데이터 분석을 위한 기초

Kim Hye Kyung
topickim@naver.com

Ⅰ 데이터 분석을 위한 기초 수학

Ⅱ 데이터 표준화

Ⅲ 압축 기법(주성분 분석)

Ⅳ 참고 용어

데이터 분석을 위한 기초 수학

수학 용어 및 개념

용어	설명
대푯값	자료 전체의 특징을 대표적으로 나타내는 값
중앙값	변량을 작은 값에서 부터 크기 순으로 나열 할 때 중앙에 오는 값
최빈값	각 변량 중에서 가장 많이 나타나는 값
산포도	분산도라고도 함 대푯값을 중심으로 자료가 흩어져 있는 정도를 하나의 수로 나타낸 값으로 분산, 표준 편차등이 있음
도수	변량의 개수

수학 기호와 발음

그리스 문자			
Αα	알파	Νν	뉴
Ββ	베타	Ξξ	크시
Γγ	감마	Οο	오미크론
Δδ	델타	Ππ	파이
Εε	엡실론	Ρρ	로
Ζζ	제타	Σσς	시그마
Ηη	에타	Ττ	타우
Θθ	세타	Υυ	입실론
Ιι	요타	Φφ	피
Κκ	카파	Χχ	키
Λλ	람다	Ψψ	프시
Μμ	뮤	Ωω	오메가

기타 문자			
Ϝ	디감마	ϛ	스티그마
Ϸ	헤타	Ϻ	산
ϙ	코파	Ϡ	삼피
Ϸ	쇼		

https://ko.wikipedia.org/wiki

대문자	소문자	영어 이름	한글표기	의미
∑	σ	sigma	시그마	수학에서 수열의 합
Δ	δ	delta	델타	수학에서 값의 작은 차이를 나타내는 기호
M	μ	mu	뮤	모평균

모집단과 표본집단

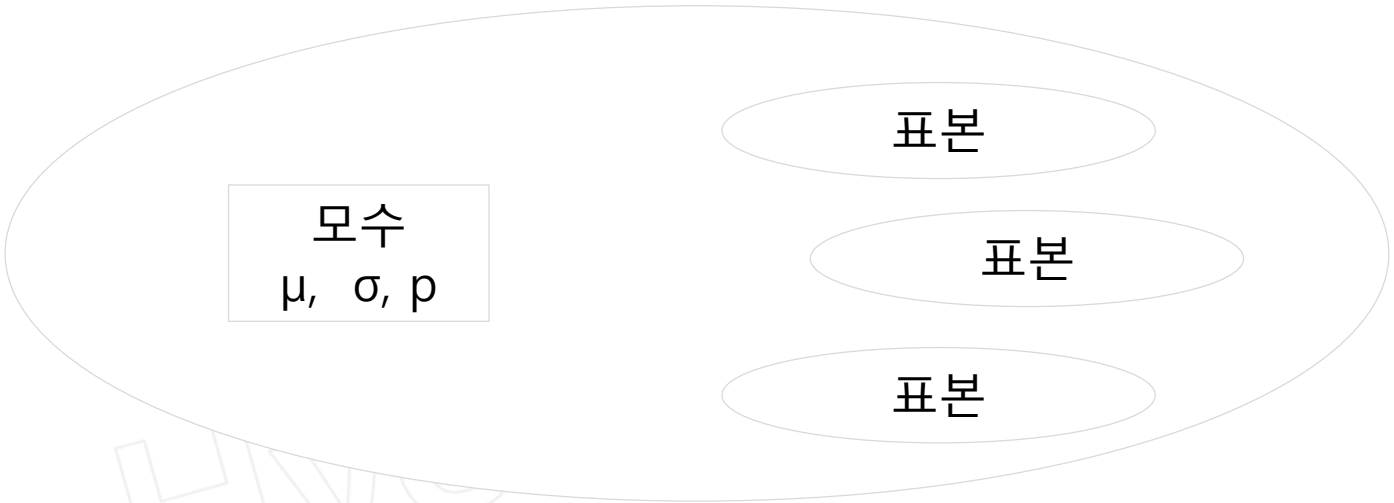
용어	설 명
모집단(population)	어떤 집단의 전체 데이터 의미
표본집단(sample)	전체 데이터 중에서 어떤 방식으로든 추출된 부분 집합을 의미

모집단과 표본집단

- 일반적으로 모집단의 정보를 구하기 어렵기 때문에 표본집단을 추출해 계산함으로써 모집단의 정보를 추정

- 예시
  - 우리나라 회사원의 연봉 평균 구하기
    - 회사원 전체를 조사 해야 하지만 현실적으론 너무나 어려움
    - 통계적으로 유효한 인원을 추출해 평균 계산 후 전체 평균의 값을 추정
    - 모집단 : 우리나라 회사원 전체
    - 표본집단 : 추출된 유효한 인원, 가령 5000명 또는 10000명등이 표본집단

