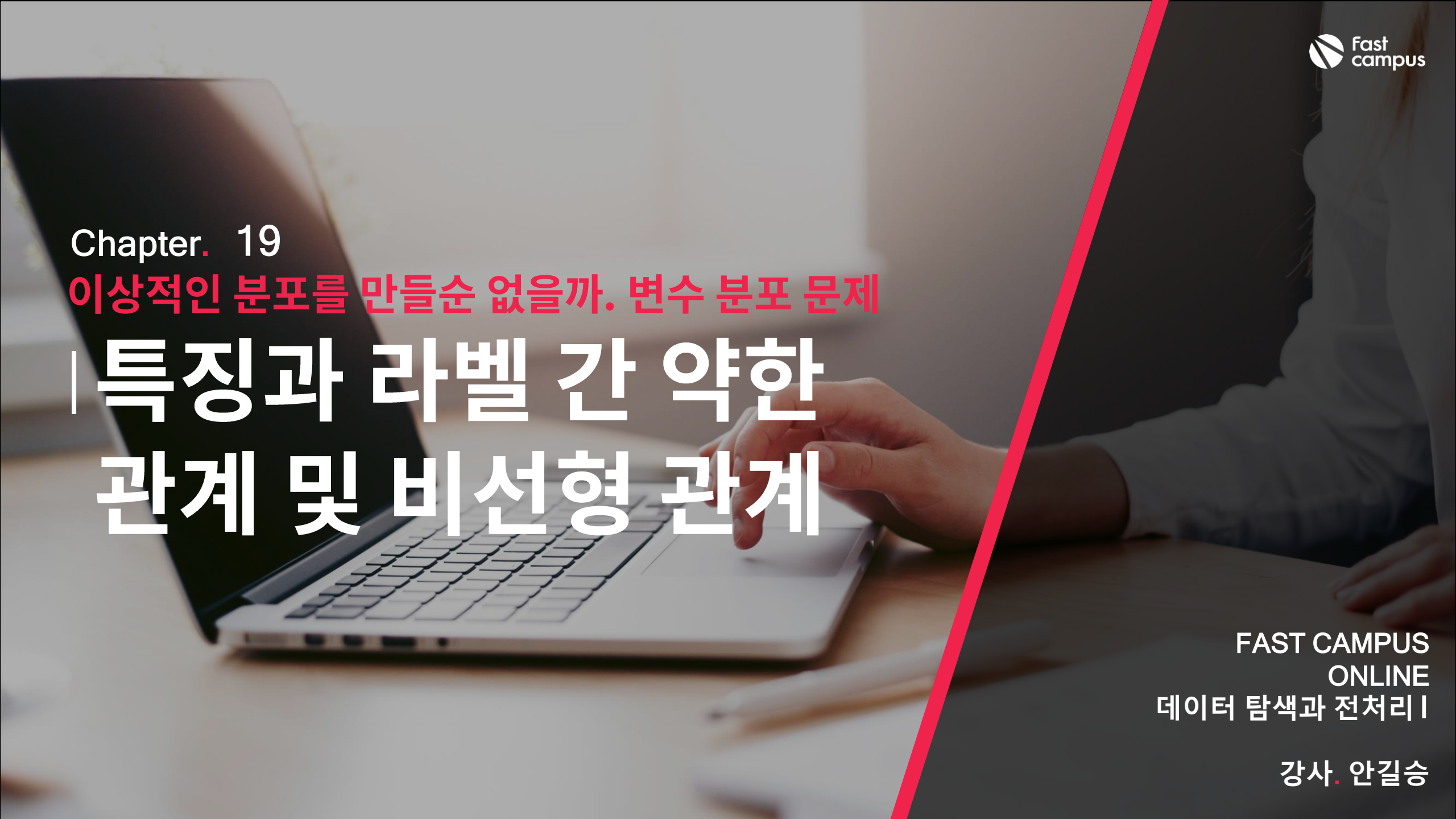


Chapter. 19

이상적인 분포를 만들순 없을까. 변수 분포 문제

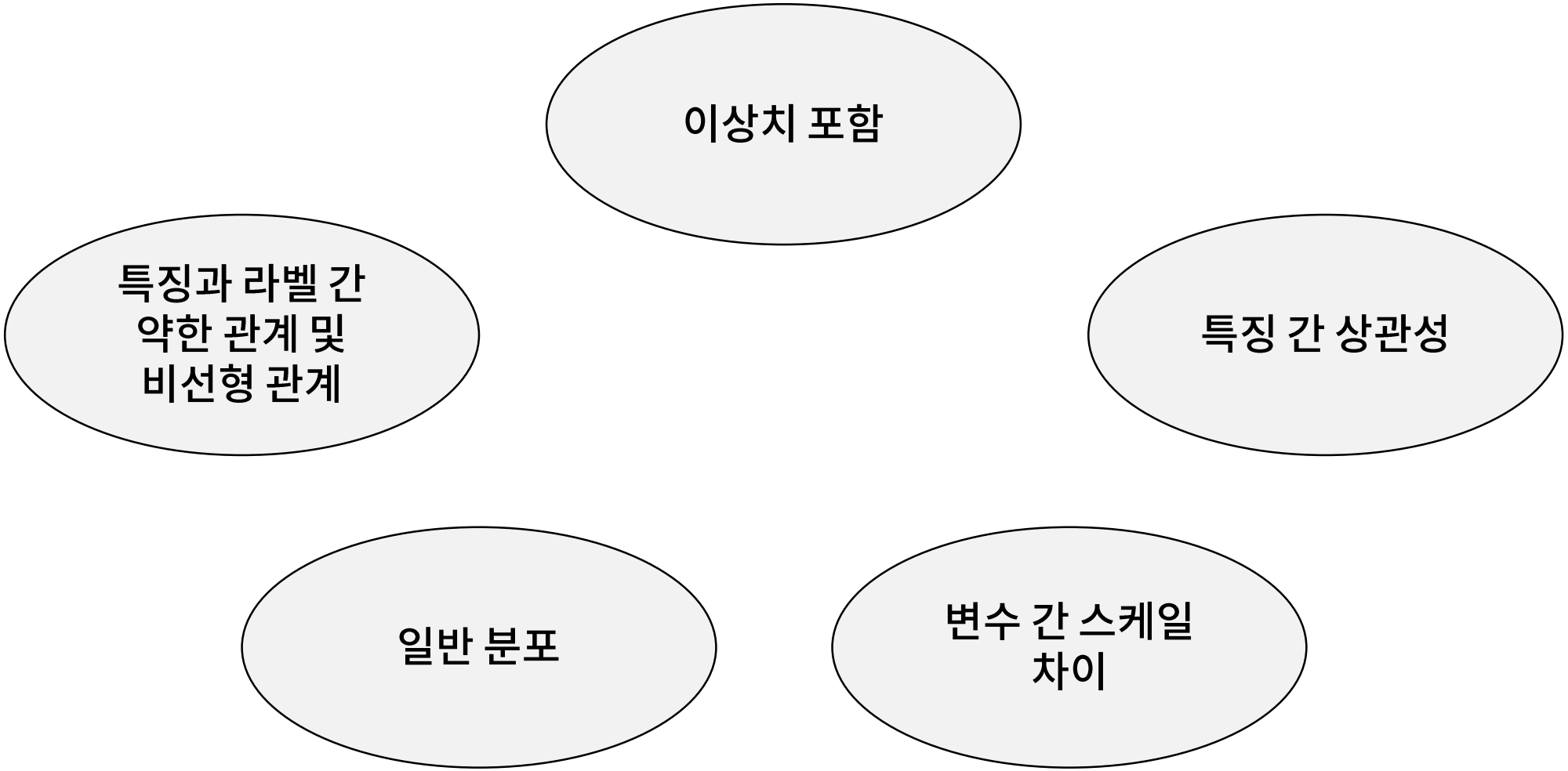
특징과 라벨 간 약한 관계 및 비선형 관계

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 