Statistical Machine Learning

2주차

담당: 18기 신인수

0. Recap: Probability

1.What is Supervised Learning?

2. Train model

3. Model Selection

0. Recap: Probability

1.What is Supervised Learning?

2. Train model

3. Model Selection

0. Recap: Probability

Probability(확률)

확률의 정의

- 표본공간 (sample space; Ω): 일어날 수 있는 모든 경우에 대한 집합
  - Ex: 주사위를 한 번 던졌을 때의 표본공간 → Ω = {1,2,3,4,5,6}

- 사건(event): 표본공간의 부분집합
  - Ex: 집합 A = {주사위를 한 번 던졌을 때 짝수가 나오는 사건} = {2, 4, 6}

- 확률:

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

: 집합의 원소 개수

Probability(확률)

확률의 정의

- 확률 변수: 사건이 가질 수 있는 값을 표현한 변수
  - X = 주사위를 한 번 던져서 나오는 값 → $X = 1, X = 2 \ldots X = 6$
  - 확률변수로 확률을 정의할 수 있다. → ex: $P(X = 1) = \frac{1}{6}$

- 이산형 확률 변수: 확률변수가 가질 수 있는 값이 이산적이다 (자연수, 정수 등)
  - $X \sim Binomial(n, p)$: 이항분포, 주사위 던지기
  - 확률은 합으로 정의된다.
    - Ex: 주사위에서 X가 3 이하로 올 확률 → $P(X = 3) + P(X = 2) + P(X = 1)$
- 연속형 확률 변수: 확률변수가 가질 수 있는 값이 연속적이다 (실수)
  - $X \sim N(\mu, \sigma^2)$: 정규분포
  - 확률은 적분 (면적)으로 정의된다.
    - Ex: 정규분포에서 X가 3 이하일 확률 → $P(X \leq 3) = \int_{-\infty}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$

Probability(확률)

기댓값 $E(X) \approx$ 평균

- **말 그대로 "기대되는 값"**
- 룰렛을 돌렸을 때,
  - -$15일 확률이 40%
  - $0 일 확률이 30%
  - $5일 확률이 20%
  - $10일 확률이 10%

내가 돈을 벌 수 있는 "기댓값"은?

$$E(X) = \sum_{x \in \Omega} xP(x)$$

→ 이산형 확률변수의 기댓값

$$E(X) = -15 \times P(X = -15) + 0 \times P(X = 0) + 5 \times P(X = 5) + 10 \times P(X = 10)$$

Probability(확률)

기댓값 $E(X) \approx$ 평균

$$E(X) = \int_{x\in\Omega} xf(x)dx$$

→ 연속형 확률변수의 기댓값
→ 연속형의 sum = 적분

Probability(확률)

분산 $Var(X)$

- **"편차 제곱의 평균"**
- 편차: 평균과의 차이 → **편차의 총합은 0**

$$Var(X) = E(|X - \mu|^2)$$

Ex: 3, 4, 5, 5, 6, 7, 7, 11
평균: 6 → 편차: -3, -2, -1, -1, 0, 1, 1, 5 → 총합: 0

편차는 별로 의미 없다 → 제곱

편차 제곱: 9, 4, 1, 1, 0, 1, 1, 25 → 분산: 5.25

Probability(확률)

★ ★ ★ 조건부 확률 ★ ★ ★

- A라는 조건 하에서 B의 확률

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- A 교수님이 있다. A 교수님은 50% 확률로 **빨강 옷** 또는 **파랑 옷**을 입는다.
  - A 교수님은 **빨강색 옷**을 입으면, 90% 확률로 기분이 좋고, 10% 확률로 기분이 안 좋다.
  - 반면 A 교수님이 **파랑색 옷**을 입으면 30% 확률로 기분이 좋고, 70% 확률로 기분이 안 좋다.

- 어느 날, 인수가 카페에서 A 교수님을 마주쳤는데, 옷 색깔이 안 보였다. 인수는 A 교수님이 기분이 안 좋을 것이라 예상했다. 인수의 생각이 맞을 확률은? → 40%
- 서연이는 A 교수님이 **빨강색 옷**을 입은 것을 발견했다. 서연이는 A 교수님이 기분이 좋을 것이라 예상했다. 서연이의 생각이 맞을 확률은? → **90%**

Probability(확률)

★ ★ ★ 조건부 확률 ★ ★ ★

- 확률의 독립
- **“조건이 있으나, 없으나 결과가 같다”**
- A 교수님이 입은 옷 색깔에 관계없이 기분이 일정한 경우 → 기분과 옷 색깔은 독립

- 아래 세 수식은 모두 같은 뜻

$$P(B|A) = P(B) \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$$

Probability(확률)

★ ★ ★ 조건부 확률 ★ ★ ★

- **조건이 있을 때 더 많은 정보를 얻을 수 있다!**
- **데이터 분석이란 변수간의 관계를 찾아내는 것!**
- **변수간의 관계란 "조건부 확률"에 해당!**

종속변수 (X) → 독립변수(Y)

X라는 데이터를 가지고 있을 때 (정보를 알 때), Y는 어떻게 변할 것인가?

Ex. 단순 선형회귀:

$$E(Y|X) = \beta_0 + \beta_1 X$$

1. What is Supervised Learning?

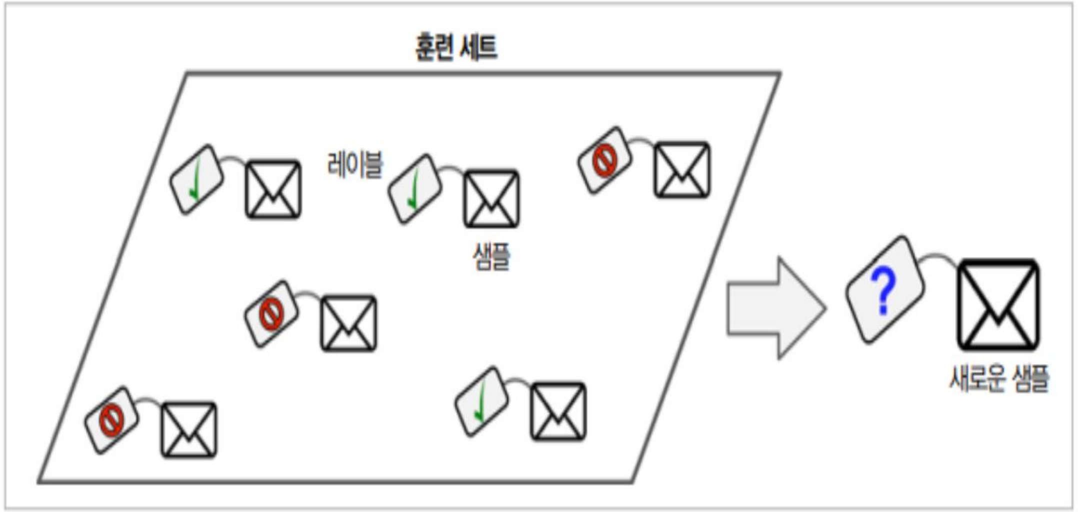
"정답이 있는 문제를 학습"

