Basic Concept of Ch.9

① 왜 어떤 ML model은 잘 generalization 되고, 어떤 모델은 overfitting 될까?

② 어떤 조건에서 empirical risk가 작다면, true risk도 작다고 보장할 수 있나?

#1. 구하고 싶은것: true Risk $R(h)$ But 데이터분포 D모름 → empirical risk $\hat{R}(h)$를 최소화해보자!
근데... $\hat{R}(h)$이 작다고 $R(h)$이 작은 증거 있어?
그럼 $R(h)$. $\hat{R}(h)$ 관계를 알아보자.

#2. Error $R(\hat{h}) - R(f^*) = \underbrace{R(\hat{h}) - R(h_H)}_{\text{estimation error}} + \underbrace{R(h_H) - R(f^*)}_{\text{approximation error}}$ f^*: ideal, h_H: H중 가장 좋은 h, $\hat{h}$: ERM 선택된 h

↳ 분석해보니

#3 Complexity - trade off
H가 단순하면 똑같이 잘 못따라해(approx. error ⇑) but 일반화 잘해 (estm. error ⇓)
H가 복잡하면 똑같이 잘 따라해(approx. error ⇓) but overfitting 되지도 (estm. error ⇑)
그럼... bias-variance tradeoff 가 필요하네?
근데 ERM 잘 하려면 조건이 필요하대!

#4 Uniform Convergence $\sup |R(h) - \hat{R}(h)| \le \varepsilon$ empirical risk 와 true risk의 차이중 가장 큰게 어떤 ε 안에 있으면, empirical risk가 작으면 true risk도 작아.

③ 근데 uniform convergence 가 안되는 경우는?

#5 SGD/SGM 같은 stochastic optimization 알고리즘! 근데 왜?

#6 SGD는 왜 잘 작동? ∵ Algorithmic Stability 때문.
여기서는 어떤 algo 가 입력 데이터를 약간 바꿔도 학습결과가 거의 안바뀌면 "stable"
stable algorithm은 generalization 잘해.
그럼 Stability가 수치적으로 어떻게 되는데?
↳ $E[R(\hat{h}) - \hat{R}(\hat{h})] \le \varepsilon$ 에서 strong convexity, Lipschitz 조건하에

$\varepsilon = \frac{2L^2}{\lambda n}$ 으로 stability 가져.

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 $\mu(x)$ $\cancel{1}, \cancel{0}$ often ≈ 0.5

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

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

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 유한하니 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: H 에서 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

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

(∵ $\hat{h}$는 리스크를 최소화하는 함수)

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

→ $|R(\hat{h}) - R(h_H)| \leq \sup_{h \in H} |\hat{R}(h) - R(h)|$ ∵ $\hat{h} \in H$ 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.

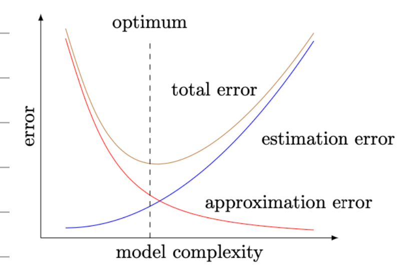
→ 100차 다항식은 1차 함수보다 더 많은 곡선 형태 내포.

= 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 안됨, 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 ε이하임을 보장.
(empirical risk와 true risk의 차이가 작다는 것을 균일하게 보장)

즉 ERM(Empirical risk Minimization)으로 찾은 $\hat{h}$도 일반화를 잘 할거란 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, $0 \le \text{loss}(f(x), y) \le B$ is bounded,

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

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

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

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

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

* Hoeffding's Inequality (Theo 1)

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

- **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 $\forall z, f, f' \in H$:

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

Moreover assume loss is bounded $\Leftrightarrow$ $0 \le \text{loss}(f, z) \le B$. 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:

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

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

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

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

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

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

* intuitive)

If $N(H, \varepsilon) = 1000$. → $\log N = \log(1000) \approx 6.9$ ⇒ 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

9.4 Stochastic Optimization and Statistical Learning Theory

$\hat{R}(\omega) = \frac{1}{n}\sum^{n}$ loss (ω, z_i)를 최소화!