들어가기전에

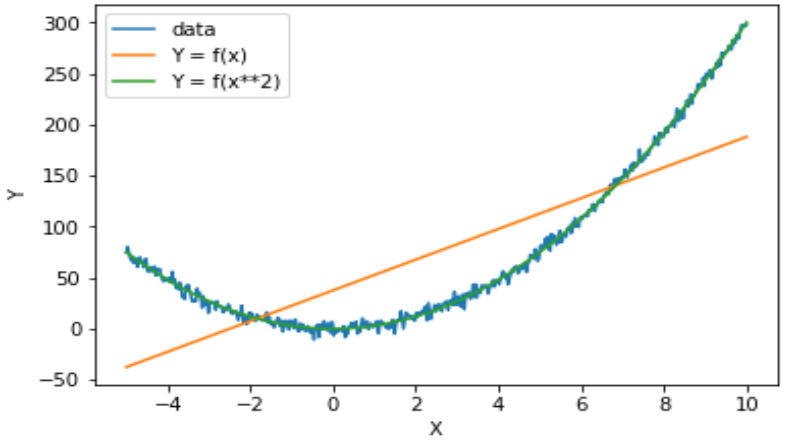
- 변수 분포 문제란 **일반화된 모델**을 학습하는데 어려움이 있는 분포를 가지는 변수가 있어, 일반화된 모델을 학습하지 못하는 문제로, 입문자가 가장 쉽게 무시하고 접근하기 어려워하는 문제