Model = "Learner"

우리가 뭔가를 알려주는 입장

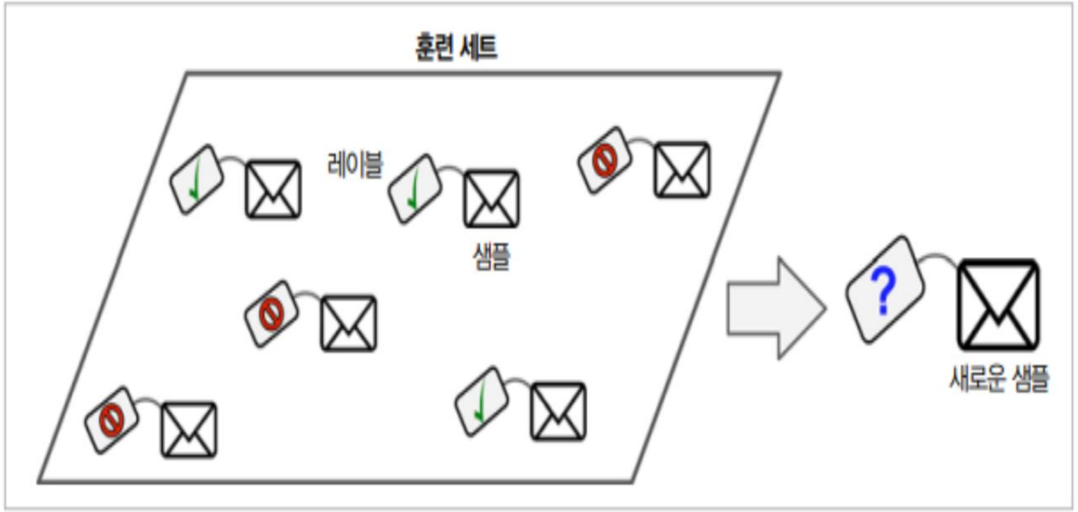
Supervised Learning (지도 학습)

Supervised Learning vs Unsupervised Learning



Supervised Learning (지도 학습)

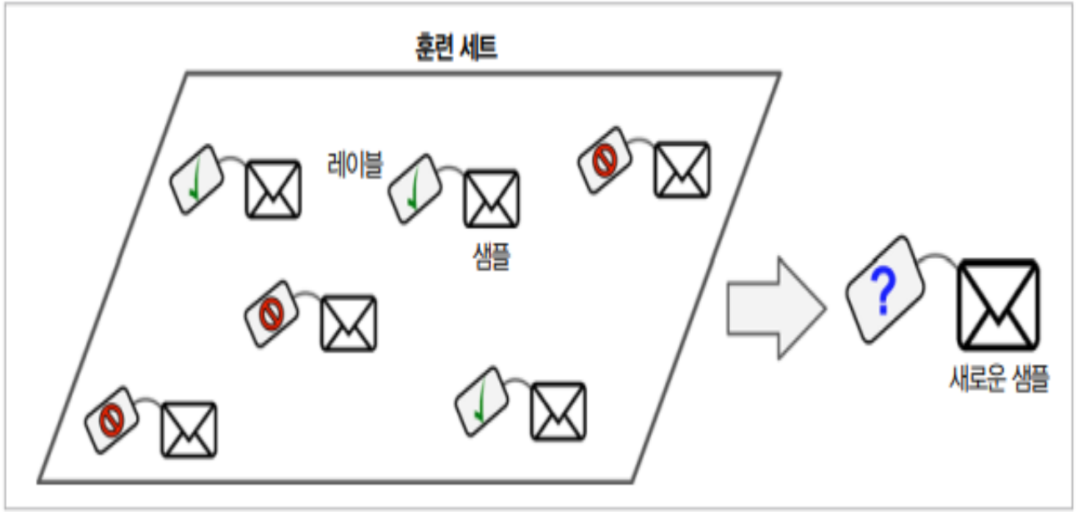
Supervised Learning



Unsupervised Learning

Supervised Learning (지도 학습)

Supervised Learning



- 스팸 메일 구분하기
- 내일 강수량(mm) 예측하기

→"정답"(label)이 있는 문제

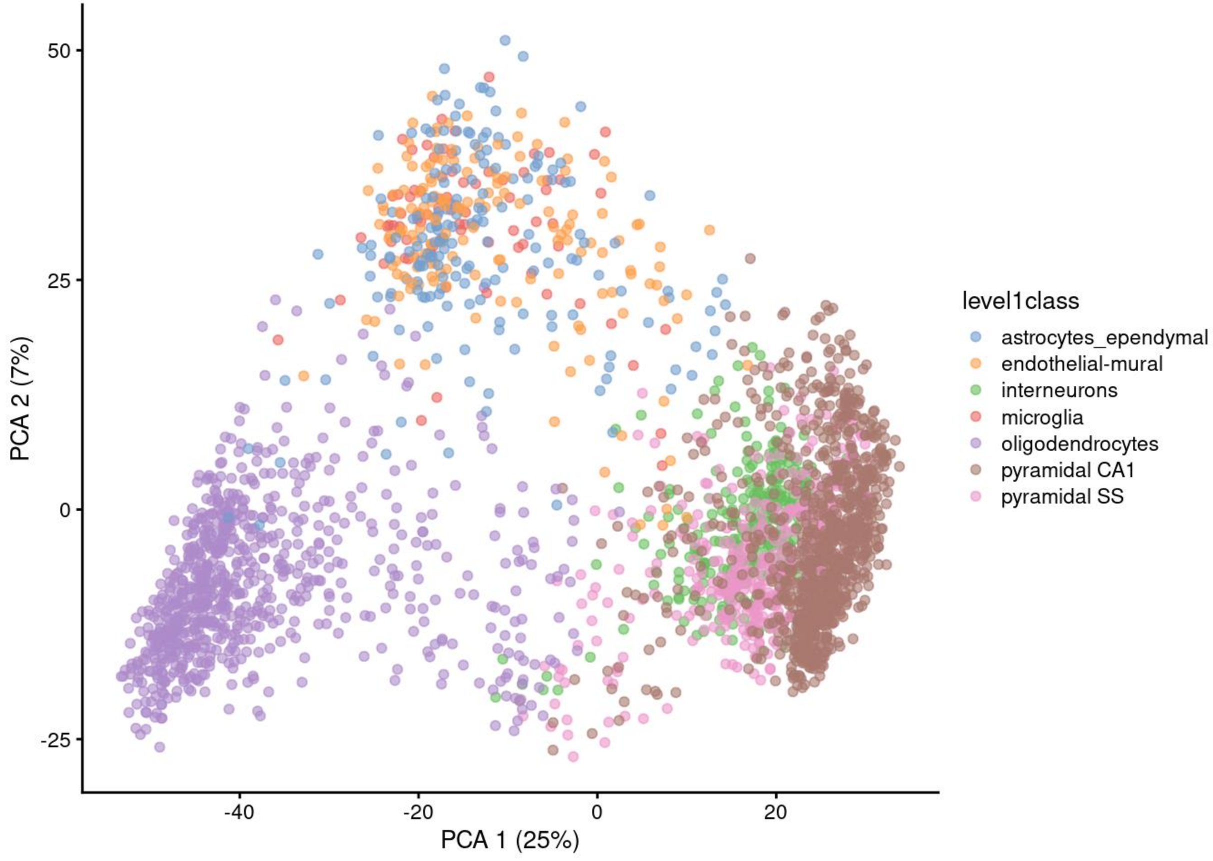
Supervised Learning (지도 학습)