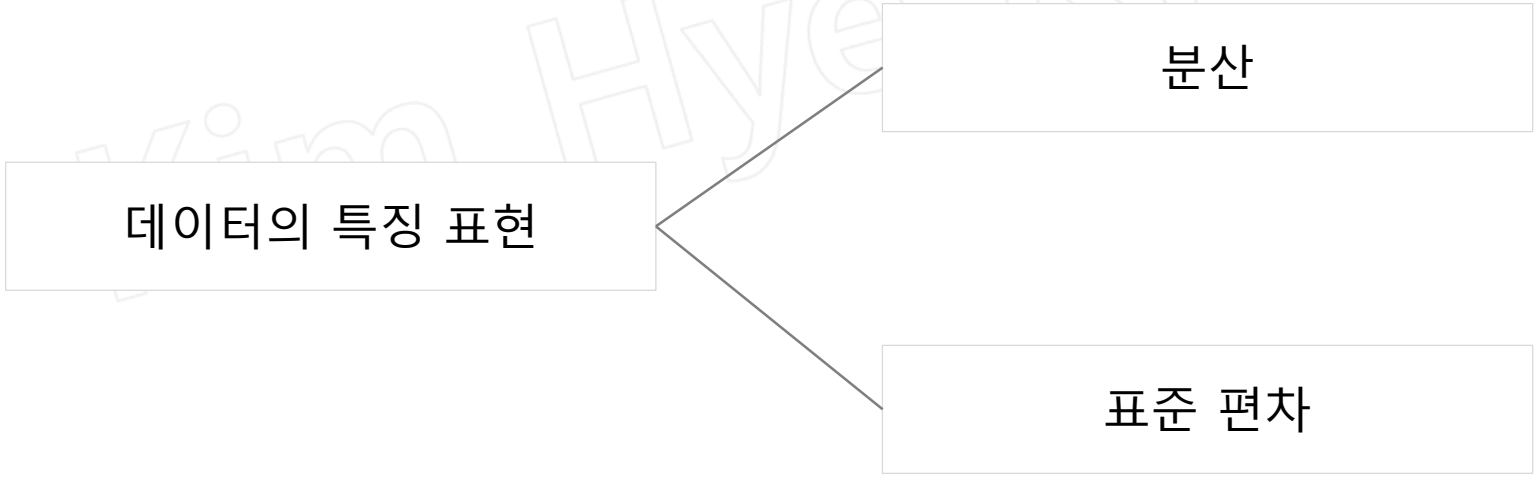
기초 통계 용어



- Population(모집단) : 관심 대상의 모든 원소의 집합
- Parameter(모수) : 모집단의 통계값(모평균 μ, 모표준편차 σ, 모비율 p)
- Sample(표본) : 모집단에서 추출된 부분 집합
- Statistic(통계량) : 표본의 통계값(표본의 평균 $\bar{x}$, 표본 표준편차 S)
- Standard deviation(표준편차) : 분산의 제곱근, 통계 집단의 변수가 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 대표적인 분포도 지수
- Sampling distribution of mean(평균의 표집분포) : 같은 모집단에서 n크기의 표본을 무한 반복하여 뽑아서 추정한 표본 평균값의 분포
- Standard error(표준오차) : 표본 평균이 모평균과 얼마나 퍼져 있는지를 나타내는 표준편차 추정치

분산(Variance)과 표준 편차(Standard Deviation)

- 데이터 특징을 표현하는 지표
- 데이터의 평균을 기준으로 어느 정도 흐트러져 있는지를 알려주는 지표
- 평균으로부터 먼 곳까지 데이터가 퍼져 있다면 분산과 표준편차의 값이 커짐
- 대부분의 데이터가 평균 근처에 위치 한다면 분산과 표준편차의 값이 작아짐



표준 오차

- 표본 평균이 모평균과 얼마나 퍼져 있는지를 나타내는 표준편차 추정치
  - 평균을 정의한 식

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$n\bar{x} = x_1 + x_2 + \ldots + x_n$$

$$0 = x_1 + x_2 + \ldots + x_n - n\bar{x}$$

$$0 = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x})$$

분산과 표준편차

- 표준을 정의 한 식을 이용한 표준 오차 이해하기

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$n\bar{x} = x_1 + x_2 + \ldots + x_n$$

$$0 = x_1 + x_2 + \ldots + x_n - n\bar{x}$$

$$0 = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x})$$

- 각 데이터와 평균과의 차이의 모든 합 = 0
- 분산과 표준 편차는 평균으로 부터 데이터가 흐트러져 있는 정도를 표현
- 문제 발생?
  - 모든 차이의 합이 0이라면 계산 불가

