목차

교육 일정

머신러닝을 활용한 나노/탄소 소재 생산공정 최적화

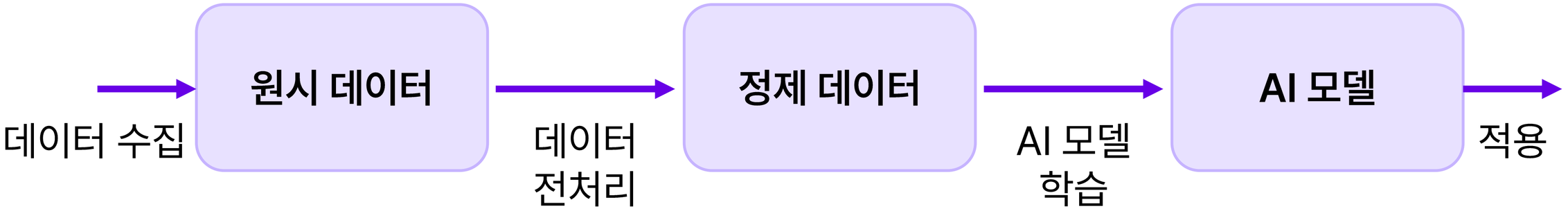
02 AI 적용을 위한 데이터셋 전처리

목차

AI 적용을 위한
데이터셋 전처리

EliceAcademy

AI 모델 적용의 순서



EliceAcademy

데이터 전처리의 종류

- **데이터 청소**
  수집 과정에서 생긴 **이상치** 및 **결측치** 제거

- **데이터 라벨링**
  AI 모델 학습에 필요한 **정답**을 데이터에 매핑

- **데이터 정리**
  불필요한 데이터를 **제거 및 간소화**

- **데이터 재구성**
  정규화, 표준화 등을 사용하여 데이터를 분석하기 적합하게 변환

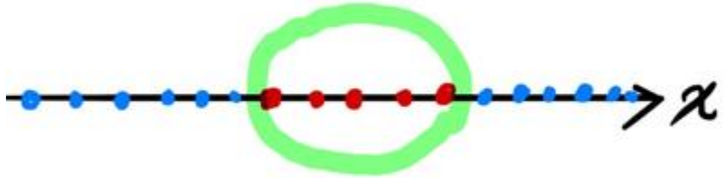
데이터 전처리의 방법

- **의미적 데이터 전처리**
  - 도메인 지식을 활용하여 데이터 전처리 과정 수행
  - *e.g.*, 사람의 성별을 예측하는 task에서 “키”는 중요한 feature로 삼고, “생년월일”은 중요하지 않은 feature로 삼음

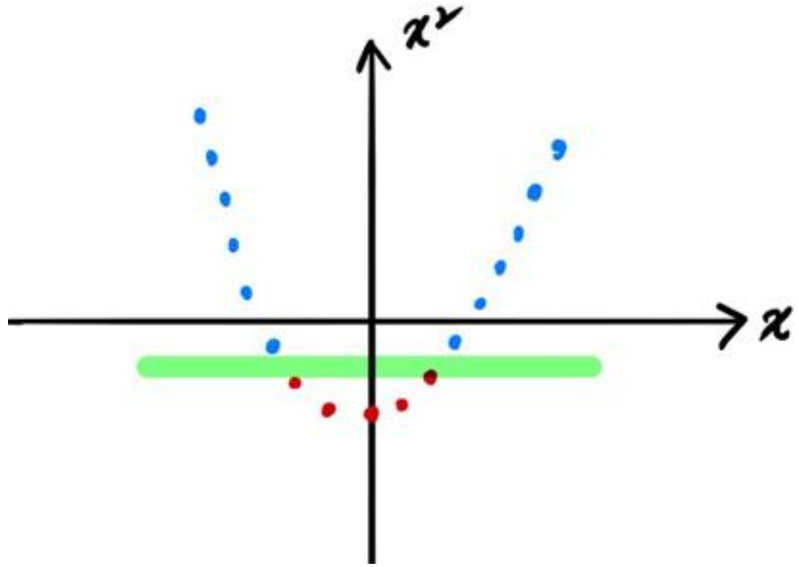
- **기계적 데이터 전처리**
  - 도메인 지식과 상관 없이 데이터 값만으로 전처리 과정 수행
  - *e.g.*, 상관관계가 높은 2개 feature를 하나의 feature로 통합, 데이터 정규화 및 표준화

EliceAcademy

데이터 전처리의 중요성



- 전처리 전 데이터
- 빨간 점과 파란 점을 나누기 위해 복잡한 모델 (원)이 필요



- 전처리 후 데이터
- 빨간 점과 파란 점을 나누기 위해 단순한 모델 (선)로 가능

https://sanghyu.tistory.com/14

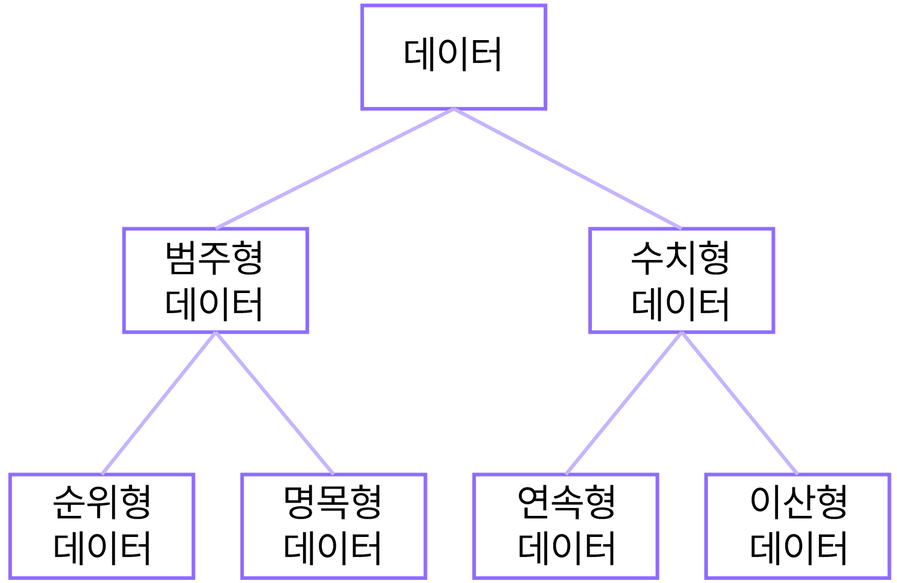
02

자료 형태에 따른 데이터 전처리

정형 데이터의 분류

- **범주형 데이터**
  - 범주로 나타내어지는 데이터
  - **순위형 데이터**
    범주 간의 순서가 있는 데이터 (*e.g.*, 선호도, 성적)
  - **명목형 데이터**
    범주 간의 순서가 없는 데이터 (*e.g.*, 혈액형, 성별)