Unsupervised Learning

- 네안데르탈인 유골 분류하기
- 암 종양 내의 세포 분류하기

→"정답"(label)이 없는 문제

Supervised Learning (지도 학습)



Unsupervised Learning

- scRNA-seq 세포 구분하기
- 깔끔하게 분류되진 않음

→ 최대한 분류해 볼 수는 있지만, **끝내 정답을 알 수는 없다.**

Supervised Learning

Objective

Regresion

지도학습: y가 연속형

Classification

지도학습: y가 범주형

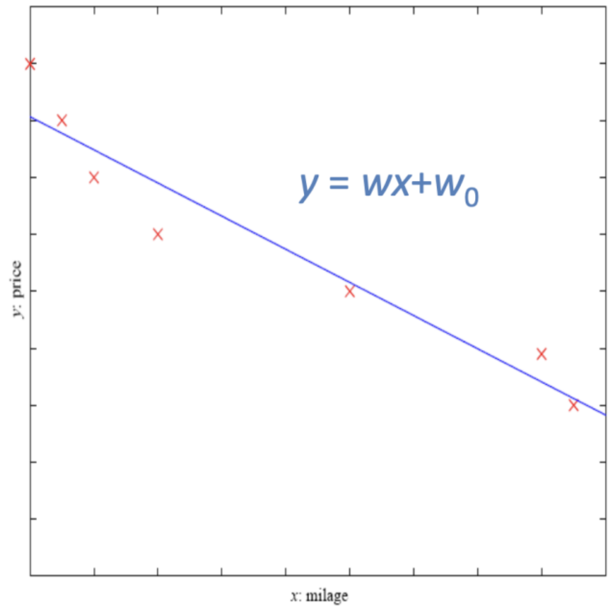
Supervised Learning

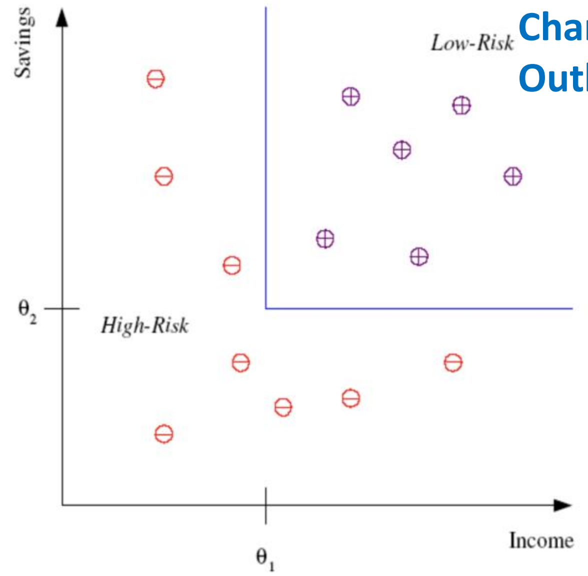
Objective

Regression

Classification

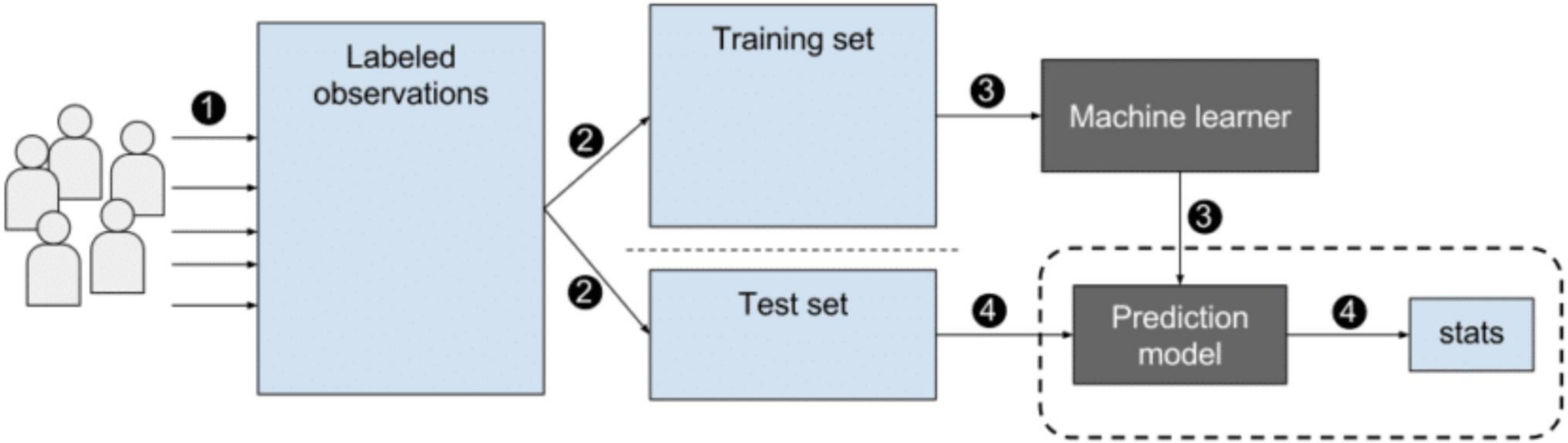
Price of car used
Housing price prediction