분산과 표준편차

- 표준을 정의한 식을 이용한 표준 오차(standard error, Se) 이해하기
- 문제
  - 모든 차이의 합이 0이라면 평균으로 부터 데이터가 흐트러져 있는 정도를 표현하는 분산과 표준 편차에 대한 계산 불가
- 각 데이터와 평균과의 차이의 정도인 표준 오차를 구하기 위한 해결책
  - 절대값과 제곱 이용하기
  - 절대값을 이용한 평균 오차(standard error) 경우

$$\text{Se} = \sum_{i=0}^{n} |x_i - \bar{x}| = |x_1 - \bar{x}| + |x_2 - \bar{x}| + |x_3 - \bar{x}| + \ldots + |x_n - \bar{x}|$$

  - 제곱을 이용한 표준 오차

$$\text{Se} = \sum_{i=0}^{n} (x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$$

분산과 표준편차

- 데이터 마이닝 기법에서는 절대값보다 제곱을 이용한 표준 오차를 사용
- why? 절대값과 제곱을 이용한 표준 오차 중에서 제곱을 사용하는 것이 데이터의 흐트러짐을 조금 더 잘 표현

예시

두 데이터 비교

x = {-2, -2, 2, 2}
y = {-3, -1, 0, 4}

x, y의 평균값 = 0

절대값을 이용해 x와 y의 오차 계산하기

Se_x = |-2-0| + |-2-0| + |2-0| + |2-0| = 8
Se_y = |-3-0| + |-1-0| + |0-0| + |4-0| = 8

제곱을 이용해 x와 y의 오차 계산하기

$Se_x = (-2-0)^2 + (-2-0)^2 + (2-0)^2 + (-2-0)^2 = 16$
$Se_y = (-3-0)^2 + (-1-0)^2 + (0-0)^2 + (4-0)^2 = 26$

분산과 표준편차

- 절대값을 이용한 오차와 제곱을 이용한 오차 비교

데이터	절대값으로 계산한 오차	제곱으로 계산한 오차
x	8	16
y	8	26

- x 와 y 의 데이터가 흐트러진 범위 비교

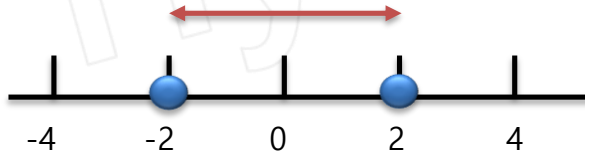
예시

두 데이터 비교

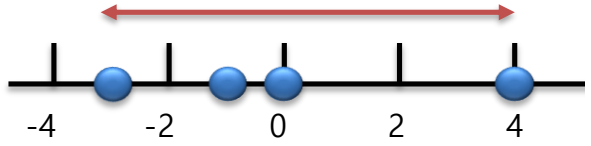
x = {-2, -2, 2, 2}

y = {-3, -1, 0, 4}

x, y의 평균값 = 0



x의 범위



y의 범위

평균 0으로 부터 x보다는 y의 데이터들이 넓게 퍼져 있음

즉 제곱으로 계산한 오차가 데이터의 퍼진 정도를 좀더 자세하게 표현

따라서 분산과 표준편차도 제곱으로 계산한 오차를 사용해서 계산

분산(Variance)과 표준편차(Standard Deviation)

- 분산(variance)
  - 편차 제곱의 평균
  - 편차 제곱의 총 합을 변량의 개수로 나누기
  - 평균을 제곱했기 때문에 단위를 안 씀

분산(Variance)과 표준편차(Standard Deviation)

- 표준 편차(standard deviation)
  - 루트 분산 즉 분산에 제곱근(루트)을 취한 것이 표준 편차
  - 원래의 데이터 단위의 길이(m)였지만 분산은 제곱을 했기 때문에 넓이(m2)로 단위가 변한 것으로 간주
  - 이렇게 바뀐 단위를 원래 데이터의 단위로 되돌리기 위해서는 분산에 제곱근(루트)을 취하면 됨
  - 표준 편차는 원래 데이터와 단위가 같아서 데이터의 흐트러진 정도를 더 쉽게 직관적으로 이해 할 수 있음
  - 분산의 양의 제곱근
  - 표준편차 구하는 순서
    - 평균 -> 편차 -> 분산 -> 표준 편차
  - 다시 돌아왔기 때문에 단위 사용