왜 stochastic optimization 이 중요한가? 많은 ML 문제가 확률 최적화 문제로 표현됨.

$$\min_{\omega \in \Theta} R(\omega) := E_{z \sim D}[\text{loss}(\omega, z)]$$

$R(\omega)$: True risk (Population Risk) / D: data 분포 / $z = (x, y)$: data point

하지만 우리는 D를 모르고, 대신 data $z_1, \cdots, z_n$ 만 가짐. 그래서 ERM 사용.

• Stochastic Gradient Method (SGM) 현실에서 큰 data를 다루기위해 stochastic gradient method 사용

$$W_{k+1} = \prod_{\theta}\left(W_k - \alpha_k \nabla \text{loss}(W_k, z_k)\right)$$

- f_w : model parameterized by $w \in \theta$,
- $R(f_w)$: risk $= E[\text{loss}(f_w(x), y)] = R(\omega) = E[\text{loss}(w, y)]$
- z_k : mini batch.
- α_k : 학습률 stepsize
- Π_θ : 제약된 영역 (projection), optional.

* 성능 분석 : SGD는 실제로 얼마나 잘 일반화 할까?

Suppose norm of SG : $E[\|\nabla \text{loss}(w_k, z_k)\|_2^2] \le B^2$, which $B = \sup\|\nabla \text{loss}(w, z)\|_2$

SGD의 iterate avg $\bar{w}_n = \frac{1}{n}\sum w_i$ 에 대해, stepsize를 $\alpha_k = D/B\sqrt{n}$ 로 설정하면:

$$E[R(\bar{w}_n)] - R(w^*) \le \frac{BD}{\sqrt{n}}$$

→ 의미: 평균 iterate risk는 $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ 만큼 w^* 보다 나쁘지 않음.

여기에 확률을 $\ge 1-\delta$ 로 설정하면 :

$$R(\bar{w}_n) - R(w^*) \le \frac{BD}{\sqrt{n}}\left(1 + \sqrt{2\log(1/\delta)}\right)$$

* 결론)

SGD는 널리 쓰이는데 이 이론이 그 일반화 성능을 수학적으로 보장.

→ loss function 의 convexity, Lipschitz, boundedness 가정함.

9.5 Algorithmic Stability

Stability ≈ Robustness

data 하나 바꾼다고 결과가 크게 바뀌지 않는 것. 왜? 일반적으로 stable algorithm은 generalization (일반화) 잘해.

* Flow 9.4) SGM이 왜 generalization 잘 하는지 수학적으로 expectation, prob. bound 로 설명.
9.5) 어떤 알고리즘이 generalization 잘하는가 를 "얼마나 안정적인가?" 라는 관점으로 설명.

• Define : Uniform Stability

Consider learning algorithm A, given $Z = \{z_1, \cdots, z_n\} \in \mathcal{Z}^n$, is ϵ-uniformly stable.

즉, training dataset Z와 그것과 단 한 Sample만 다른 $Z^{(i)}$ 에 대해 :

$$\forall z, \quad E\left[\,\left|\text{loss}(A(Z), z) - \text{loss}(A(Z^{(i)}), z)\right|\,\right] \le \epsilon$$

→ data 하나 바뀌어도 output function은 거의 바뀌지 않음 & 기댓값은 알고리즘의 randomness 에 대해 취함.

• Proposition 3 : Stability ⇒ Generalization

let A ε-uniformly stable, then generation error of A is bounded by $E[R(A(z)) - \hat{R}(A(z))] \le \varepsilon$

⇒ 즉, 일반화 오차의 기댓값이 작다 ≈ overfitting 위험이 작다.

• Proposition 4: Regularized ERM is stable

let A(Z) minimizing regularized empirical loss of model. 이 알고리즘이 다음 최적화 문제를 푼다면, :

$\min_w \frac{1}{n}\sum^n \text{loss}(w, z_i) + \frac{\lambda}{2}\|w\|^2$ ⇝ 이 algorithm은 $\varepsilon = \frac{2L^2}{\lambda n}$.

- loss가 L-Lipschitz 이고,
- λ-strongly convex 라면

* interpretation

- $n \Uparrow \rightarrow$ good stability
- $\lambda \Uparrow \rightarrow$ good stability

* Conclusion

1) Stability는 data 수 n 과 regularization λ 에 의해 제어됨.
2) SGD 나 Regularized ERM 처럼 실제 사용하는 알고리즘의 일반화 성능을 Stability 관점에서 해석할수 있음.
3) uniform convergence 없이 generalization 을 설명