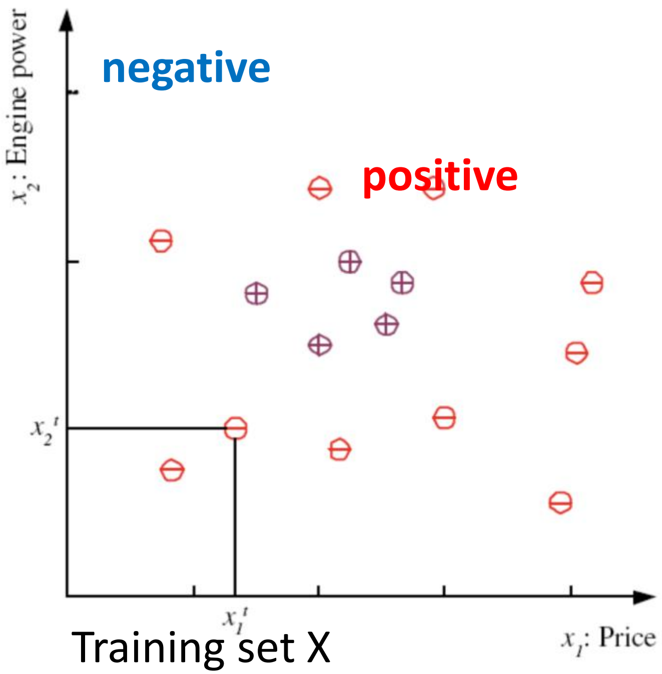
Face Recognition
Character Recognition
Outlier / Novelty Detection

Supervised Learning



2. Train Model

Learning a Class



Training set X

Class C : Family car

→ Is this car x a family car? = classification task

Input representation:

X_1 :price, X_2 : engine power

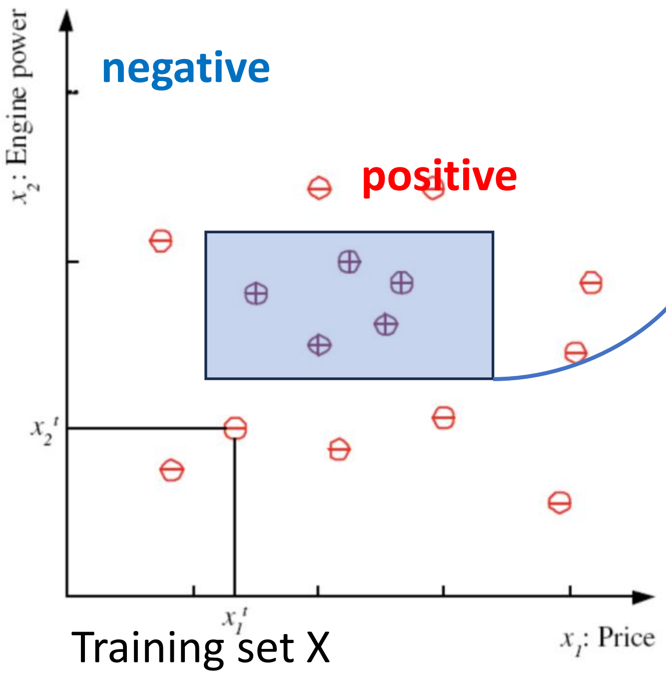
"Learning a Class"

= input feature을 통해서 class를 서술하는 것

Output:

positive(+) or negative(-)

Learning a Class



Training set X

$(p_1 \leq \text{price} \leq p_2) \ \& \ (e_1 \leq \text{engine power} \leq e_2)$

학습의 목표: Class description → example classification

Inductive Bias:
- 학습이 가능하도록 하는 장치→ Aligned Rectangle
- 학습 시에는 만나보지 않았던 상황에 대하여 정확한 예측을 하기 위해 사용하는 추가적인 가정
- Parameter : {p1, p2, e1, e2}

!! 결국 classification을 위해서 {p1, p2, e1, e2} 만 찾으면 된다 !!

수 많은 {p1, p2, e1, e2} 조합 = Hypothesis H = Assumption = Model

Q. 이 중에서 "최적"의 선택은 어떤 것…?

Inductive Bias (중요x)

어떤 머신러닝 모델이건 새로운 데이터를 분석하기 위한 가정들이 존재한다.

예시:

- Linear regression: input과 output 사이에 선형적인 관계를 이룬다.
- KNN (k-nearest neighbors) : 가까운 거리에 있는 input은 output도 가깝다.

Dimensions of a Supervised Learner

모델을 훈련한다= Task를 이행하기 위해서, 훈련 데이터 셋에 **가장 잘 맞도록** 모델 파라미터를 설정

1. Model: $g(\mathbf{x} \mid \theta)$

2. Loss function: $E(\theta \mid \mathcal{X}) = \sum_t L\left(r^t, g\left(\mathbf{x}^t \mid \theta\right)\right)$

3. Optimization procedure: $\theta^* = \arg\min_\theta E(\theta \mid \mathcal{X})$

모델을 설정.

↓

정답을 통한 모델의 성능을 측정

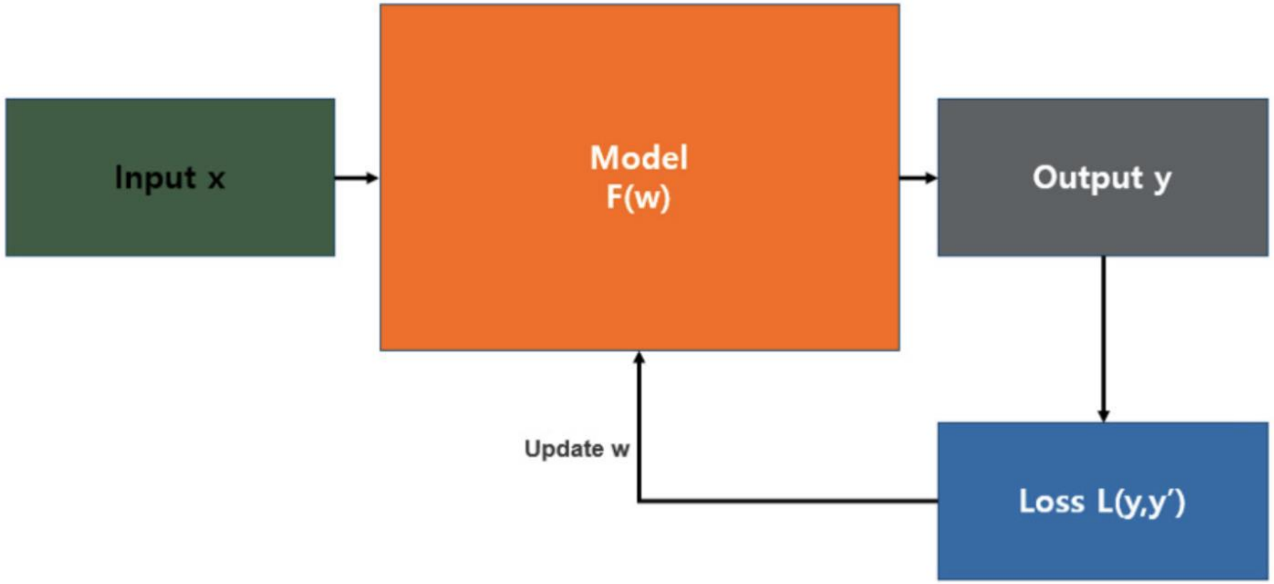
↓

정답을 가장 잘 맞추는 모델 파라미터를 찾는다.

