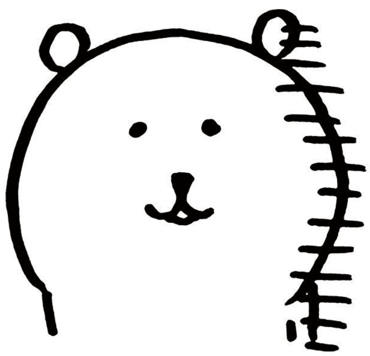
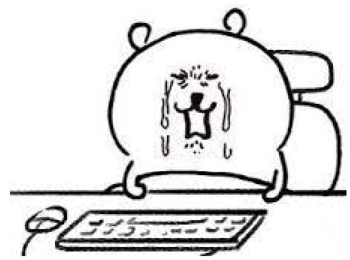
🍎

9강: Approx/Estimation Error & ERM



우선.... 이 강의에 대한 감상을 짧로 남긴다...
난 있잖아... 통계가 세상에서 제일 싫어 하늘 땅만큼



렉쳐노트 보고 교안 쓰는 내 모습
다행히 뒤져보니 렉쳐노트에 있다..!

😇 🔫 목차

Bias and Variance

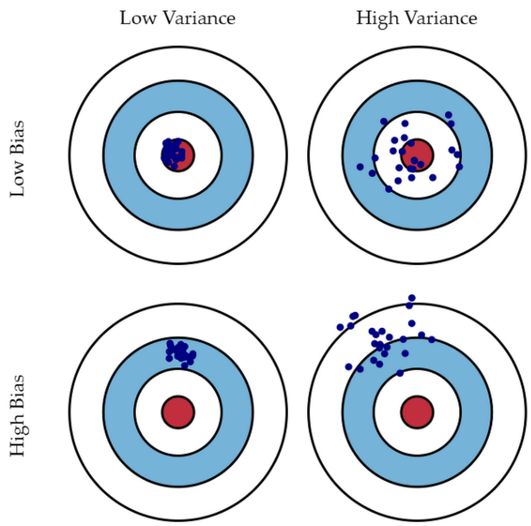


Fig. 1 Graphical illustration of bias and variance.

- **bias**
  - 모델을 통해 얻은 예측값과 실제 정답과의 차이의 평균

$$Bias[f(\hat{x}_0)] = E[f(\hat{x}_0) - f(x)]$$

  - 지나치게 단순한 모델로 인한 error

- **variance**
  - 다양한 데이터 셋에 대하여 예측값이 얼만큼 변화할 수 있는 지에 대한 양(Quantity)의 개념
  - 지나치게 복잡한 모델로 인한 error

$$Var[f(\hat{x})] = E[(f(\hat{x}) - E[f(\hat{x})])^2] = E[f(\hat{x})^2] - E[f(\hat{x})]^2$$

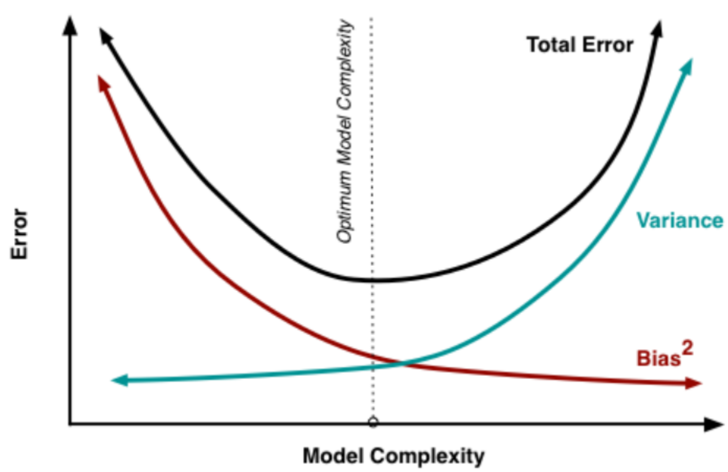
- $m \to \infty \Rightarrow Var[\hat{\theta}] \to 0$
- “Statistical Efficiency” : rate of $Var[\hat{\theta}] \to 0$
- consistent : $\hat{\theta} \to \theta^* (\text{true parameter})$ as $m \to \infty$
- unbiased estimator : $E[\hat{\theta}] = \theta^* \text{ for all m}$

