h_\theta(x) = \theta^T x\}$,

5차 다항식 $H = \{h(x) = a_0 + a_1 x + \cdots + a_5 x^5\}$

Neural network 더 큰 H.

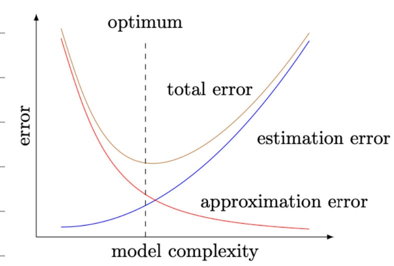
→ 100차 다항식은 1차 함수보다 더 많은 곡선 형태 내포.

= H안에 진짜 최적함수 f^*와 가까운 함수가 들어있을 가능성 높음. ≈ approximation 가까워짐 ⇒ approx. error ⇓

↳ But, data를 외워버릴 위험이 커짐 (overfitting) data가 유한하면 일반화가 안될수도. e.g) 5차식 모델인데 data가 3개뿐이면...

(복잡한 모델은 학습해야하는 parameter 수 ↑, 즉 data 1개당 설명해야하는 정보량이 적어져서 noise 민감해짐)

complexity	Class H	Approx. Error	Estimation Error
complex	↓	↑ (표현력 부족)	↓
simple	↑	↓	↑ (추정이 잘 안되 학습 ↑)



9.3 Uniform Convergence

Empirical Risk를 잘 줄이는 모델이 true risk도 잘 줄이나? (항상 참은 아님) → 증명해주기.

- Definition : < Uniform Convergence >

$$\sup_{h \in H} |\hat{R}(h) - R(h)| \le \varepsilon \text{ (with high probability)}$$

Class H의 모든 가설에 대해 $\hat{R}(h)$와 $R(h)$ 차이가 오이함을 보장.
(empirical risk와 true risk의 차이가 작다는 것을 균일하게 보장)

즉 ERM(Empirical risk Minimization)으로 찾은 $\hat{h}$도 일반화를 잘 할거라는 이론적 보장

↳ Prop 1	finite H , generalization 보장
Theorem 1	Infinite H , covering number를 통해 보장
Prop 2	Lipschitz loss + covering number 상황일때 bound 제공

Prop 1 + Theorem 1 + Prop 2 } uniform convergence ✓ ⇒ ERM 일반화 보장 ⇒ empirical risk 작으면 true risk 작다는 말이 Prop 1, Theorem 1, Prop 2 를 만족할때 성립.

• Prop 1 : Finite Hypothese Class

Suppose H finite, $0 \le loss(f(x),y) \le B$ is bounded,

with probability (at least) $1-\delta$, for all $h \in H$:

$$R(f) \le \hat{R}(f) + B\sqrt{\frac{\log(|H|/\delta)}{2n}}$$

$$P\left[\sup_h |R(h)-\hat{R}(h)| \ge \varepsilon\right] \le 2|H|e^{-2n\varepsilon^2}$$

* finite H에 대해 Höffding's Inequality + union Bound로 bound 가능

$$P\left[\max_i (R(f_i)-\hat{R}(f_i)) \ge t\right] = P\left[\bigcup_{i=1}^{|H|}(R(f_i)-\hat{R}(f_i) \ge t)\right]$$

$$\overset{*}{\le} \sum_i^{|H|} P\left[R(f_i)-\hat{R}(f_i) \ge t\right] \overset{*}{\le} \sum_i^{|H|} e^{-\frac{2nt^2}{B^2}} = |H|e^{-\frac{2nt^2}{B^2}}$$

* Hoeffding's Inequality (Theo 1)

• Theorem 1 : Höffding's Inequality

Let $z_1, \cdots, z_n$ independent random variables taking values in $[a,b]$. Then for $\forall \beta > 0$:

$$P\left[\frac{1}{n}\sum_i^n (z_i - E[z_i]) \ge +\beta\right] \le e^{-\frac{2n\beta^2}{(b-a)^2}}$$

infinite H에서는 union bound X 대신 복잡도 측정 도구 필요.

ε - covering number 개념 도입. 즉 H를 몇개의 함수로 근사적으로 덮을 수 있나? $N(H, \varepsilon, n)$

→ H가 작게 덮이면 (covering number 小) uniform conv. 성립.

• Prop 2 : Lipschitz loss + covering number

suppose loss is Lipschitz, for $\forall z, f, f' \in H$:

$$|loss(f,z) - loss(f',z)| \le L\|f-f'\|$$

Moreover assume loss is bounded $\Leftrightarrow 0 \le loss(f,z) \le B$. Then,

$\sup_f R(f) - \hat{R}(f) \le \varepsilon$ with probability $1 - N(H, \frac{\varepsilon}{4L}) \cdot e^{-2n\varepsilon^2/B^2}$

Also:

$$P\left[\sup |R(f)-\hat{R}(f)| \ge \varepsilon\right] \le N(H, \frac{\varepsilon}{4L}) \cdot e^{-\frac{2n\varepsilon^2}{B^2}}$$

Prof). Let $S = \frac{\varepsilon}{4L}$ for H. $f \in H : \exists f' \in S$ s.t.

$|loss(f,z) - loss(f',z)| \le L\|f-f'\| \le \frac{\varepsilon}{4}$.

Then) $P\left[\exists f \in H : \frac{1}{n}\sum_i^n (E[loss(f,z)] - loss(f,z_i)) \ge \varepsilon\right]$

$\le P\left[\exists f \in S : \frac{1}{n}\sum_i^n (E[loss(f,z)] - loss(f,z_i)) \ge \frac{\varepsilon}{2}\right]$

$\le N(H, \frac{\varepsilon}{4L}) \cdot e^{-\frac{2n\varepsilon^2}{B^2}}$ □

* intuitive)

If $N(H,\varepsilon) = 100$. → $\log N = \log(1000) \approx 6.9$ ⇒ 필요한 data $n \ge \frac{6.9}{\varepsilon^2}$

If $N(H,\varepsilon) = 10^6$ → $\log N \approx 13.8$ ⇒ 동일한 generation bound를 얻으려면 2배 data 필요