Loss Function

- What is Loss Function?   예측값과 실제값(레이블)의 차이를 구하는 기준

  Quantifies the error between output of the algorithm and given target value.



Loss Function

Loss function penalizes bad predictions.

Regression

- Mean Squared Error

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

Others:
Mean absolute error and mean bias error

Classification

- Binary Cross Entropy

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

- Categorical Cross Entropy

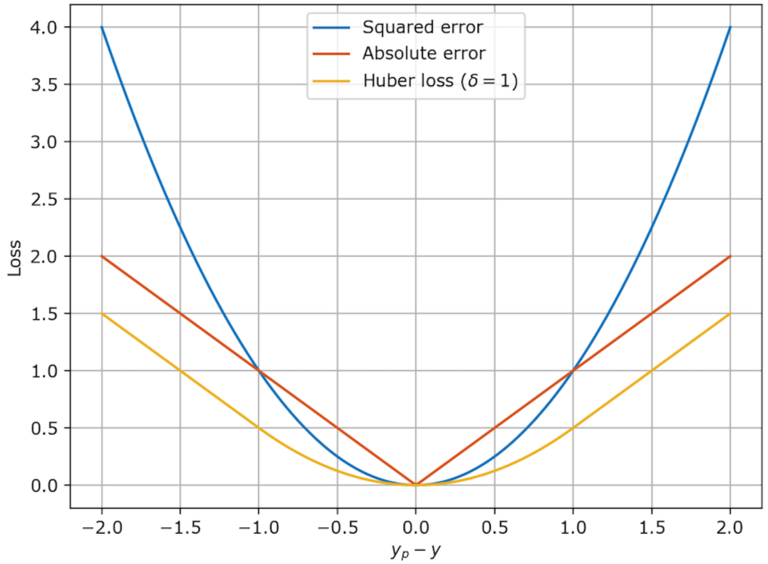
$$CCE = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot \log(\hat{y}_j) + (1 - y_j) \cdot \log(1 - \hat{y}_j)$$

Others:
Hinge loss / SVM loss.

MSE

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$



Idea는 분산과 동일:
- 분산: “편차 제곱의 평균”
- MSE: “잔차(residual) 제곱의 평균”

→“거리의 평균을 수치화”
→ 마음에 안 들면 거리의 개념을 바꾸면 그만
→ MAE, Huber loss, quantile loss 등

제곱을 거리로 삼을 것이냐, 절댓값을 거리로 삼을 것이냐의 차이

MSE: 거리가 크면 penalty가 더 세다
MAE: 거리가 커도 penalty가 덜하다.

Entropy

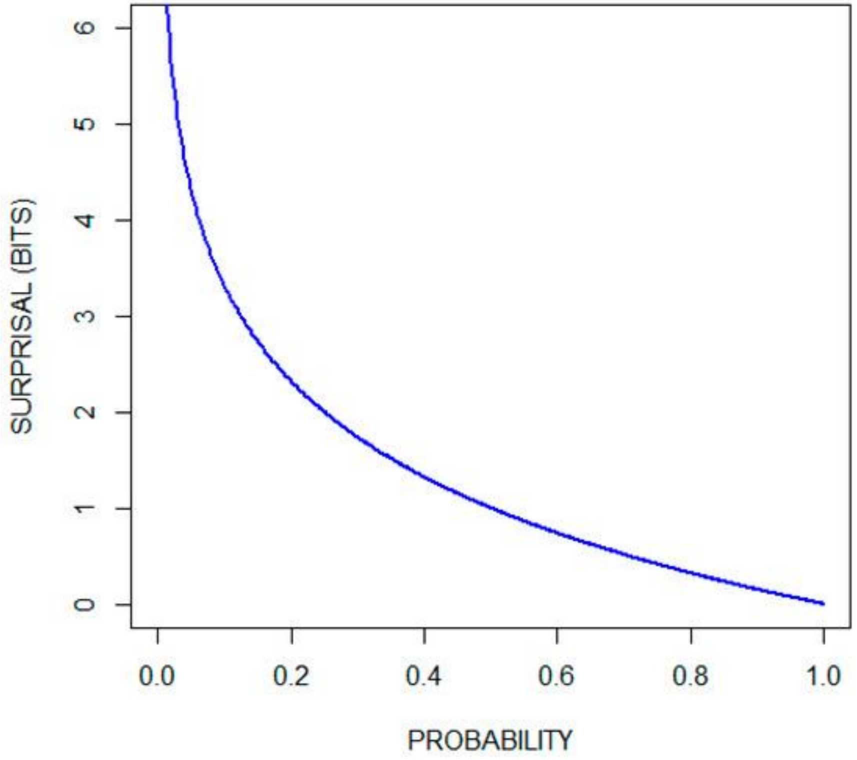
열역학 제2법칙: 엔트로피 증가의 법칙
“Expectation of Surprise” : 놀라움의 정도를 수치화 ≒ 무질서도

$$H(X) = -\sum p(x)\log p(x)$$

$$= \sum p(x)\log\frac{1}{p(x)}$$

Entropy

열역학 제2법칙: 엔트로피 증가의 법칙
"Expectation of Surprise" : 놀라움의 정도를 수치화 ≒ 무질서도



흰공 9개, 검은공 1개가 있다. 어떤 경우가 더 놀라운가?
- 흰공을 뽑았을 때
- 검은공은 뽑았을 때

→ 확률이 높아지면 놀라움 (surprise)는 작아진다.
→ 놀라움은 확률과 반비례
→ $surprise \propto \frac{1}{p(x)}$
→ 그런데 $p(x) = 1$이면 전혀 놀랍지 않음에도 surprise는 양수
→ 로그

$$H(X) = \sum p(x) \log \frac{1}{p(x)}$$

Cross Entropy

엔트로피에서 착안하여: “내가 틀리면 놀랍다”
→ 틀릴 때마다 증가

$$H(p,q) = -\sum p(x) \log q(x)$$

$$= \sum y \log \frac{1}{\hat{y}} + (1-y) \log \frac{1}{1-\hat{y}}$$

Dimensions of a Supervised Learner

모델을 훈련한다. = Task를 이행하기 위해서, 훈련 데이터 셋에 **가장 잘 맞도록** 모델 파라미터를 설정

1. Model: $g(\mathbf{x} \mid \theta)$

모델을 설정.