분산(Variance)과 표준 편차(Standard Deviation)

- 분산과 표준편차는 편차들의 평균이라는 의미 따라서 데이터 개수로 나누어야 함
- 분모가 n 또는 n-1
- 이는 보유하고 있는 데이터가 모집단인지 표본집단인지에 의해 결정됨

분산

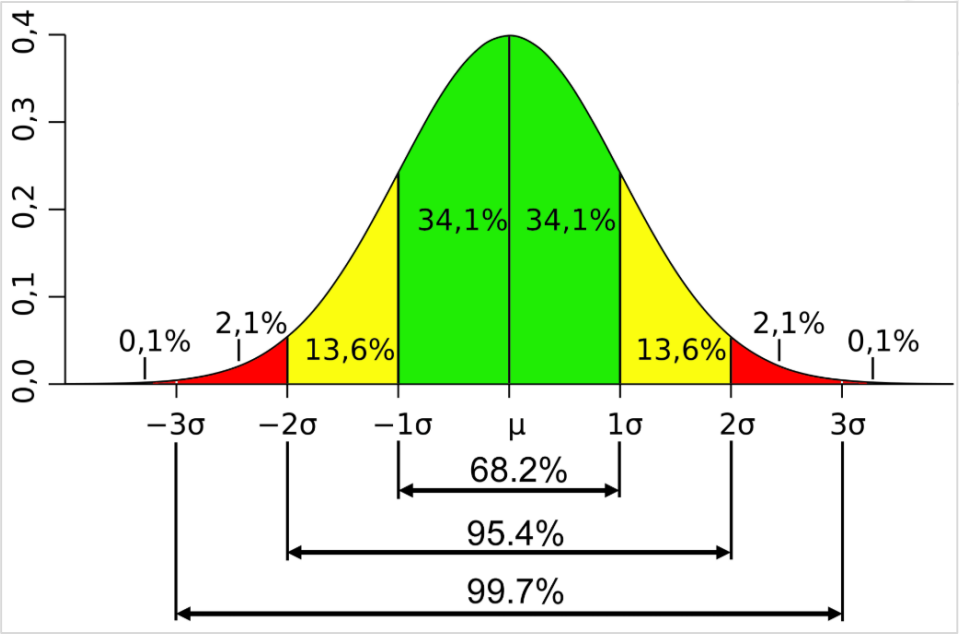
$$s^2 = \frac{\Sigma(y-\bar{y})^2}{n-1}$$

표준편차

$$s = \pm\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

정규 분포(Normal distribution)

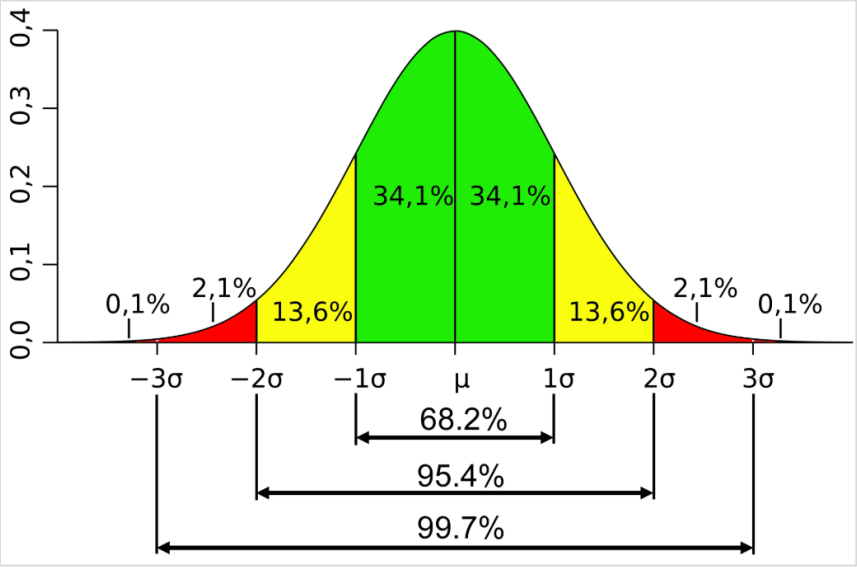
- 자연계의 많은 데이터를 히스토그램으로 표현시 데이터의 분포가 종 모양(bell-shape)처럼 평균을 중심으로 좌우 대칭된 분포를 의미
- 통계에서 가장 중요
- 유명한 수학자 가우스가 제안(가우스 분포[Gaussian Distribution]라고도 함)



중앙을 중심으로 50:50 으로 구분

같은 간격(표준 편차)를 기준으로 나뉘어져 있음

정규 분포(Normal distribution)