- **수치형 데이터**
  - 수치로 측정되는 데이터
  - **연속형 데이터**
    실수로 표현 가능한 연속적인 데이터 (*e.g.*, 키, 몸무게, 온도, 습도)
  - **이산형 데이터**
    정수로 표현 가능한 셀 수 있는 데이터 (*e.g.*, 참여 인원, 나이)

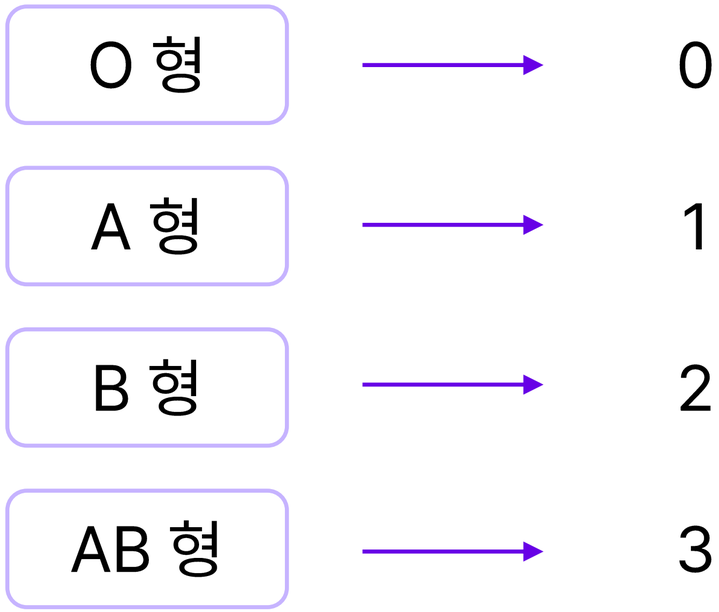


타이타닉 데이터

변수 명	변수 설명
PassengerID	승객의 고유 번호 **(수치형 – 이산형)**
Survived	생존 여부 (0: 사망, 1: 생존) **(범주형 – 명목형)**
Pclass	객실 등급 (Upper, Middle, Lower) **(범주형 – 순위형)**
Name	이름 **(범주형 – 명목형)**
Sex	성별 **(범주형 – 명목형)**
Age	나이 **(수치형 – 이산형)**
SibSp	동반 형제자매와 배우자 수 **(수치형 – 이산형)**
Parch	동반 부모와 자식의 수 **(수치형 – 이산형)**
Ticket	티켓의 고유 번호 **(수치형 – 이산형)**
Fare	요금 **(수치형 – 이산형)**
Cabin	객실 번호 **(수치형 – 이산형)**
Embarked	승선한 항 (C, Q, S) **(범주형 – 명목형)**

범주형 데이터 전처리

- 수치 매핑 변환



EliceAcademy

범주형 데이터 전처리

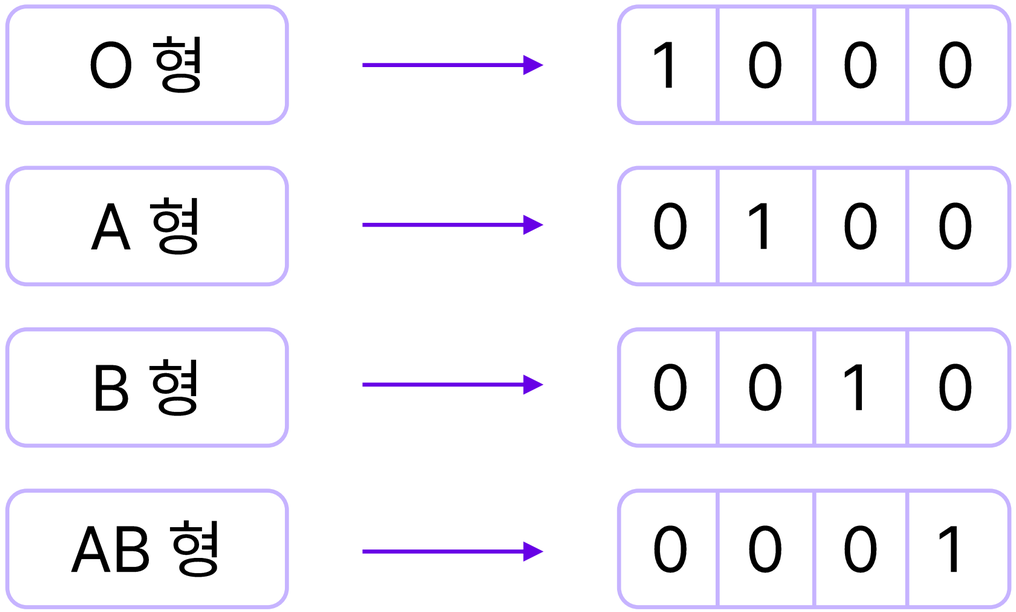
```
titanic['SexMapping'] = titanic['Sex'].replace({'male': 0, 'female': 1})
```

	Sex	SexMapping
0	male	0
1	female	1
2	female	1
3	female	1
4	male	0

- 수치 매핑 변환
- replace 메소드 활용
- 남자는 0, 여자는 1로 매핑

범주형 데이터 전처리

- 더미 기법



범주형 데이터 전처리

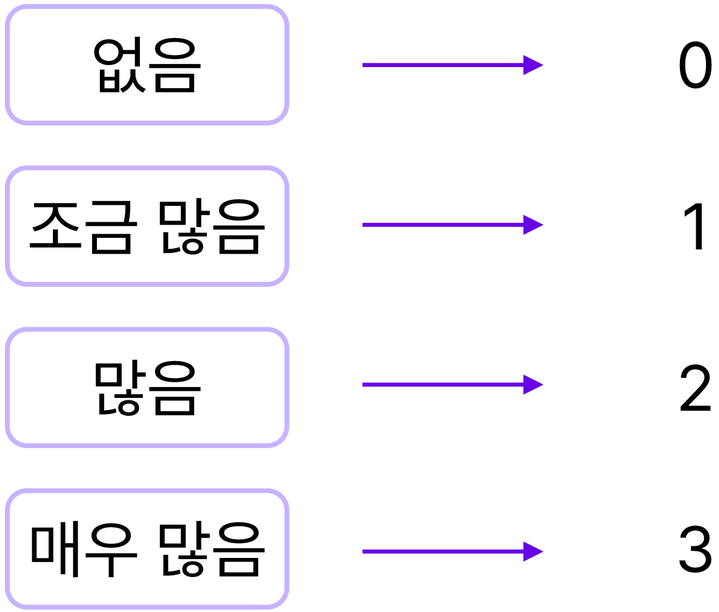
```
dummies = pd.get_dummies(titanic['Sex'])
```