2. Loss function: $E(\theta \mid \mathcal{X}) = \sum_t L\left(r^t, g\left(\mathbf{x}^t \mid \theta\right)\right)$

정답을 통한 모델의 성능을 측정

3. Optimization procedure: $\theta^* = \arg\min_\theta E(\theta \mid \mathcal{X})$

정답을 가장 잘 맞추는 모델 파라미터를 찾는다.

고려해야 할 point!

1. 어떤 모델을 써야 할까
2. Task에 맞는 loss function 을 써야할까
3. Parameter estimation하는 어떤 estimator 을 써야 할까

Maximum Likelihood Estimator

From Bayes Theorem

모델을 훈련한다. = Task를 이행하기 위해서, 훈련 데이터 셋에 **가장 잘 맞도록** 모델 파라미터를 설정

훈련 데이터 셋에 잘 맞는 파라미터

관측 훈련 데이터 셋이 주어졌을 때, 특정 파라미터의 그럴듯함.

Posterior Probability of parameters

$$P(\theta|X) = \frac{\underbrace{P(X|\theta)}_{\text{MLE}}\,\underbrace{p(\theta)}_{\text{MAP}}}{P(X)}$$

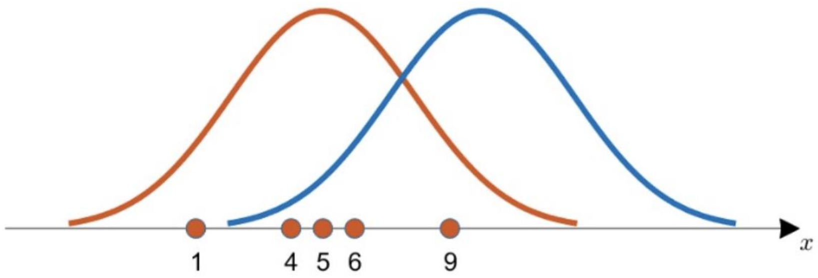
Likelihood Function

$$\underbrace{P(\theta|X)}_{\text{Unknown}} = \frac{P(X|\theta)p(\theta)}{P(X)} \propto \underbrace{P(X|\theta)}_{\text{Likelihood function}}$$

다음과 같이 5개의 데이터를 얻었다고 가정하자.

$$x = \{1, 4, 5, 6, 9\}$$

이 때, 아래의 그림을 봤을 때 데이터 x는 주황색 곡선과 파란색 곡선 중 어떤 곡선으로부터 추출되었을 확률이 더 높을까?



Log Likelihood Function

Definition (Likelihood)

For $X_1, \cdots, X_n \overset{iid}{\sim} f_X(x;\theta)$, where θ denotes a parameter of interest. The likelihood function is

$$L(\theta; \mathbf{X}) = L(\theta; X_1, \cdots, X_n) = \prod_{i=1}^{n} f_X(X_i;\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} P(X|\theta)$$
$$= \arg\max_{\theta} \prod_i P(x_i|\theta)$$

0~1 사이 값으로 이루어진 확률값들의 곱

→ 0으로 가까워져 버린다.

$$\theta_{MLE} = \arg\max_{\theta} \log P(X|\theta)$$
$$= \arg\max_{\theta} \log \prod_i P(x_i|\theta)$$

Log Likelihood function

$$= \arg\max_{\theta} \sum_i \log P(x_i|\theta)$$

Maximum Likelihood Estimator

- What is MLE?

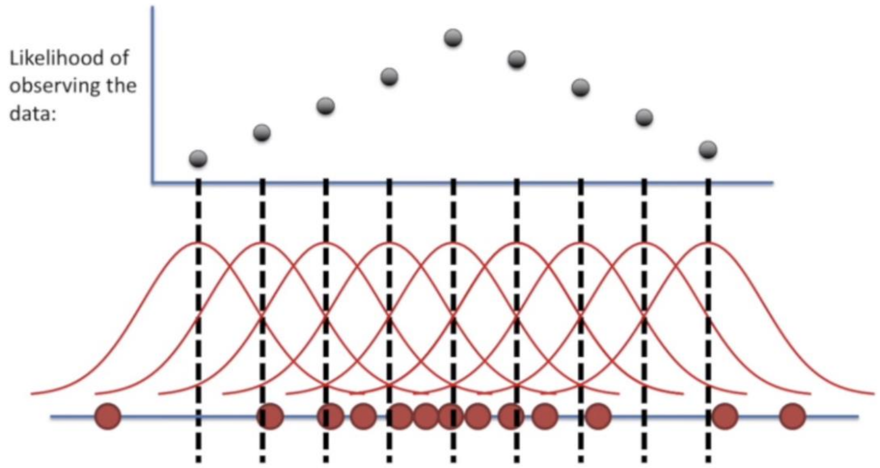
Definition (Maximum likelihood estimator, MLE)

For $X_1, \cdots, X_n \overset{iid}{\sim} f_X(x;\theta)$, the MLE of θ is

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta; \mathbf{x}).$$

which is equivalent to maximize the logarithm of $L(\theta; \mathbf{x})$ which we call the log-likelihood

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}).$$



Log Likelihood Function

- **Bernoulli distribution**

$$\log L(p) = \sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p))$$

- **Binomial distribution**

$$\log L(p) = \log\binom{n}{c} + \sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p))$$

- **Multinomial distribution**

$$\log L(p) = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij} \log p_j$$

- **Normal distribution**

$$\log L(\mu) \approx -\frac{\sum_{i=1}^{n}(y_i - \mu)}{\sigma^2}$$

MLE → Loss Function

정답에 가까운 distribution을 찾아주는 것이 **MLE**

우리가 만든 모델과 정답과의 차이를 보여주는 것이 **loss function**

Loss function penalizes bad predictions.

Regression

- Mean Squared Error

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

Others:
Mean absolute error and mean bias error

Classification

- Binary Cross Entropy

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)$$

- Categorical Cross Entropy

$$CCE = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot \log(\hat{y}_j) + (1-y_j) \cdot \log(1-\hat{y}_j)$$

Others:
Hinge loss / SVM loss.

3. Model Selection

우리의 목표

Train Error / Test Error을 최대한 낮추고 싶다!

우리의 목표

Train Error / Test Error을 최대한 낮추고 싶다!