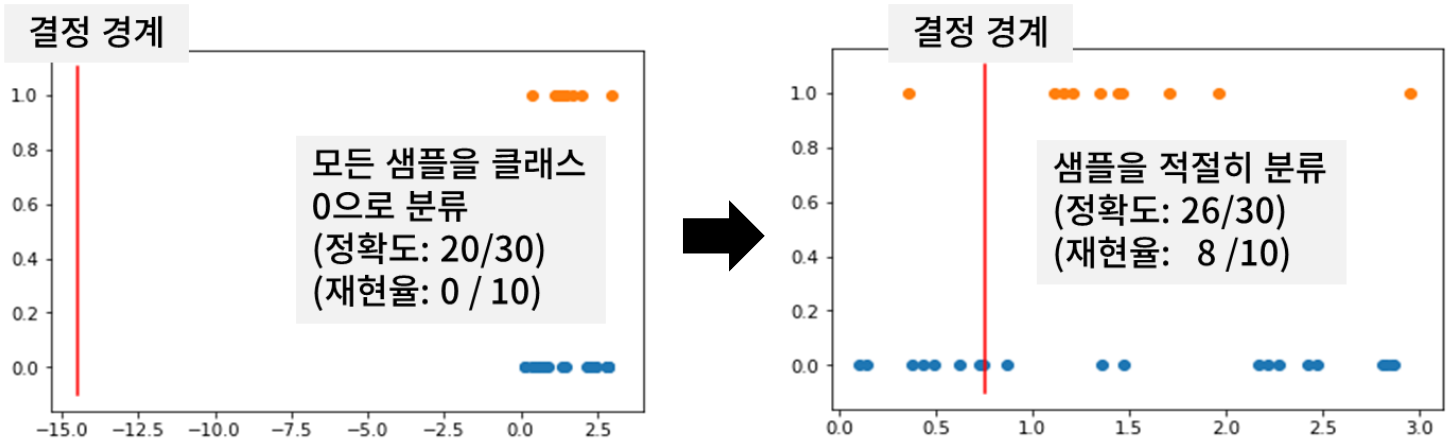
I 문제 정의

- 특징과 라벨 간 관계가 없거나 매우 약하다면, 어떠한 전처리 및 모델링을 하더라도 예측력이 높은 모델을 학습할 수 없음

- 그러나 **특징과 라벨 간 비선형 관계**가 존재한다면, **적절한 전처리**를 통해 **모델 성능을 크게 향상**시킬 수 있음



선형 회귀 모델 사례



로지스틱 회귀 모델 사례

- Tip. 대다수의 머신러닝 모델은 **선형식**을 포함함

I 해결 방안

- 가장 이상적인 해결 방안은 각 특징에 대해, 특징과 라벨 간 관계를 나타내는 그래프를 통해 적절한 특징 변환을 수행해야 함

- 하지만 특징 개수가 많고, 다른 특징에 의한 영향도 존재하는 등 그래프를 통해 적절한 변환 방법을 선택하는 것은 쉽지 않아, 다양한 변환 방법을 사용하여 특징을 생성한 뒤 특징 선택을 수행해야 함