	Sex	female	male
0	male	0	1
1	female	1	0
2	female	1	0
3	female	1	0
4	male	0	1

- 더미 기법
- get_dummies 메소드 활용
- 남자는 [0, 1], 여자는 [1, 0] 으로 매핑

범주형 데이터 전처리

- 순서형 자료: 수치 매핑 방식



EliceAcademy

수치형 데이터의 전처리

	Fare
0	7.2500
1	71.2833
2	7.9250
3	53.1000
4	8.0500
...	...
886	13.0000
887	30.0000
888	23.4500
889	30.0000
890	7.7500

정규화 or 표준화

	FareNormalized
0	0.014151
1	0.139136
2	0.015469
3	0.103644
4	0.015713
...	...
886	0.025374
887	0.058556
888	0.045771
889	0.058556
890	0.015127

	FareStandarized
0	-0.502163
1	0.786404
2	-0.488580
3	0.420494
4	-0.486064
...	...
886	-0.386454
887	-0.044356
888	-0.176164
889	-0.044356
890	-0.492101

03

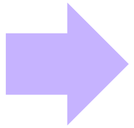
데이터 정규화와 표준화

수치형 데이터 전처리

- 정규화
  - $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$
  - 데이터를 0~1 사이의 값으로 변환

	신장
0	175
1	170
2	182
3	158
4	162
5	174

데이터 정규화 전



	신장
0	0.708333
1	0.500000
2	1.000000
3	0.000000
4	0.166667
5	0.666667

데이터 정규화 후

수치형 데이터 전처리

```
titanic['FareNormalized'] = (titanic['Fare'] - titanic['Fare'].min()) / (titanic['Fare'].max() - titanic['Fare'].min())
```

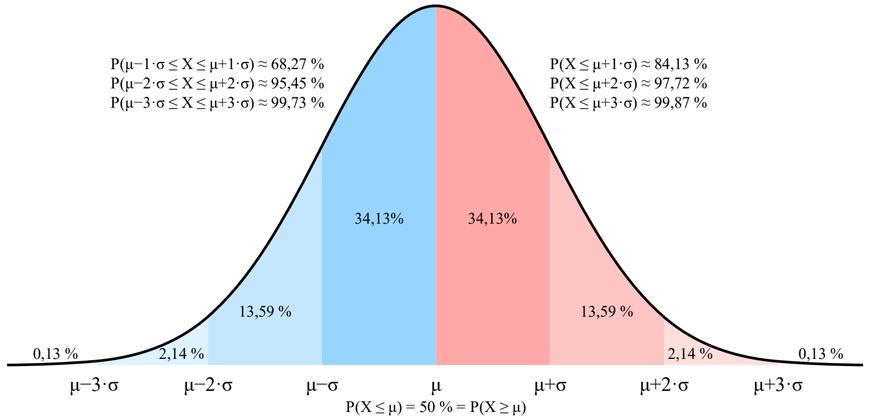
	Fare	FareNormalized
0	7.2500	0.014151
1	71.2833	0.139136
2	7.9250	0.015469
3	53.1000	0.103644
4	8.0500	0.015713

- 데이터 정규화
- 데이터를 0~1 사이로 변환
- 최소값은 0, 최대값은 1

EliceAcademy

수치형 데이터 전처리

- 표준화
  - $X' = \frac{X - \mu}{\sigma}$
  - μ 는 X 의 평균
  - σ 는 X 의 표준편차
  - 데이터를 정규분포를 따르도록 변환



	신장
0	175
1	170
2	182
3	158
4	162
5	174

데이터 표준화 전

	신장
0	0.545290
1	-0.018803
2	1.335020
3	-1.372627
4	-0.921352
5	0.432471

데이터 표준화 후

수치형 데이터 전처리

```
titanic['FareStandarized'] = (titanic['Fare'] - titanic['Fare'].mean()) / titanic['Fare'].std()
```

	Fare	FareStandarized
0	7.2500	-0.502163
1	71.2833	0.786404
2	7.9250	-0.488580
3	53.1000	0.420494
4	8.0500	-0.486064

- 데이터 표준화
- 데이터를 정규분포로 변환
- 평균은 0, 표준편차는 1

EliceAcademy

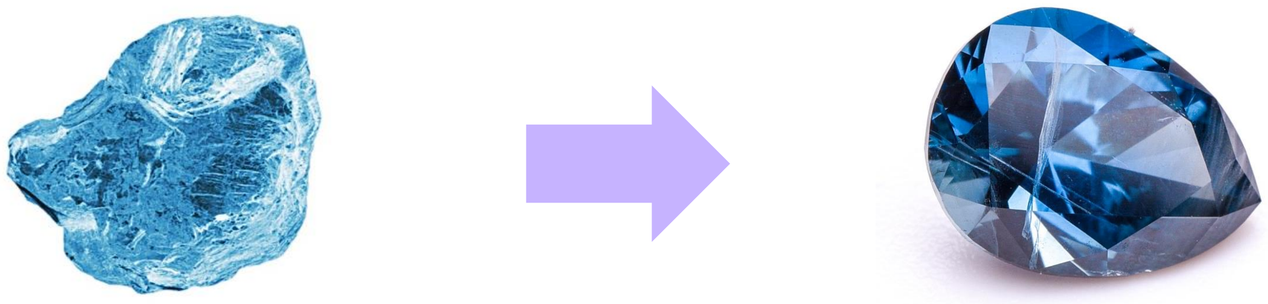
04

피쳐 엔지니어링

EliceAcademy

피쳐 엔지니어링이란?