MSE Decomposition

$$
\begin{aligned}
E[y_0 - \hat{f(x_0)}]^2 &= \underbrace{Var(\hat{f(x_0)}) + [Bias(\hat{f(x_0)})]^2}_{reducible\ error} + \underbrace{Var(\epsilon)}_{irreducible\ error} \\
&= E[\{y_0 - E(\hat{f(x_0)})\} - \{\hat{f(x_0)} - E(\hat{f(x_0)})\}]^2 \\
&= \underbrace{E([y_0 - E(\hat{f(x_0)})]^2}_{E(f(x_0)+\epsilon - E(\hat{f(x_0)}))^2} + E[\hat{f(x_0)} - E(\hat{f(x_0)})\}]]^2 - 2E[(y_0 - E(\hat{f(x_0)}))(\hat{f(x_0)} - E(\hat{f(x_0)}))] \\
&= E[f(x_0) - E(\hat{f(x_0)})]^2
\end{aligned}
$$

$$
\begin{aligned}
E[Y - \hat{Y}]^2 &= E[(f+\varepsilon-\hat{f})^2] \\
&= E[(f+\varepsilon-\hat{f}+E[\hat{f}]-E[\hat{f}])^2] \\
&= E[(f-E[\hat{f}])^2] + E[\varepsilon^2] + E[(E[\hat{f}]-\hat{f})^2] \\
&\quad + 2E[(f-E[\hat{f}])\varepsilon] + 2E[\varepsilon(E[\hat{f}]-\hat{f})] + 2E[(E[\hat{f}]-\hat{f})(f-E[\hat{f}])] \\
&= (f-E[\hat{f}])^2 + E(\varepsilon^2) + E[(E(\hat{f})-\hat{f})^2] \\
&= Bias[\hat{f}]^2 + Var(\varepsilon) + Var[\hat{f}] = Bias[\hat{f}]^2 + Var[\hat{f}] + \sigma^2 \\
&(= [f(x) - \hat{f}(x)]^2 + Var(\varepsilon))
\end{aligned}
$$

편향과 분산 간에는 trade-off 관계가 있다는 것을 알 수 있다.



Fighting Variance

1. 데이터 사이즈 키우기 ($M \to \infty$) ~~말이 쉽다~~
2. regularization : l_1,l_2

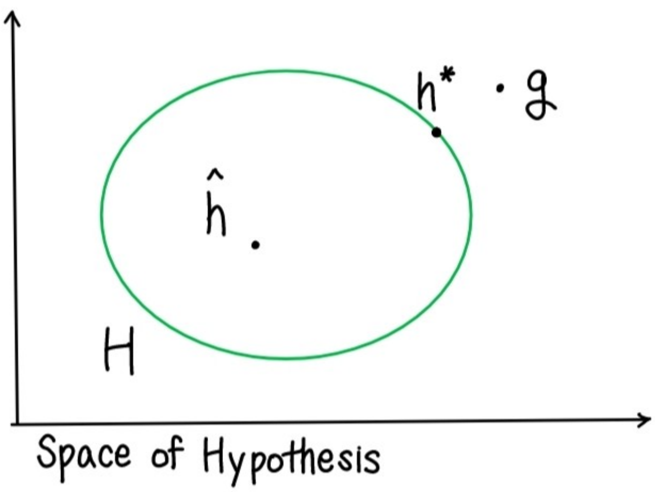
Fight High Bias

1. make your space of Hypothesis $\mathcal{H}$ bigger
2. Empirical Risk Minimization (ERM)

→ 이 두 내용이 어떤 내용인지는 뒷 내용을 통해 알아보자…!

Space of Hypothesis

개념의 이해를 위해 classification 상황을 가정하고 일반화해보자.