Chapter. 19

이상적인 분포를 만들순 없을까. 변수 분포 문제

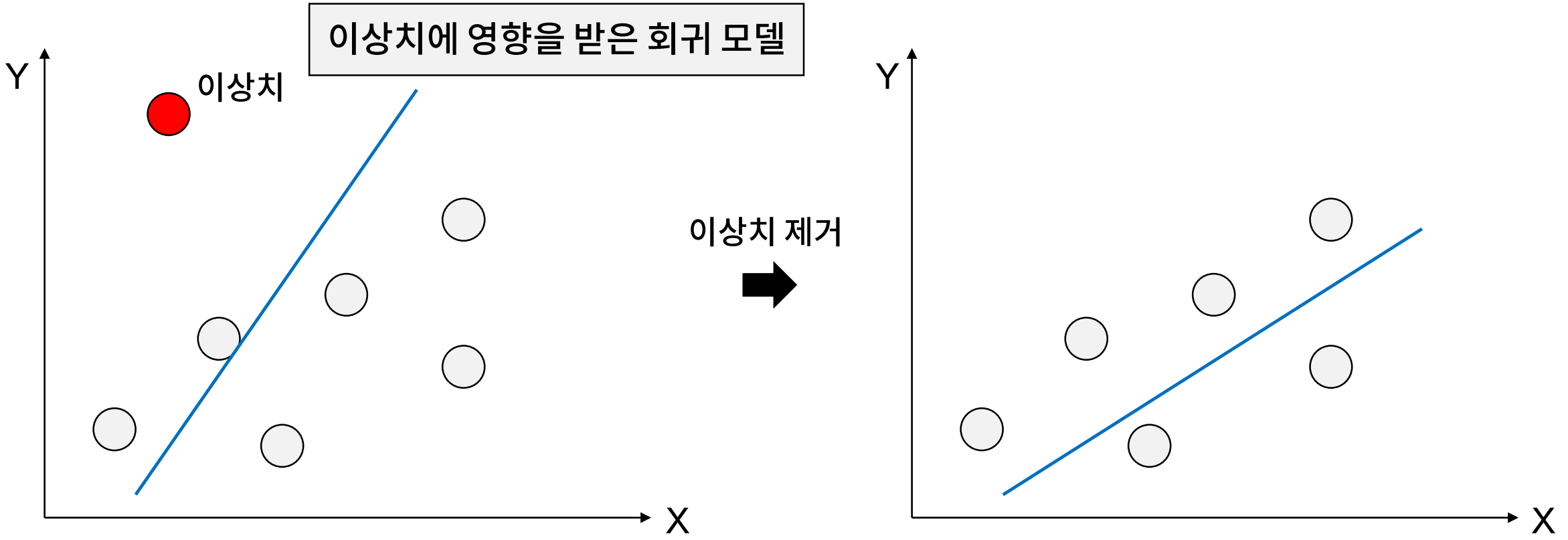
이상치 제거

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

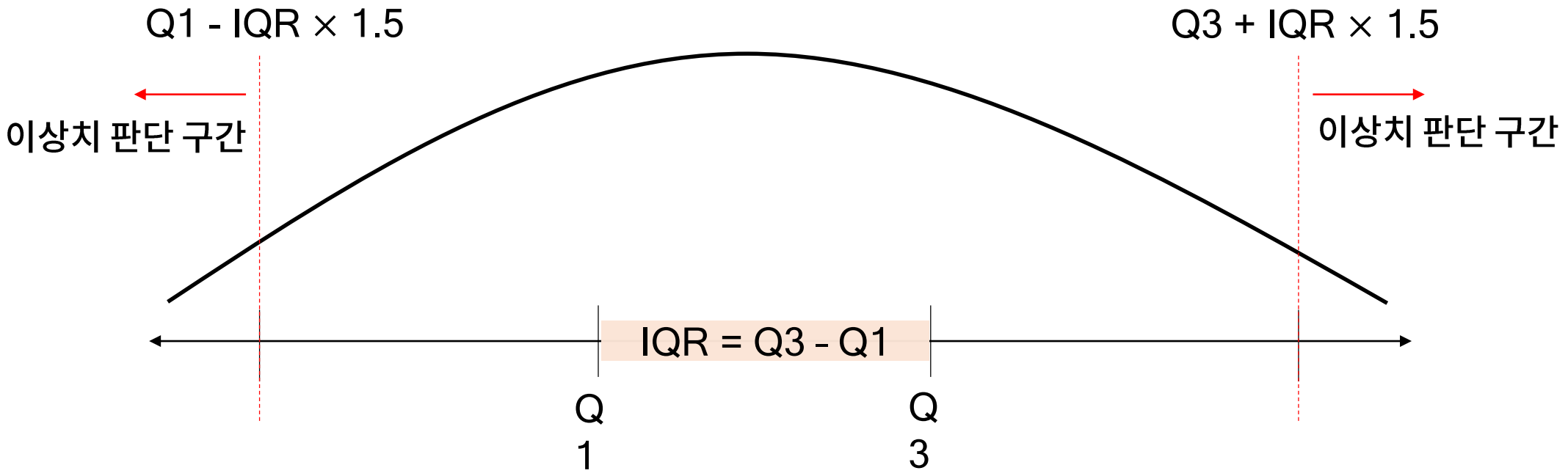
I 문제 정의 및 해결 방안

- 변수 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값으로, **일반화된 모델을 생성하는데 악영향**을 끼치는 값으로 **이상치를 포함하는 레코드를 제거**하는 방법으로 이상치를 제거함 (절대 추정의 대상이 아님에 주의)