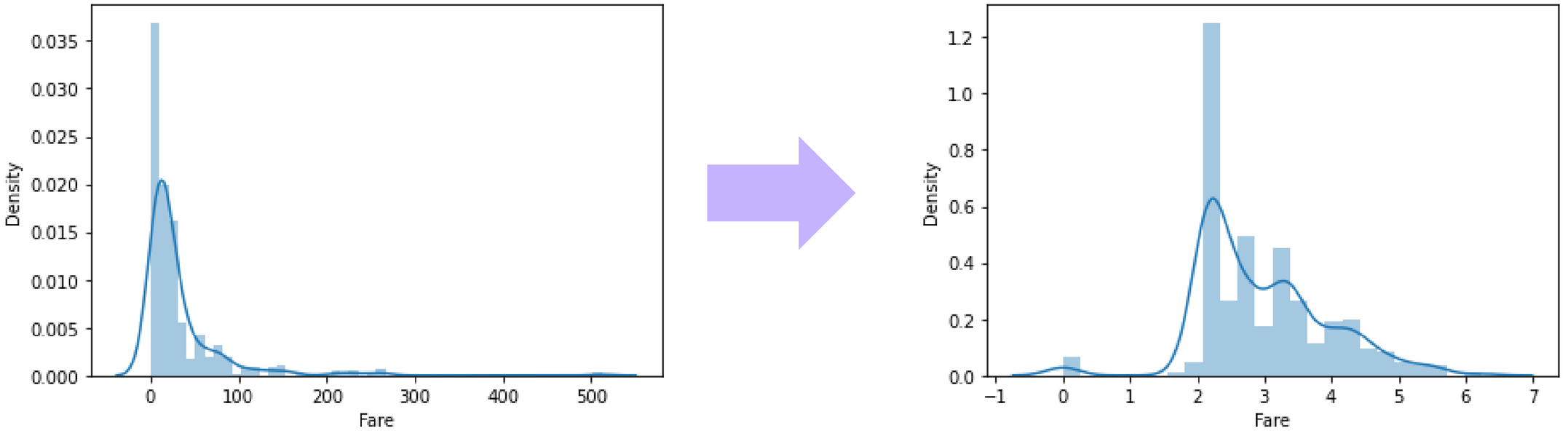
- 도메인 지식을 활용하여 원시 데이터 (raw data)로부터 적절한 피쳐 (feature)를 만드는 과정
- 이러한 피쳐들은 머신러닝 알고리즘의 성능을 향상시킴
- 데이터 표준화와 정규화도 피쳐 엔지니어링의 일종



EliceAcademy

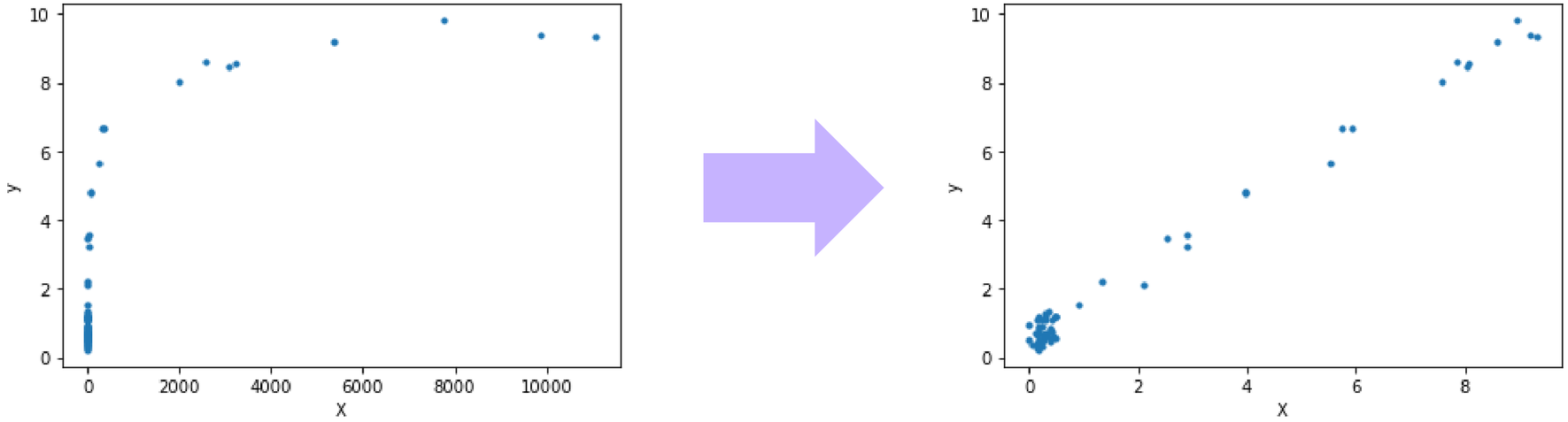
피쳐 엔지니어링의 예시

- 로그 변환
  타이타닉 데이터의 Fare 칼럼



피쳐 엔지니어링의 예시

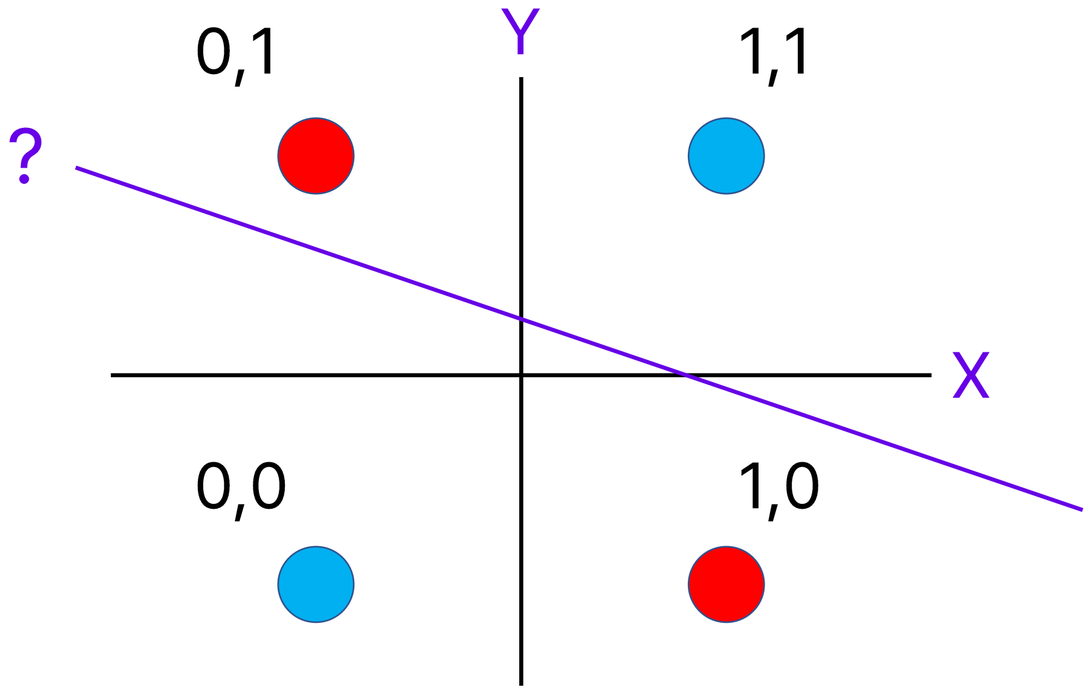
- 로그 변환
  - 변환 전에는 X와 Y 사이의 관계를 추론하기 어려움.
  - 변환 후에는 X와 Y 사이에 양의 상관관계가 있음을 추론 가능.