Let $\mathcal{H}$ class of Hypothesis

- g : Best possible Hypothesis $\mathcal{H}$
- h^* : Best in class $\mathcal{H}$
- $\hat{h}$: learned from finite data

1. **Empirical Risk** (a.k.a Empirical Error,**training error**)

$$\hat{\varepsilon}(h) = \frac{1}{n}\sum_{i=1}^{n} I\{h(x^{(i)}) \neq y^{(i)}\}$$

- h가 오분류하는 training example의 비율
- training set S에 의존하는 값임을 명시하기 위해서 $\hat{\varepsilon}_s(h)$라 표현하기도 함

2. **Generalization Risk** (a.k.a Generalization Error, **test error**)

$$\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

분포 $\mathcal{D}$에서 새로운 샘플 (x, y)를 끌고 오면 h가 오분류할 확률

3. **Bayes Error** $\varepsilon(g)$: irreducible error

4. **Approximation Error** $\varepsilon(h^*) - \varepsilon(g)$

> *“ What is the price that we are paying for limiting ourselves to some class? ”*

5. **Estimation Error** $\varepsilon(\hat{h}) - \varepsilon(h^*)$

이를 종합한다면 다음이 성립한다는 것을 알 수 있다.

$$\begin{aligned}\varepsilon(\hat{h}) &= \underbrace{\varepsilon(\hat{h}) - \varepsilon(h^*)}_{\text{Estimation Error}} + \underbrace{\varepsilon(h^*) - \varepsilon(g)}_{\text{Approx error}} + \underbrace{\varepsilon(g)}_{\text{irredicible error}} \\ &= Variance + Bias + Irreducible\ error\end{aligned}$$

Assumptions about train data and test data

1. Data Distribution $\mathcal{D}$: $(x, y) \sim \mathcal{D}$ ← train set과 test set은 같은 분포에서 온다.

cf) PAC assumptions : “probably approximately correct”

2. All this samples are **independent** samples

$$\underbrace{\begin{matrix} x^{(1)}\ y^{(1)} \\ \vdots \\ x^{(m)}\ y^{(m)} \end{matrix}}_{sample} \Rightarrow D \Rightarrow \text{Learning Algorithm} \Rightarrow \hat{h}, \hat{\theta}$$

- D : Deterministic Function
- Learning Algorithm : estimator
- $\hat{h}, \hat{\theta}$: hypothesis → random variables

Empirical Risk Minimization

그럼 위의 두가지를 가정할 때, 어떤 파라미터를 선택해야 할까?

하나의 접근법은, **train error**값을 최소화하는 파라미터를 선택하면 된다.

$$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta)$$

이를 hypothesis space $\mathcal{H}$와 연결지어 적어보자면 다음과 같다.

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

우리는 이런 과정을 **Empirical Risk Minimization (ERM)**이라 부르기로 했어요

당연한 소리 지껄이고 있으니 짜증이 난다. 응씨 가만 안둬 ~

Sample complexity bounds

train/test set이 샘플링된 모집단의 분포가 동일하다는 것,
training error를 최소화하는 파라미터를 고르는 방법이 적절하다는 것을 어떻게 보장할 수 있을까?

이걸 bound로 증명한 미친것들이 있다! 알아보도록하자

Preliminaries

1. **The Union Bound**

Let $A_1, A_2, ... A_k$ be k different events (that may not be independent).

$$P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \cdots + P(A_k)$$

cf) In probability theory, it is usually stated as an axiom

2. **Hoeffding inequality(Chernoff bound)**

Let $Z_1, ... Z_n$ be n independent and identically(iid) distributed random variables from Bernoulli(ϕ) distribution ,$\hat{\phi} = (1/n) \sum_{i=1}^{n} Z_i$ (=mean of random variables) , $\gamma > 0$ fixed

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2exp(-2\gamma^2 n)$$

n이 커지면서 추정값과 true value의 값 차이가 클 확률이 점점 줄어든다.

(동전을 던지는 횟수 n이 커질수록 앞면이 나올 확률 ϕ에 가까워지는 것을 생각해보라)

The case of finite $\mathcal{H}$

앞서서와 똑같이 이진분류 상황을 가정해보자

- $\mathcal{H} = \{h_1, \ldots . h_k\}$ ← K개의 hypotheses
- $\mathcal{H}$: 함수 k개의 집합 , $\mathcal{X} \rightarrow \{0, 1\}$
- empirical risk minimization : select $\hat{h}$ (∵ the smallest training error)
- sample $(x, y) \sim \mathcal{D}$
- $Z = I\{h_i(x) \neq y\} \rightarrow Z_j = I\{h_i(x^{(j)}) \neq y^{(j)}\}$

이때, training error는 $\hat{\varepsilon}(h_i) = \frac{1}{n}\sum_{i=1}^{n} Z_j$로 표현가능하다.
(= mean of
n random variables $Z_j \sim_{iid} Bernoulli(\varepsilon(h_i))$

Hoeffding inequality를 적용하면 다음과 같은 부등식을 얻을 수 있다.