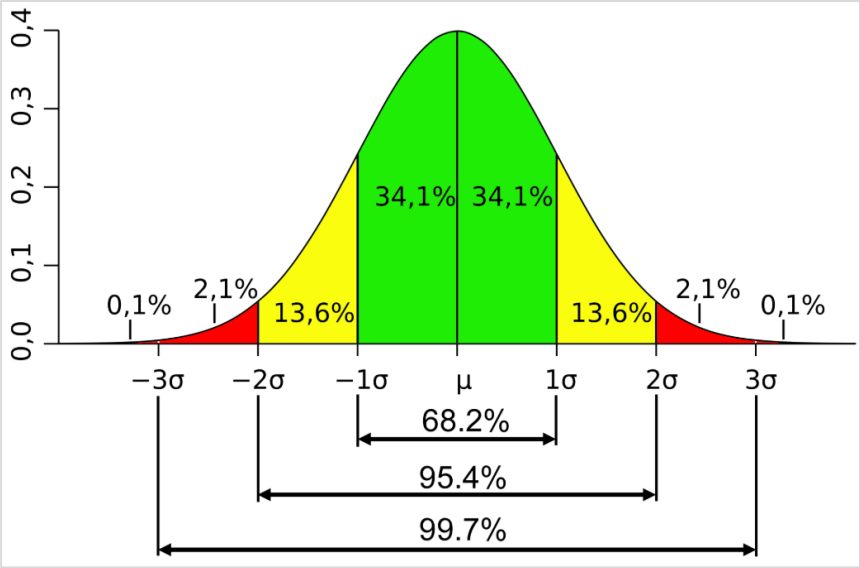


1. 총 데이터의 68.2%가 $\pm\sigma$ 범위 내에 존재
- 데이터가 $\pm\sigma$ 범위 내에 포함될 확률은 68.2%

2. 총 데이터의 95.4%의 데이터가 $\pm 2\sigma$ 범위 내에 존재
- 데이터가 $\pm 2\sigma$ 범위 내에 포함될 확률은 95.4%

3. 총 데이터의 99.7%의 데이터가 $\pm 3\sigma$ 범위 내에 존재
- 데이터가 $\pm 3\sigma$ 범위 내에 포함될 확률은 99.7%

정규 분포(Normal distribution)



예시

시그마 범위	시그마별 키 허용 범위	포함 확률
$\pm 1\sigma$	163.52 ≤ 키 ≤ 174.88	68.3%
$\pm 2\sigma$	157.84 ≤ 키 ≤ 180.56	95.5%
$\pm 3\sigma$	152.16 ≤ 키 ≤ 186.24	99.7%

우리나라의 40대가 100만 명이라 가정

99.7%인 997,000명이 152.16cm ~ 186.24cm 사이에 포함

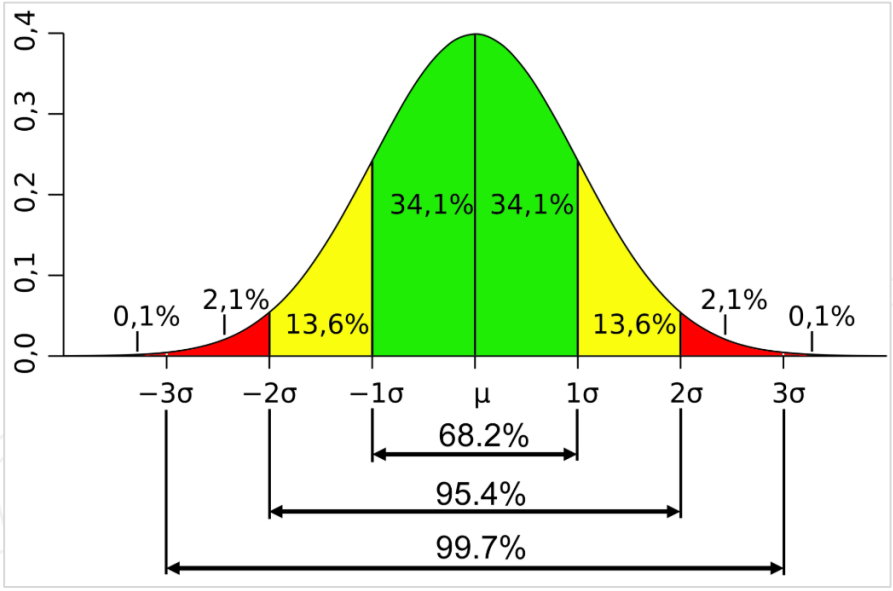
100만명 중 3000명만이 $\pm 3\sigma$ 에 포함되지 않음

통계적인 관점 : 정확히 3000명이 아닌 오차가 존재함

자신의 키가 152.16cm 미만이거나 186.24cm를 초과한다면 '통계적으로 $\pm 3\sigma$ 를 기준으로 정상 범위에 포함되지 않음

정규 분포(Normal distribution)