피쳐 엔지니어링의 예시

- XOR 분류

X	Y	
0	0	0
1	0	1
0	1	1
1	1	0

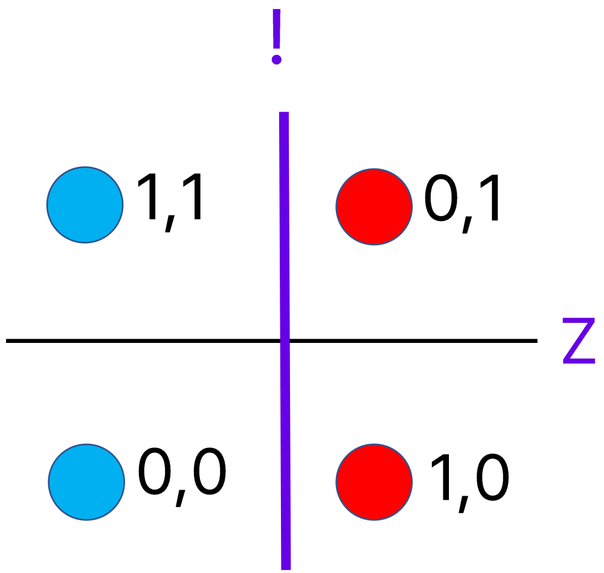


피쳐 엔지니어링의 예시

- XOR 분류

$$Z = (X - Y)^2$$

X	Y		Z 피쳐		
0	0	→	$(0-0)^2 = 0$	→	0
1	0	→	$(1-0)^2 = 1$	→	1
0	1	→	$(0-1)^2 = 1$	→	1
1	1	→	$(1-1)^2 = 0$	→	0



피쳐 엔지니어링의 예시

- 상호작용 피쳐
  - 한 개의 피쳐만을 사용하는 것보다 **여러 개의 피쳐**를 사용하면 모델의 성능이 향상될 수 있음
  - 누군가의 기대수명을 예측하려 할 때, 성별만 아는 것보다, 성별과 흡연 여부를 함께 알면 더 정확하게 예측 가능
  - 그러나, **모델의 복잡도가 증가**하여 요구되는 메모리와 연산량이 증가됨.
  - 어떠한 피쳐를 사용할지 잘 선택하는 것이 중요

EliceAcademy

피쳐 엔지니어링의 예시

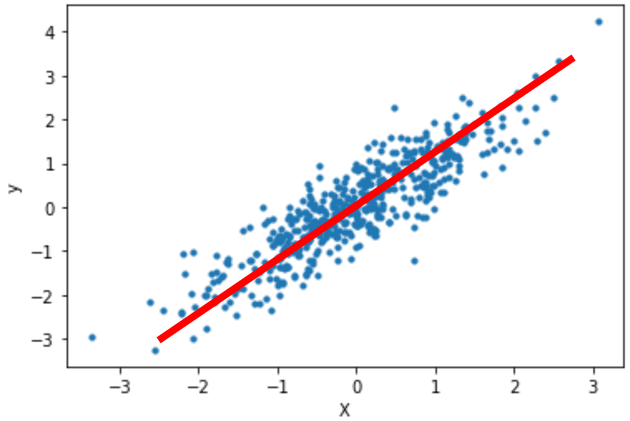
- 피쳐 선택

기대 수명

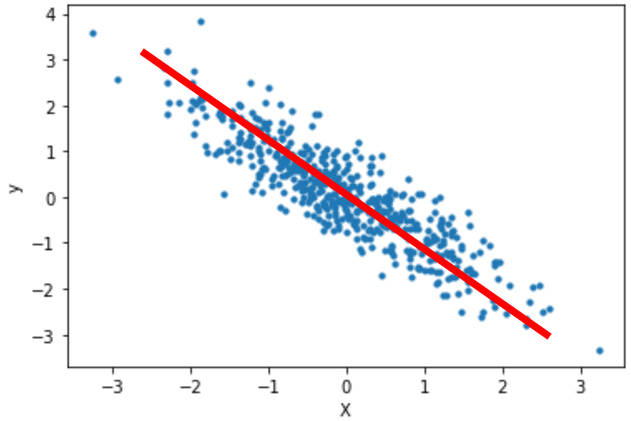
- 도메인 지식 이용
  - 성별과 흡연 여부는 기대 수명에 영향을 미칠 것이다.
  - 태어난 날의 기온과 기대 수명은 관련이 없을 것이다.
- 상관 관계 분석
  - 혼인 여부와 기대 수명은 관계가 있을까?