I 이상치 판단 방법 1. IQR 규칙 활용

- 변수별로 IQR 규칙을 만족하지 않는 샘플들을 판단하여 삭제하는 방법



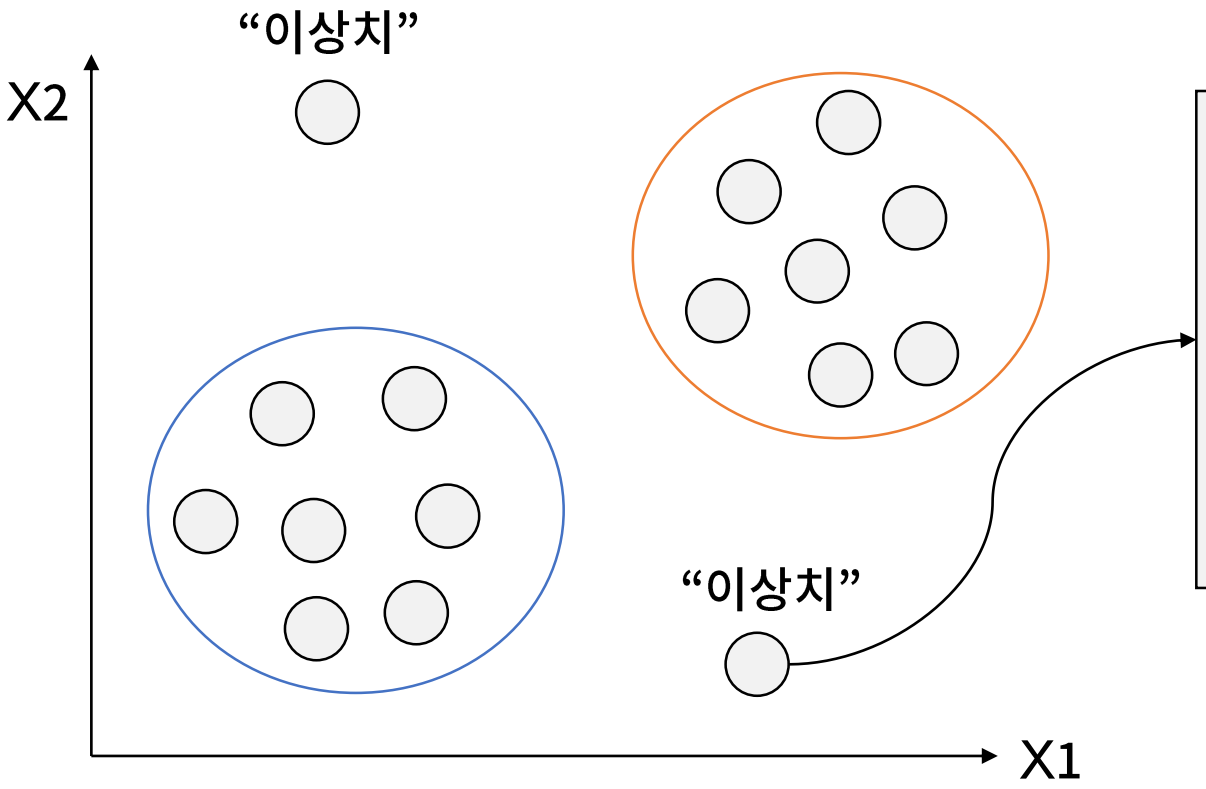
- 직관적이고 사용이 간편하다는 장점이 있지만, 단일 변수로 이상치를 판단하기 어려운 경우가 있다는 문제가 있음 (이전 페이지의 그림에서 표시된 이상치의 X값은 이상치라고 보기 힘든 구간에 있었음에 주목)

numpy.quantile

- Array의 q번째 quantile을 구하는 함수
- 주요 입력
  - a: input array (list, ndarray, array 등)
  - q: quantile (0과 1사이)