- 이상점(outlier) 검출



3시그마 규칙(3 Sigma Rule)

어떤 상품을 $\pm 3\sigma$로 품질을 관리한다는 품질 조사 결과 $\pm 3\sigma$ 밖에 상품의 데이터가 존재할 확률이 0.3%이기 때문에 이 범위를 벗어나는 상품은 불량으로 간주

데이터 표준화

데이터 표준화(Data Standardization)

- 데이터끼리 비교하기 위해서는 서로 같은 기준이나 척도가 적용되어야 함

홈런 20개를 친 타자

안타 100개를 친 타자

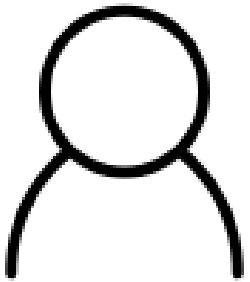
누가 잘 했나?

데이터 표준화(Data Standardization)

- 데이터끼리 비교하기 위해서는 서로 같은 기준이나 척도가 적용되어야 함

누가 잘 했나?

홈런 20개를 친 타자



안타 100개를 친 타자

어떤 선수는 득점 능력이 뛰어남

어떤 선수는 도루 능력이 뛰어남 ..

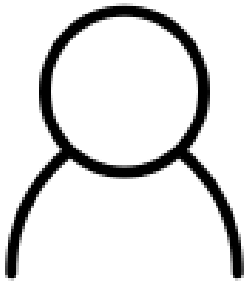
선수마다 뛰어난 능력이 다름 이처럼 서로 다른 득점이나 도루는 어떻게 비교하면 좋을까?

데이터 표준화(Data Standardization)

- 데이터끼리 비교하기 위해서는 서로 같은 기준이나 척도가 적용되어야 함

누가 잘 했나?

홈런 20개를 친 타자



안타 100개를 친 타자

해결책 :

데이터 표준화 - 서로 다른 기준이나 척도를 가진 데이터를 비교하기 위해서 사용하는 방법

데이터 표준화(Data Standardization)

$x = \{x_1 + x_2 + x_3 + ..., x_n\}$ 일 때의 표준화

$$z_i = \frac{(x_i - \bar{x})}{\sigma}$$

데이터를 데이터의 평균과의 오차를 계산 한 후 표준편차(σ)로 나누는 것

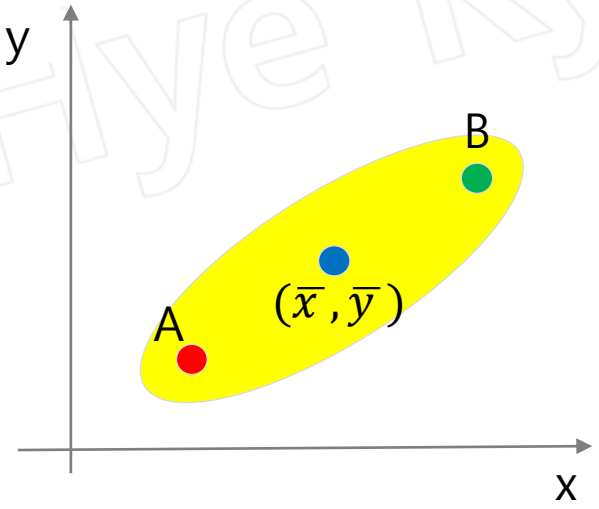
데이터 표준화 : 데이터를 표준편차로 나누어 단위 없앰

z-value 또는 Z score라고 하므로 z로 표시

* 모집단의 표준편차는 (σ, 시그마)로, 표본의 표준편차는 S(에스)로 표현 *

표준화 전후 데이터 분포

- 가정 사항
  - 데이터가 타원형으로 분포하고 있음
  - 평균보다 작은 값을 가지는 점 A와 평균보다 큰 값을 가지는 점 B가 있고, 데이터 분포의 중심은$(\bar{x},\ \bar{y})$