상관 분석

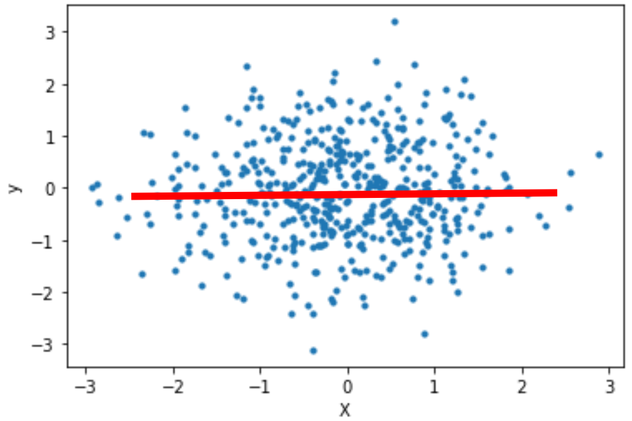
- 상관 관계
  한 변수가 증가하면 다른 변수도 선형적으로 증가하거나 감소하는지를 나타내는 지표



양의 상관관계



음의 상관관계



상관관계 없음

상관 분석

- 상관 계수
  - 상관 관계의 크기를 나타내는 값
  - -1 ~ 1 사이의 값
    - 1: 매우 높은 양의 상관 관계
    - 0: 상관 관계 없음
    - -1: 매우 높은 음의 상관 관계
  - 피어슨 상관 계수
    - $\rho = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}}$

EliceAcademy

상관 분석

- 타이타닉 데이터의 상관 분석

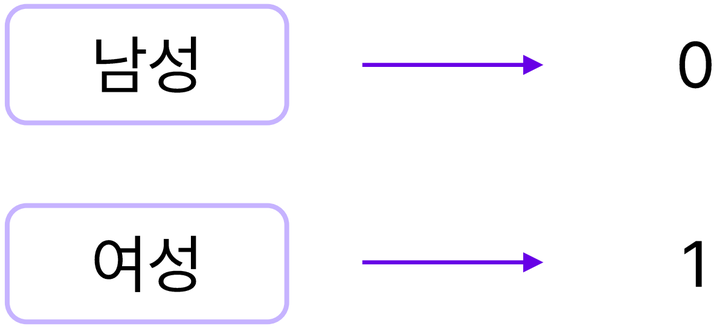
```
titanic.corr()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

- **승객의 고유번호 (PassengerId)**와 생존율 (Survived)는 가장 낮은 상관관계를 보였다.
- **객실의 등급 (Pclass)**과 생존율은 가장 높은 상관관계를 보였다.
- 생존율과 생존율의 상관관계는 1이다.

상관 분석

- 상관 분석은 수치형 자료만 가능!
- 범주형 자료를 수치 매핑 변환하여 수치형 자료로 만들 수 있음



EliceAcademy

상관 분석

```
titanic.corr().style.background_gradient()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	SexMapping
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658	-0.042939
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307	0.543351
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500	-0.131900
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067	-0.093254
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651	0.114631
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225	0.245489
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000	0.182333
SexMapping	-0.042939	0.543351	-0.131900	-0.093254	0.114631	0.245489	0.182333	1.000000

- 범주형 자료인 성별을 **수치 매핑 변환**하여 수치형 자료로 변환
- 성별과 생존율에 가장 큰 상관 관계가 있음.

EliceAcademy

상관 분석의 활용

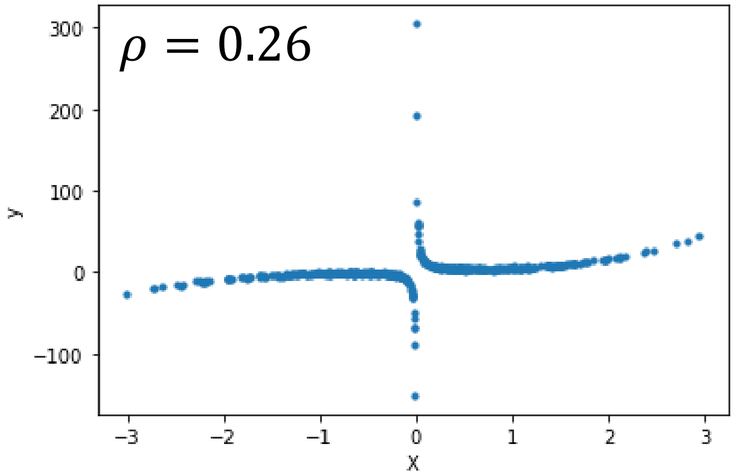
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	SexMapping
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658	-0.042939
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307	0.543351
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500	-0.131900
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067	-0.093254
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651	0.114631
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225	0.245489
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000	0.182333
SexMapping	-0.042939	0.543351	-0.131900	-0.093254	0.114631	0.245489	0.182333	1.000000

- **필요 없는 피쳐 제거**
  - PaasengerId 는 생존율과의 상관 관계가 낮으므로 제거
- **상관관계가 높은 피쳐 중복 제거**
  - Fare는 생존율과 상관 관계가 높음
  - 그러나, Pclass와 Fare 역시 상관관계가 높음
  - Fare와 생존율의 상관관계가 높은 원인이 Pclass 일 수 있음.
  - 따라서, Fare는 Pclass와 의미적으로 중복 될 수 있으므로 제거

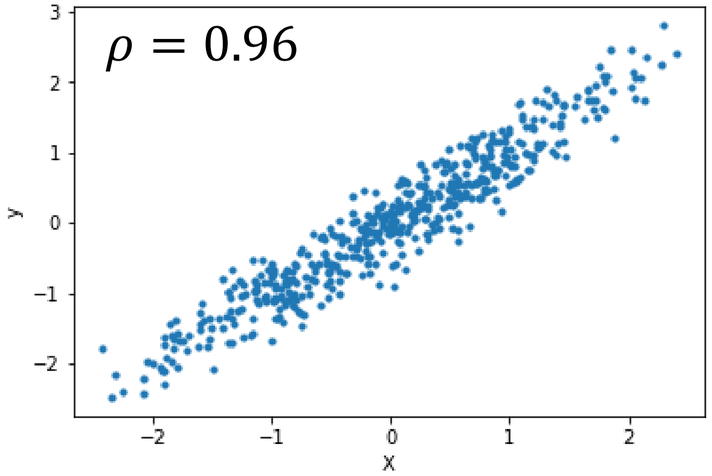
상관 분석의 한계

- 상관 분석은 두 변수 간 선형 관계의 파악만 가능

  비선형적 관계에 있는 두 변수의 상관관계를 파악하지 못함



$$y = e^X + X^3 + \frac{1}{X}$$



$$y = X$$

상관 분석의 한계

- 상관 관계가 높다고 해서, 무조건 두 변수가 유의미한 관계가 있는 것은 아님
  - 두 변수에 동시에 영향을 미치는 다른 변수에 의해 상관 계수가 높은 것일 수 있음
  - *e.g.*, 노인들 중 악력이 센 사람이 건강하다는 상관관계 발견
    - 악력이 세서 건강하다?
    - 꾸준한 운동으로 인해 악력이 세고, 건강하다?