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2exp(-2\gamma^2 n)$$

이는 특정 h_i에 대해서 , n이 크다고 가정할 때 , training error는 높은 확률로 test error에 가까워짐을 보여준다.

하지만, 우리는 특정한 하나의 h_i에 대해서만 이걸 증명하고 싶은 것이 아니다. 이 때 union bound를 써보자.

$$\begin{aligned} P(\exists h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\ &\leq \sum_i^k P(A_i) \\ &\leq \sum_i^k 2exp(-2\gamma^2 n) \\ &= 2k\ exp(-2\gamma^2 n) \end{aligned}$$

양변을 1에서 빼면

$$\begin{aligned} P(\neg\exists h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall\exists h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\ &\geq 1 - 2k\ exp(-2\gamma^2 n) \end{aligned}$$

적어도 $1 - 2k\ exp(-2\gamma^2 n)$의 확률로 **모든** $h \in \mathcal{H}$에 대해 $\varepsilon(h)$는 $\hat{\varepsilon}(h)$의 γ 이내에 있을 것이다. (γ is margin of a error)

그럼 n이 얼마나 커야 적어도 $1 - \delta$의 확률로 training error는 test error의 γ 이내에 있을 수 있을까?
(
δ는 probability의 error라고 할 수 있다.)

$\delta = 2k\ exp(-2\gamma^2 n)$라 하고 이를 n에 대하여 풀면,

$$n \geq \frac{1}{2\gamma^2} log \frac{2k}{\delta}$$

적어도 $1 - \delta$의 확률로 모든 $h \in \mathcal{H}$에 대해서 $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$임을 알 수 있다.

이처럼 특정 수준의 수행을 기대하기 위해 필요한 training set의 개수 n을 **algorithm's sample complexity**라고 한다.

$h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h)$ 를 $\mathcal{H}$에서 가능한 최고의 hypothesis라고 하자.

그럼 다음이 성립한다.

$$\begin{aligned}\varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma\end{aligned}$$

Theorem. Let $|\mathcal{H}| = k$, and let any n, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h\in\mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2n}\log\frac{2k}{\delta}}.$$

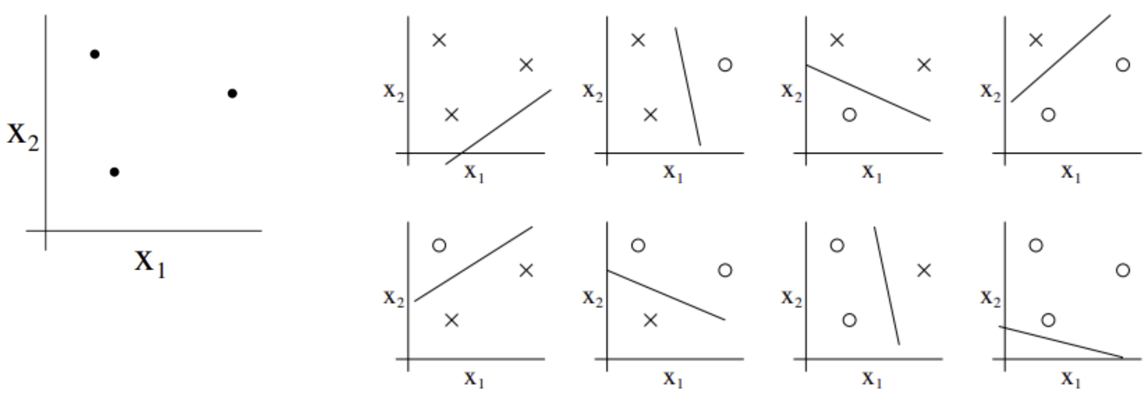
Corollary. Let $|\mathcal{H}| = k$, and let any δ, γ be fixed. Then for $\varepsilon(\hat{h}) \leq \min_{h\in\mathcal{H}} \varepsilon(h) + 2\gamma$ to hold with probability at least $1 - \delta$, it suffices that

$$\begin{aligned}n &\geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right),\end{aligned}$$