I 이상치 판단 방법 2. 밀도 기반 군집화 수행

- DBSCAN 등의 밀도 기반 군집화 기법은 군집에 속하지 않은 샘플을 이상치라고 간주하므로, 밀도 기반 군집화 결과를 활용하여 이상치를 판단할 수 있음



DBSCAN은 중심점과 경계점이 아닌 모든 샘플을 이상치라고 판단함

- 중심점: 설정된 반경 (eps) 안에 들어오는 샘플 수가 기준치 (min_samples) 이상인 샘플
- 경계점: 중심점의 이웃에 속하는 샘플

- 다만 DBSCAN 등의 밀도 기반 군집화 모델의 파라미터 튜닝이 쉽지 않다는 단점이 있음

sklearn.cluster.DBSCAN

- DBSCAN 군집화를 수행하는 인스턴스를 생성하는 함수
- 주요 입력
  - eps: 이웃이라 판단하는 반경
  - min_samples: 중심점이라 판단하기 위해, eps 내에 들어와야 하는 최소 샘플 수
  - metric: 사용하는 거리 척도
- 주요 attribute
  - .labels_: 각 샘플이 속한 군집 정보 (-1: 이상치)

Chapter. 19

이상적인 분포를 만들순 없을까. 변수 분포 문제

특징 간 상관성 제거

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 문제 정의

- 회귀 모델, 신경망, SVM과 같이 $wx + b$ 형태의 선형식이 모델에 포함되는 경우, 특징 간 상관성이 높으면 **강건한 파라미터 추정이 어려움** (즉, 추정할 때마다 결과가 달라질 수 있음)

x_1과 x_2를 이용하여 y를 예측하는 회귀 모델

- $y = 2x_1$라는 관계를 알고 있음
- $x_2 = x_1$라는 선형 관계가 존재함

→

$$y = w_1 x_1 + w_2 x_2$$

$$y = w_1 x_1 + w_2 x_1$$

$$y = (w_1 + w_2) x_1$$

$$\Leftrightarrow w_1 + w_2 = 2$$ **(무수히 많은 해)**

- 트리 계열의 모델은 사실 특징 간 상관성이 높다고 해서 모델 예측 성능에 영향을 받지 않지만, 상관성이 높은 변수 중 소수만 모델에 포함되기 때문에 **설명력에 크게 영향**을 받을 수 있음

I 해결 방법 (1) VIF 활용

- Variance inflation factors (VIF)는 **한 특징을** 라벨로 간주하고, 해당 라벨을 예측하는데 **다른 특징을 사용한 회귀 모델이 높은 R^2**을 보이는 경우 해당 특징이 다른 특징과 상관성이 있다고 판단함