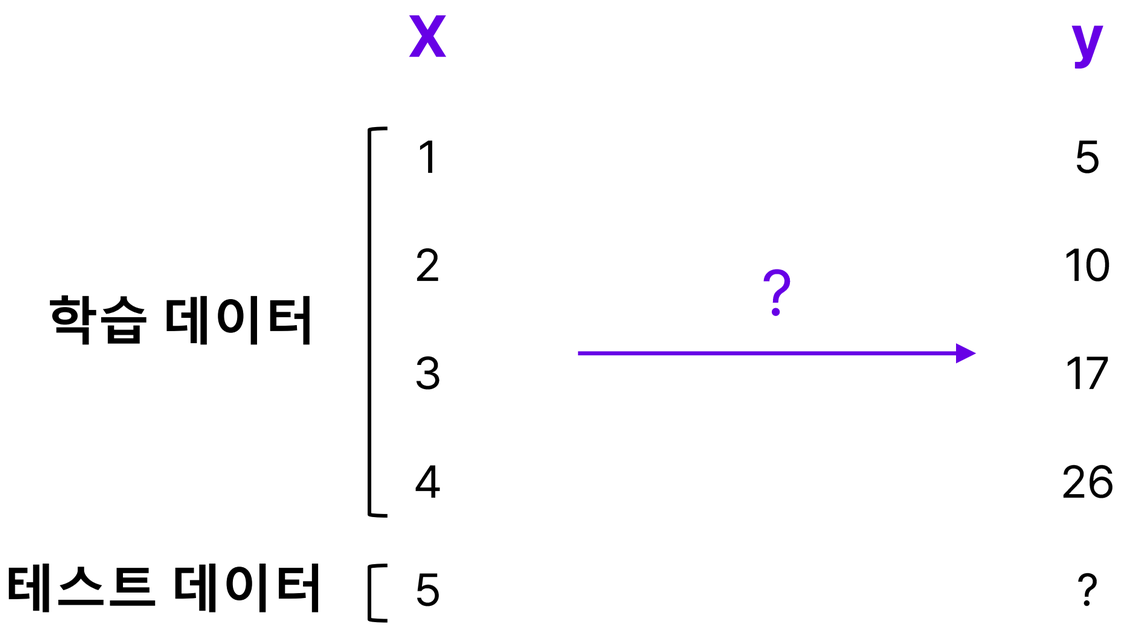
05

데이터 분할

EliceAcademy

회귀 분석

- 입력 변수와 출력 변수 사이의 관계를 모델링하여, 입력값에 대한 출력값을 예측하는 일.



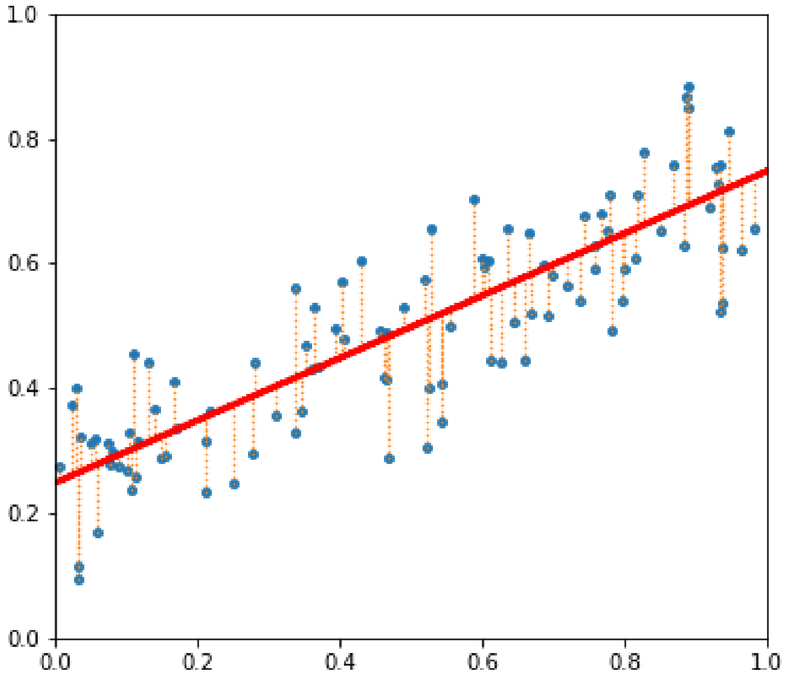
회귀 분석

- 입력 변수와 출력 변수 사이의 관계를 모델링하여, 입력값에 대한 출력값을 예측하는 일.

	X		y
학습 데이터	1	$(X+1)^2+1$	5
	2		10
	3		17
	4		26
테스트 데이터	5		**37!**

회귀 분석

- 회귀의 기본 원칙은 잔차를 최소화 하는 것



- 파란 점: 학습 데이터
- 빨간 선: 회귀 모델
- 주황 점선: 잔차

과적합

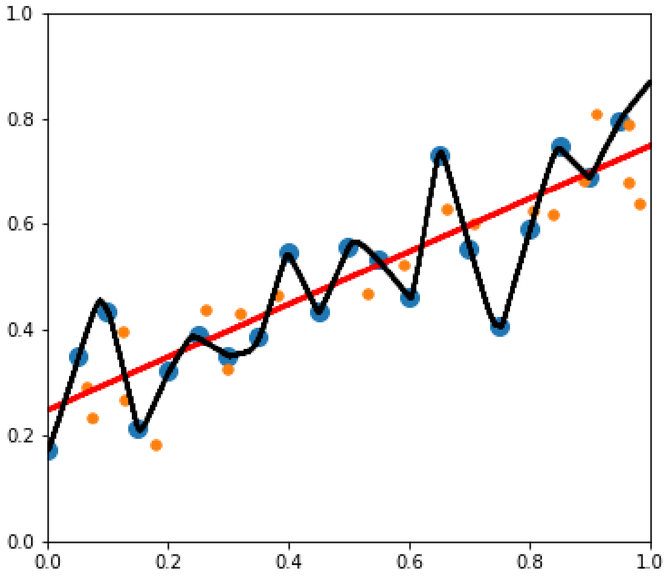
- 학습용 데이터의 잔차를 무조건 최소화 하면 좋은가?

  NO!