The case of infinite $\mathcal{H}$

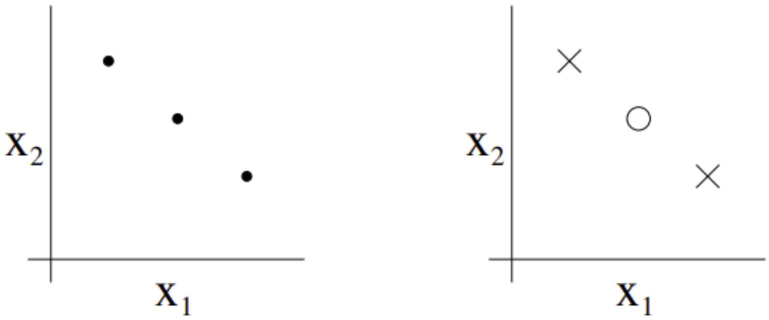
Vapnik-Chervonenkis dimension VC($\mathcal{H}$)

: size of the largest set that is shattered by $\mathcal{H}$ (shatter: 주어진 데이터셋에 라벨링을 하는 정도)



선형적으로 분류한다고 하면, 주어진 데이터를 라벨링할 수 있는 최대의 집합의 개수는 3이므로 **VC($\mathcal{H}$) =3이다.**

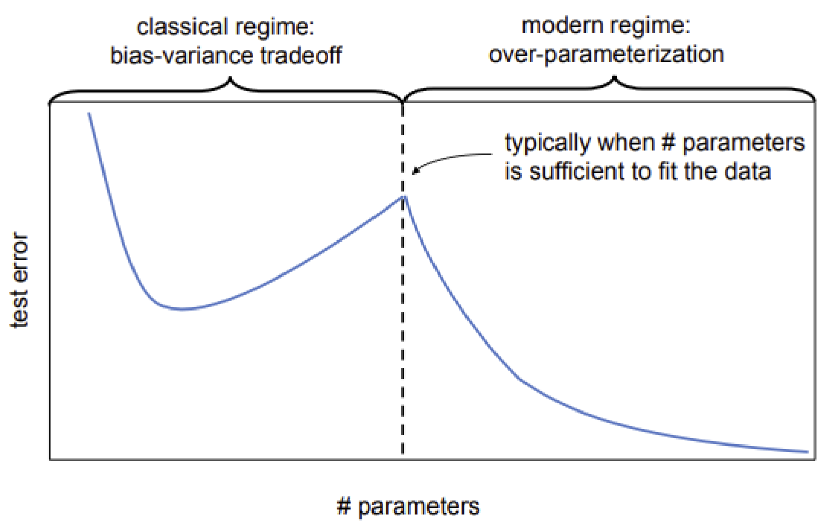
참고) 선형적인 분류가 어려운 경우 (VC($\mathcal{H}$)$\neq 3$ 인 경우)



적어도 $1-\delta$의 확률로, 모든 $h \in \mathcal{H}$에 대해서

- $|\varepsilon(h) - \hat{\varepsilon}(h)| \le O(\sqrt{\frac{D}{n} log \frac{n}{D} + \frac{1}{n} log \frac{1}{\delta}}$
- $\varepsilon(\hat{h}) \le \varepsilon(h^*) + O(\sqrt{\frac{D}{n} log \frac{n}{D} + \frac{1}{n} log \frac{1}{\delta}}$
- $n = O_{\gamma,\delta}(D)$

▼ The double descent phenomenon



Deep Double Descent: Where Bigger Models and More Data Hurt

고전적인 Bias-Variance Trade-Off와, 이와 다른 Double Descent를 포괄할 수 있는 가설 제시 방대한 양의 실험으로 제시한 가설 Effective Model Complexity의 타당성을 보임 이전에 다루어 지지 않은 Epoch-wise Double Descent와 Sample-wise Non-Monotonicity 현상을 발견 구글에 Bias-Variance Trade-Off 를 검색하

https://creamnuts.github.io/short%20review/deep_double_descent/