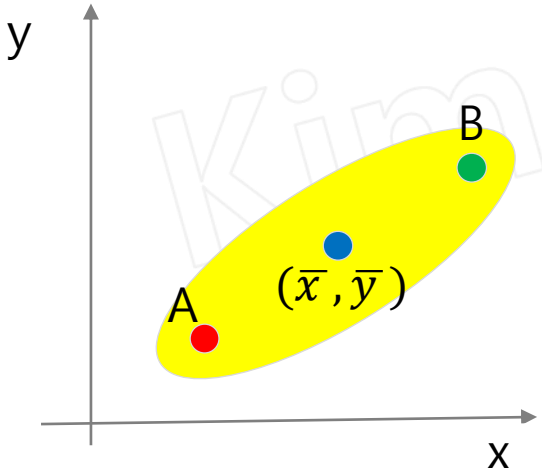


원래의 데이터 분포

표준화 전후 데이터 분포

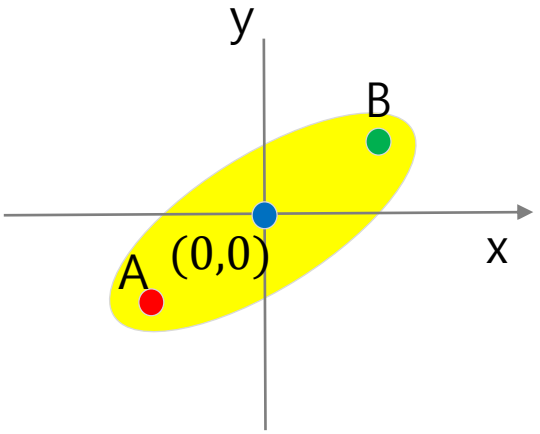
- 표준화 계산시
  - **데이터 분포의 중심을 (0, 0)으로 이동 시킴**
  - 각 데이터를 데이터의 평균과의 오차를 계산한 후 표준편차(σ)로 나눔

    (Z-value or Z score)



원래의 데이터 분포

$$z_i = \frac{(x_i - \bar{x})}{\sigma}$$



표준화 후 데이터 분포

표준화 전후 데이터 분포

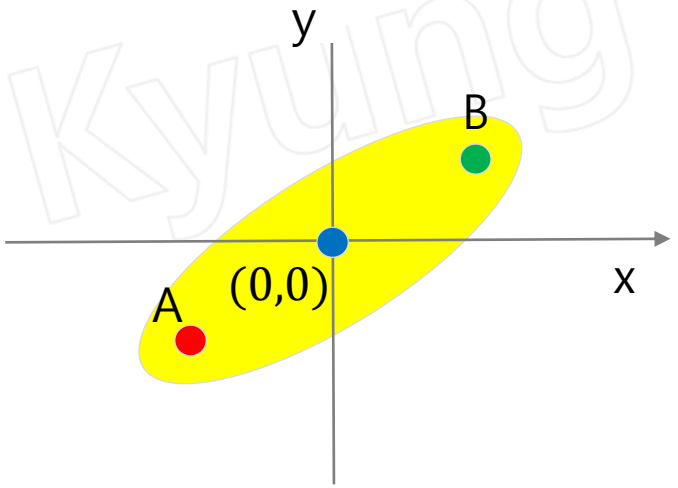
- 표준화 계산시
  - 데이터 분포의 중심은 (0, 0)으로 이동
  - 평균보다 작은 점 : 음수(-)
  - 평균보다 큰 점 : 양수(+)

$$0 = \frac{(\bar{x} - \bar{x})}{\sigma}$$

$$0 = \frac{(\bar{y} - \bar{y})}{\sigma}$$

표준화된 데이터 분포

1. 평균 $\bar{x} = 0$
2. 표준편차 $\sigma = 1$
3. 단위 없음



표준화 후 데이터 분포

공분산(covariance)과 상관계수(correlation)

- 공분산 개념
  - 두 개의 데이터 간의 관계 파악이 가능한 값
  - 예시
    - 습도가 높을수록 짜증 지수가 높아진다
    - 공부에 집중한 시간과 상위권 대학의 합격률은 비례한다

- 두 개의 확률 변수의 분포가 결합된 결합확률분포의 분산, 방향성은 나타내지만, 결합 정도에 대한 정보로서는 유용하지 않음

공분산(covariance)과 상관계수(correlation)

- 공분산이 0보다 큰 경우
  - 두 변수는 같은 방향으로 움직임
- 공분산이 0보다 작은 경우
  - 다른 방향으로 움직임
- 공분산이 0인 경우
  - 두 변수간에는 아무런 선형 관계가 없으며 두 변수는 서로 독립적인 관계

- 그러나! 두 변수가 독립적이라면 공분산은 0이 되지만, 공분산이 0이라고 해서 항상 독립적이라고 할 수 없음
- **두 데이터의 비례, 반비례 관계 여부를 알 수 있으나, 데이터 간의 관계가 어느 정도인지는 모름**
- **따라서 어느 정도인지 알기 위한 상관계수 필요**