- 학습용 데이터로 훈련하여 평가용 데이터에 적용

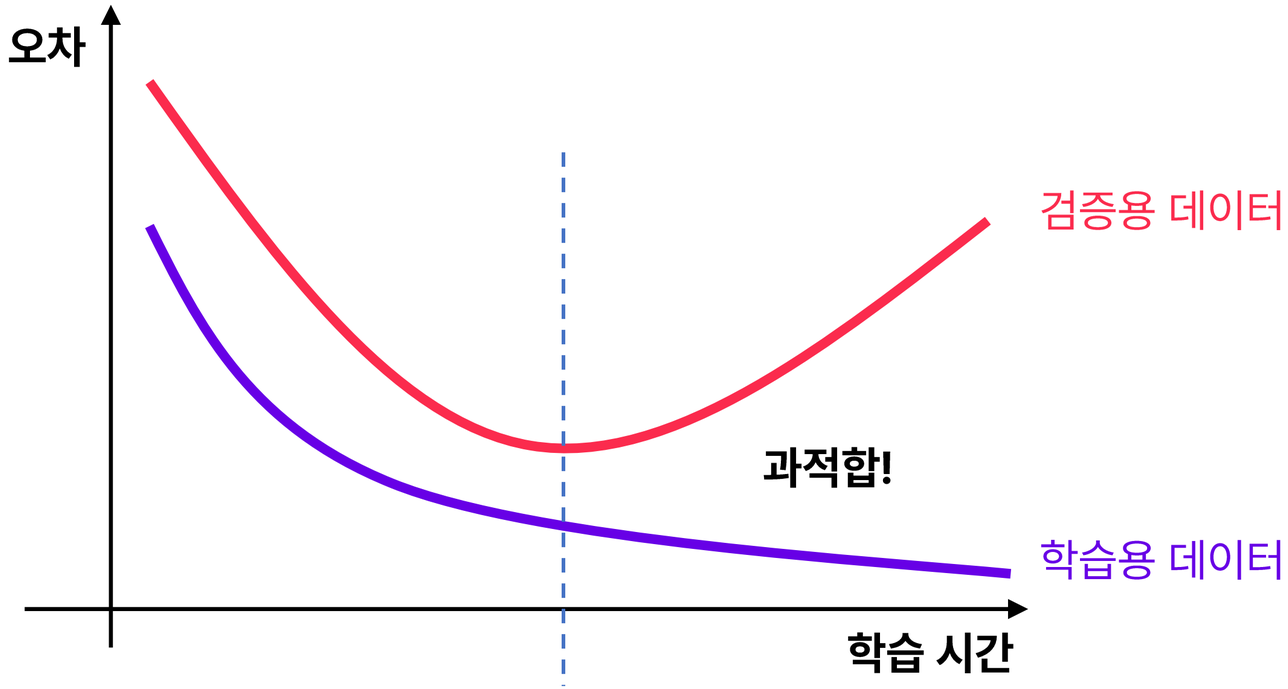
  학습 데이터에 과도하게 적합 된 모델은 일반화가 어려움

과적합



- 파란 점: 학습 데이터
- 주황 점: 테스트 데이터
- 검은 선: 과적합 된 모델
  - 학습 오차: 0
  - 테스트 오차: 0.19
- 빨간 선: 일반화가 잘 된 모델
  - 학습 오차: 0.07
  - 태스트 오차: 0.08

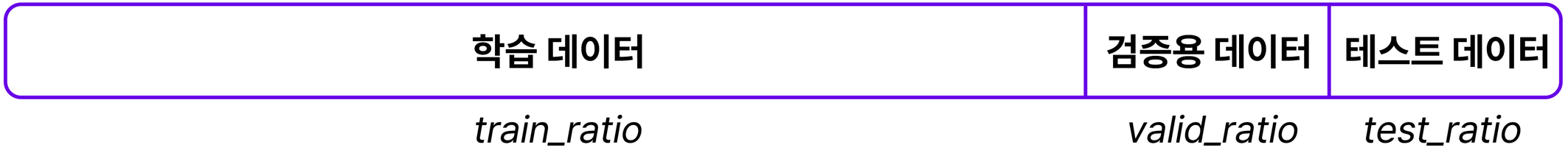
과적합



데이터의 분할

- 학습 데이터
  - 전체 데이터 중 대부분은 모델 학습에 사용
- 검증용 데이터
  - 전체 데이터 중 일부는 검증에 사용
  - 검증용 데이터에 대한 오차가 증가하면 과적합
- 테스트 데이터
  - 전체 데이터 중 일부는 테스트에 사용
  - 최종 모델 성능 평가 목적

데이터의 분할




```
train = titanic[:int(len(titanic) * train_ratio)]
valid = titanic[int(len(titanic) * train_ratio) : int(len(titanic) * train_ratio) + int
(len(titanic) * valid_ratio)]
test = titanic[int(len(titanic) * train_ratio) + int(len(titanic) * valid_ratio):]
```

EliceAcademy

06

머신러닝 학습을 위한 데이터 정제

Pandas Dataframe의 Numpy array 변환

- 머신러닝 모델 학습을 위해서 Pandas Dataframe 데이터 형식을 Numpy array로 변환해주어야 함.
- Stage 1
  - 입력: **Machine 1, 2, 3 + Combiner + AmbientConditions**
  - 출력: **Stage1 측정값**
- Stage 2
  - 입력: **Stage1 측정값 + Machine 4, 5 + AmbientConditions**
  - 출력: **Stage2 측정값**

Pandas Dataframe의 Numpy array 변환

```
data_np = data.values # numpy array 변환
np.save('./Data/train_data.npy', data_np)
```

```
train_data = np.load('./Data/train_data.npy', allow_pickle = True)
```

- Pandas dataframe 을 Numpy array로 변환
- 변환한 Numpy array를 npy 파일로 저장
- 저장된 npy 파일을 로드하여 numpy array 읽어오기

Quiz

- 범주형 데이터의 전처리 방법으로 옳은 것을 모두 고르세요.

① 더미 기법

② 수치 변환 기법

③ 정규화

④ 표준화

Quiz

- 수치형 데이터의 전처리 방법으로 옳은 것을 모두 고르세요.

① 더미 기법

② 수치 변환 기법

③ 정규화

④ 표준화

Quiz

- 원시 데이터를 적절한 피쳐로 가공하는 일련의 과정을 무엇이라 부르나요?
  - 피쳐 엔지니어링

Quiz

- 전체 데이터를 머신러닝에 활용하기 위해 어떠한 데이터 셋들로 분할하나요?
  - 학습용 데이터
  - 검증용 데이터
  - 테스트 데이터