Error

- Error : deviation from an actual value by a prediction or expectation of that value
- Loss function: 모델의 학습 과정에서 최소화되어야 하는 함수로서 모델의 **오류(Error)**를 정량화하는 역할

~~Error = Variance + Bias~~

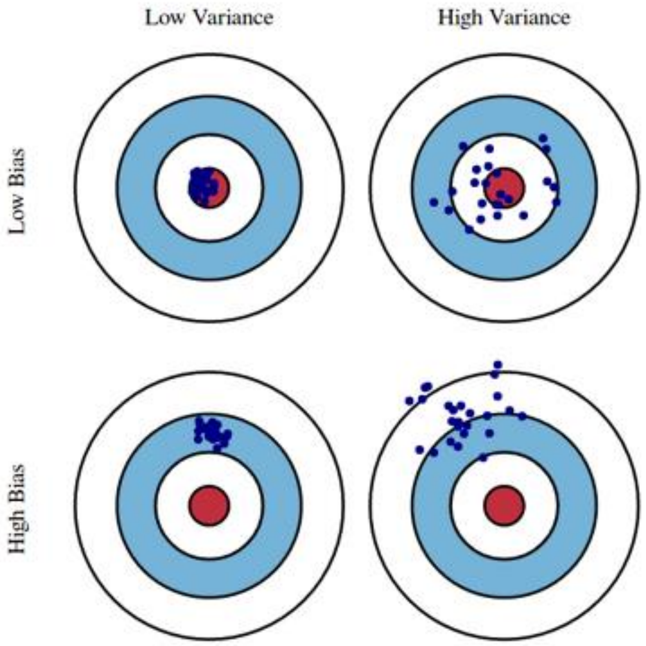


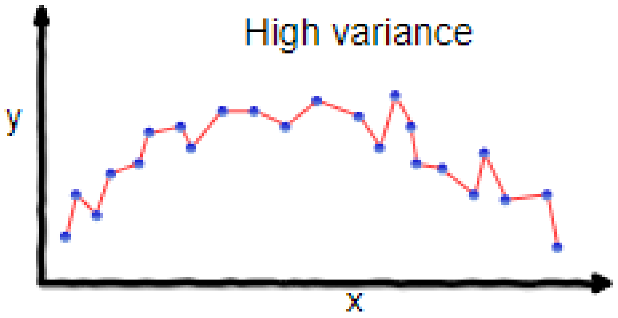
Fig. 1 Graphical illustration of bias and variance.

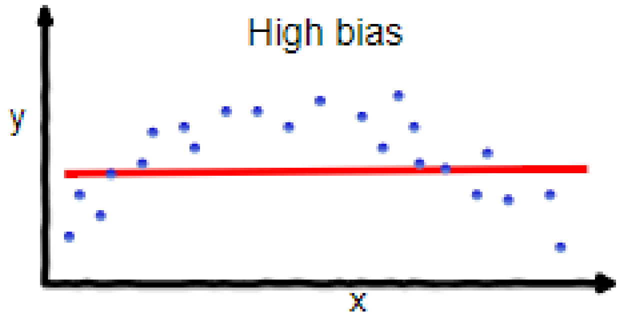
Error = 추정값 - 참값

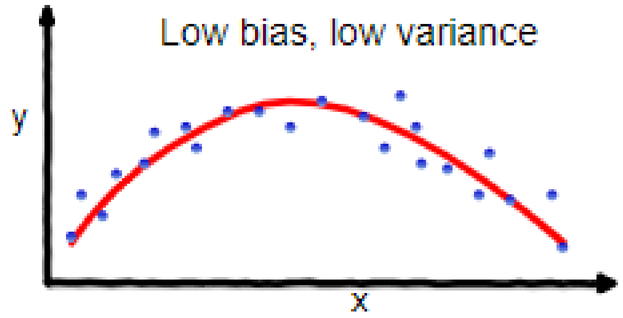
$$y = f(x) + \epsilon$$

$$\begin{aligned} MSE &= E[y - \hat{f}(x)^2] \\ &= E[f(x) + \epsilon - \hat{f}(x)]^2 \\ &= E[f(x) - \hat{f}(x)]^2 + E[\epsilon]^2 + 2E[\epsilon(f(x) - \hat{f}(x)] \\ &= [Var(\hat{f}(x)) + Bias(\hat{f}(x))^2] + Var(\epsilon) \\ &= Reducible\ Error + Irreducible\ Error \end{aligned}$$

Underfitting vs Overfitting







overfitting **underfitting** **Good balance**

이 식으로부터

$$
\begin{aligned}
MSE &= E[y - \hat{f}(x)^2] \\
&= E[f(x) + \epsilon - \hat{f}(x)]^2 \\
&= E[f(x) - \hat{f}(x)]^2 + E[\epsilon]^2 + 2E[\epsilon(f(x) - \hat{f}(x)] \\
&= \boxed{[Var(\hat{f}(x))} + \boxed{Bias(\hat{f}(x))^2]} + Var(\epsilon) \\
&= Reducible\ Error + Irreducible\ Error
\end{aligned}
$$

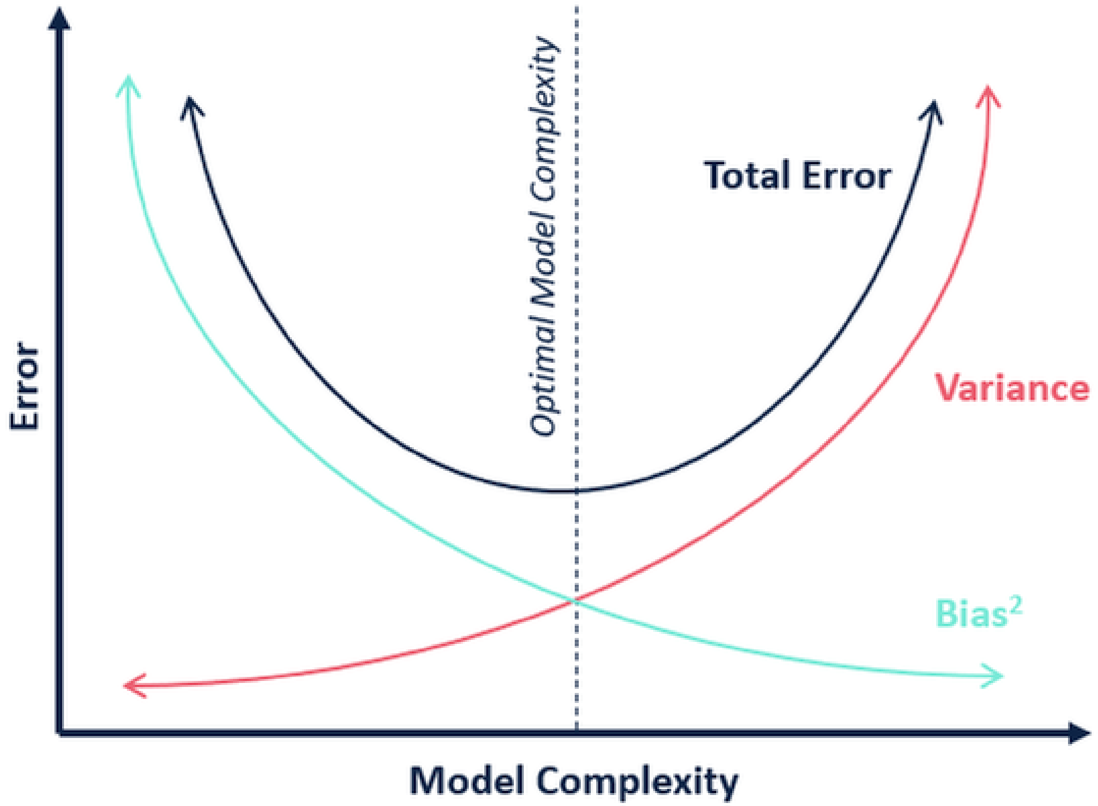
복잡한 모델(overfitting):