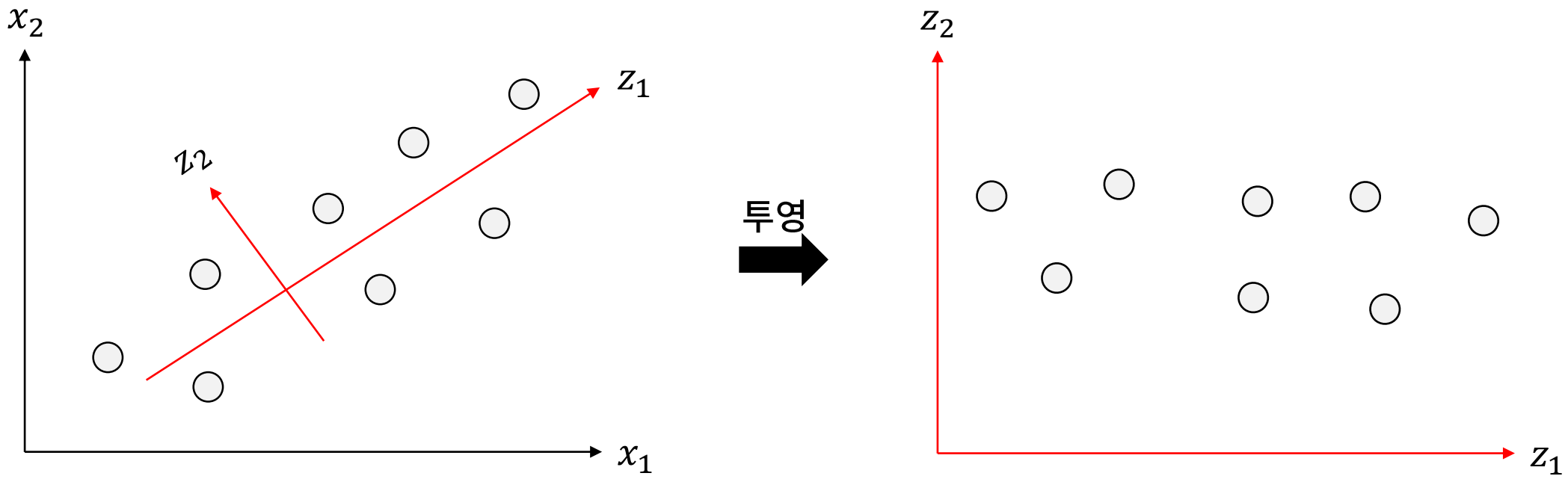
x_1	x_2	x_3	y

모델 개발 → $\hat{x}_3 = f_1(x_1, x_2)$ → VIF 계산 → $\frac{1}{1-R^2}$ → x_3 삭제

- VIF가 높은 순서대로 특징을 제거하거나, VIF가 10이상인 경우 주로 삭제함

l 해결 방법 (2) 주성분 분석

- 주성분 분석을 이용하여 **특징이 서로 직교**하도록 만들어 특징간 상관성을 줄이는 방법도 존재



- n차원의 데이터는 총 n개의 주성분이 존재하지만, 차원 축소 등을 위해 분산의 대부분을 설명하는 $m < n$ 주성분만 사용

sklearn.decomposition.PCA

- 주성분 분석을 수행하는 인스턴스를 생성하는 함수
- 주요 입력
  - n_components: 사용할 주성분 개수를 나타내며, 이 값은 기존 차원 수보다 작아야 함
- 주요 attribute
  - .explained_variance_ratio_: 각 주성분이 원 데이터의 분산을 설명하는 정도