- 데이터가 달라질 때 예측값 변동폭이 큼 → high variance
- 참값을 대부분 맞힘 → low bias

단순함 모델(underfitting):

- 데이터가 달라져도 예측값의 변동폭이 작음 → low variance
- 참값을 대부분 벗어남 → high bias

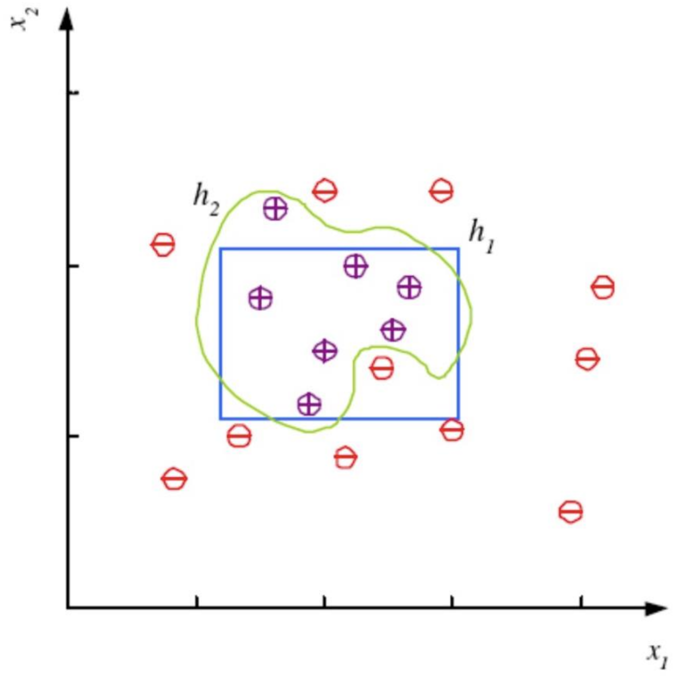
Trade-off



Bias / Variance dilemma : Geman et al. 1992

Model Selection

Inductive Bias : Occam's Razor



<u>If performances are similar,</u>

Use the simpler one because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance)

Cross Validation

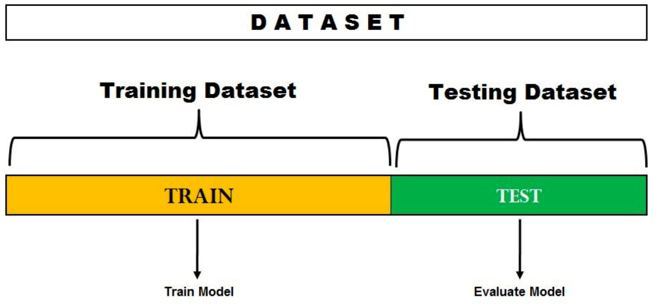
To estimate generalization error, we need data unseen during training.

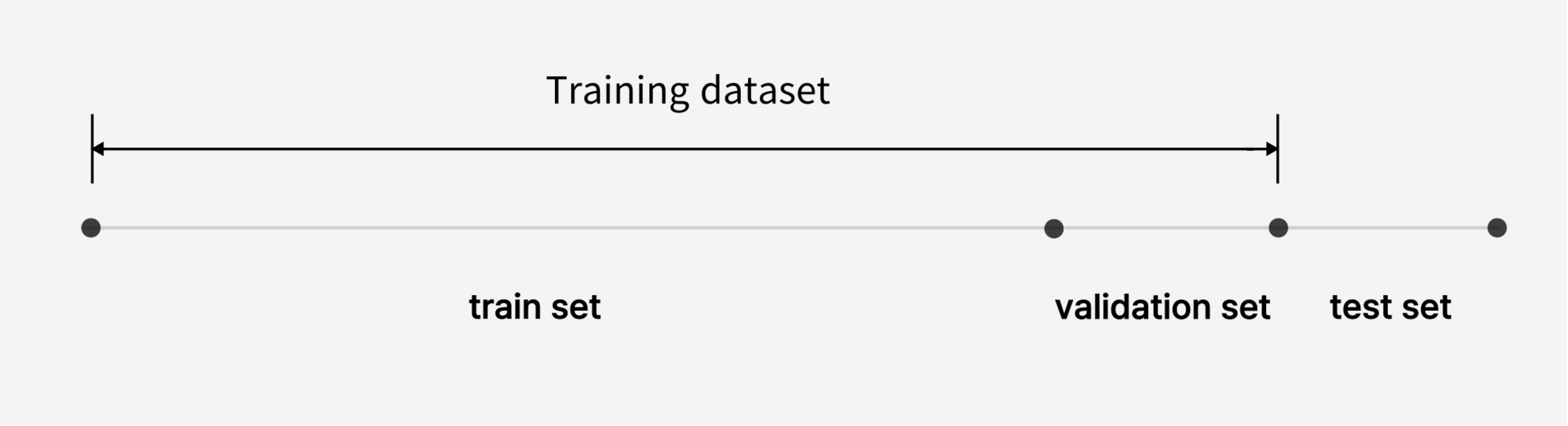
We split the data as

- – Training set (50%)
- – Validation set (25%)
- – Test (publication) set (25%)

Measure generalization accuracy by testing on data unused during training

Hold out





Regularization

Penalize complex models

- E'=error on data + λ*model complexity

* If λ increases, variance decreases, but bias increases

In regression...

Regularization (L2): $$E(\mathbf{w}|\mathcal{X})=\frac{1}{2}\sum_{t=1}^{N}\left[r^t-g(x^t|\mathbf{w})\right]^2+\lambda\sum_i w_i^2$$

수고하셨습니다!

해당 세션자료는 KUBIG Github에서 보실 수 있습니다!

다음은 이번 주차 과제 설명이 있습니다!