Chapter. 19

이상적인 분포를 만들순 없을까. 변수 분포 문제

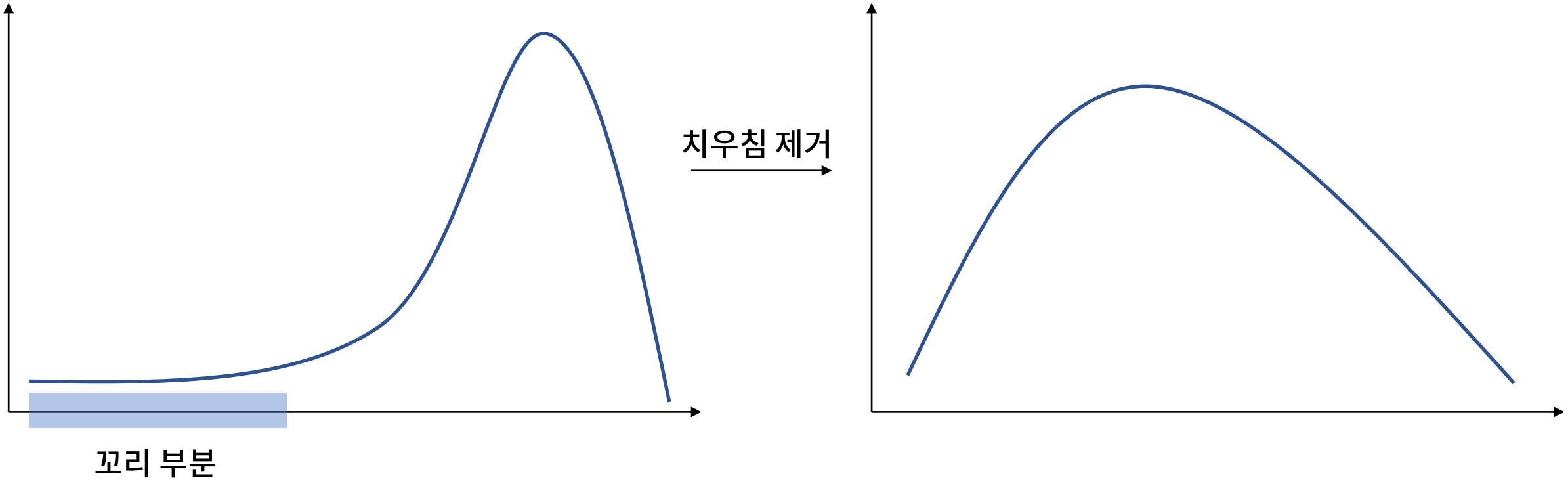
변수 치우침 제거

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

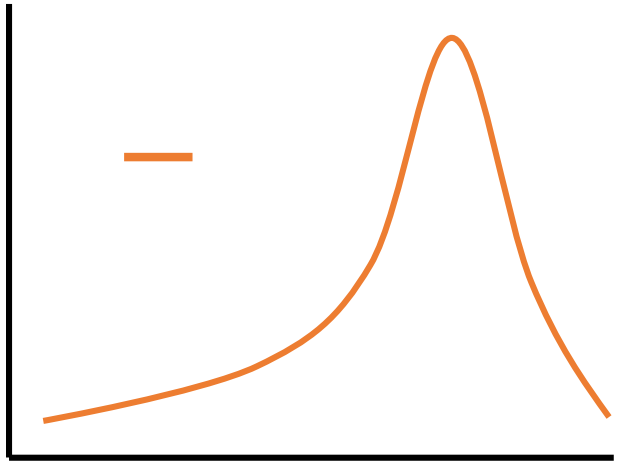
I 문제 정의

- 모델링에 가장 적합한 확률 분포는 정규 분포이나, 실제로 많은 변수가 특정 방향으로 **치우쳐 있음**
- 한 쪽으로 치우친 변수에서 **치우친 반대 방향의 값 (꼬리 부분)**들이 **이상치처럼 작용**할 수 있으므로, 이러한 치우침을 제거해야 함



I 탐색 방법: 왜도 (skewness)

- 변수 치우침을 확인하기 가장 적절한 척도로는 왜도(skewness)가 있음
- 왜도는 분포의 비대칭도를 나타내는 통계량으로, 왜도 값에 따른 분포는 다음과 같음



왜도 < 0

왜도 = 0

왜도 > 0

- 보통 **왜도의 절대값이 1.5 이상**이면 치우쳤다고 판단함

scipy.stats

- 다양한 확률 통계 관련 함수를 제공하는 모듈
- scipy.stats.mode: 최빈값을 구하는 함수
- scipy.stats.skew: 왜도를 구하는 함수
- scipy.stats.kurtosis: 첨도를 구하는 함수

I 해결 방안

- 변수 치우침을 해결하는 기본 아이디어는 **값 간 차이를 줄이는데** 있음
- 대표적인 처리 방법은 다음과 같음

$$\log(x - \min(x) + 1)$$

Log Transform

$$\sqrt{(x - \min(x))}$$

Square Root Transform

Chapter. 19

이상적인 분포를 만들순 없을까. 변수 분포 문제

스케일링

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 문제 정의

- 특징 간 **스케일이 달라서 발생하는 문제**로, 스케일이 큰 변수에 의해 혹은 스케일이 작은 변수에 의해 모델이 크게 영향을 받는 문제를 의미
  - ➢ 스케일이 큰 변수에 영향을 받는 모델: k-최근접 이웃
  - ➢ 스케일이 작은 변수에 영향을 받는 모델: 회귀모델, 서포트 벡터 머신, 신경망
  - ➢ 스케일에 영향을 받지 않는 모델: 나이브베이즈, 의사결정나무 (이진 분지에 한함)

I 해결 방법

- 스케일링을 사용하여 변수 간 스케일 차이를 줄이는 방법으로 해결할 수 있음

$$\frac{x-\mu}{\sigma}$$

Standard Scaling

$$\frac{x-\min(x)}{\max(x)-\min(x)}$$

Min-max Scaling

- 모델에 따른 스케일러 선택
  - ➢ Standard Scaler: 특징의 정규 분포를 가정하는 모델 (예: 회귀모델, 로지스틱회귀모델)
  - ➢ Min-Max Scaler: 특정 분포를 가정하지 않는 모델 (예: 신경망, k-최근접 이웃)

sklearn.preprocessing.MinMaxScaler & StandardScaler

- Min max scaling과 standard scaling을 수행하는 인스턴스를 생성하는 함수
- 주요 메서드
  - fit: 변수별 통계량을 계산하여 저장 (min max scaler: 최대값 및 최소값, standard scaler: 평균 및 표준편차)
  - transform: 변수별 통계량을 바탕으로 스케일링 수행
  - inverse_transform: 스케일링된 값을 다시 원래 값으로 변환

Chapter.

이상적인 분포를 만들순 없을까. 변수 분포 문제

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승