공분산(covariance)과 상관계수(correlation)

- 상관계수(Correlation coefficient)
  - 두 가지 데이터가 어느 정도의 관계를 가지고 있는지를 알려주는 계수
    - 예시 : 기온이 올라가서 아이스크림 소비량이 증가
    - 결론 : 기온과 아이스크림 소비량의 상관관계가 높다는 의미

  - 두 개의 확률 변수 사이의 선형적 관계 정도를 나타내는 척도
  - 방향성과 선형적 결합 정도에 대한 정보를 모두 포함하고 있음
  - 두 변수의 공분산을 각 변수의 표준편차로 모두 나누어 구할 수 있으며, -1과 1사이에서 그 값이 결정됨
  - 공분산은 원래의 단위의 곱이 되기 때문에 경우에 따라서 이를 표준화할 필요가 있으며, 표준화한 결과가 상관계수가 됨

상관계수(correlation)

- 양과 음의 상관 관계

양의 상관 관계

x와 y가 비례 관계이고

x가 2가 증가할 경우 y도 2가 증가하는 것 처럼

증가 비율이 동일하다는 의미

음의 상관 관계

x와 y가 반비례 관계

x가 증가할 때 같은 비율로 y가 감소한다는 의미

미분과 편미분

- 미분
  - 데이터 마이닝에서 매우 중요
  - 어떤 점에서의 기울기(순간 변화량)
  - 최대값, 최소값을 구하기 위한 수단으로 자주 사용

- 편미분
  - 미분의 확장형
  - 어떤 함수가 여러 가지 변수를 가지고 있을 때 각 변수에 대해서 미분을 하는 방식 의미

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \frac{\mathrm{f(x} + \Delta x) - f(x)}{\Delta x}$$

미분과 편미분

- 미분
  - y = f(x)가 있을 때
  - 함수 f(x)를 x 에 대해서 미분한다는 의미
  - x가 아주 조금 변했을 때 y가 얼마나 변했는지 구하는 식

식 : A

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \frac{\mathrm{f}(\mathrm{x} + \Delta x) - f(x)}{\Delta x}$$

  - Δx : x의 변화량
  - $\lim_{\Delta x \to 0}$: x의 변화량 Δx 를 0에 한없이 가깝게 한다는 의미

    즉 **x 의 변화량을 거의 0에 가까울 정도로 작게 하라는 의미**

  - 일차 방정식의 기울기(기본적으로 식 A와 같은 의미)

식 : B

$$\text{기울기} = \frac{y\text{의 변화량}}{x\text{의 변화량}} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

미분과 편미분

- 편미분
  - 변수가 두 개 이상인 함수를 하나의 변수에 대해서 미분하는 것 의미
  - 미분과 같이 기울기 의미
  - 단, 변수가 여러 개이므로 특정 변수에 대해서 편미분 할 때에는 관계가 없는 다른 변수들을 상수로 취급해서 미분하면 됨

  - 편미분의 표기

$$\frac{\partial f(x)}{\partial x} \text{ 또는 } \frac{\partial}{\partial x} f(x)$$

  - 변수가 많을 경우 다른 변수들을 고정한 상태에서 그 변수에 대한 변화량을 의미

미분과 편미분

- 편미분의 중요성
  - 딥러닝의 기본 구조인 신경망에는 무수히 많은 파라미터 존재
  - 오차역전파(back propagation)이라는 학습 방법 이용
  - 오차역전파법은 신경망의 손실함수(loss function)가 최솟값을 가지도록 각 파라미터의 최적값을 찾는 학습 방법이며 경사 하강법(gradient decent)를 이용하여 구현
  - 편미분은 경사 하강법에서 각 파라미터들을 학습 시킬 때 이용하는 방법으로 알고리즘의 핵심

압축 기법(주성분 분석)

주성분 분석

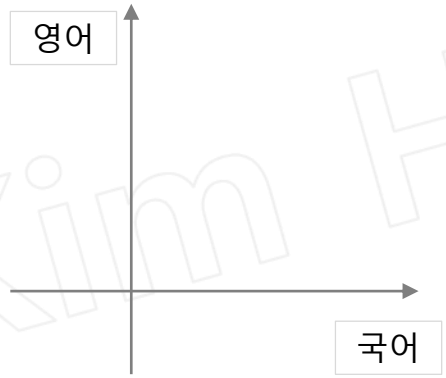
비지도 알고리즘 중 가장 광범위하게 사용되는 것 중 하나

차원 축소 알고리즘이지만 노이즈 필터링과 특징 추출 공학 등에서도
유용하게 사용되는 도구

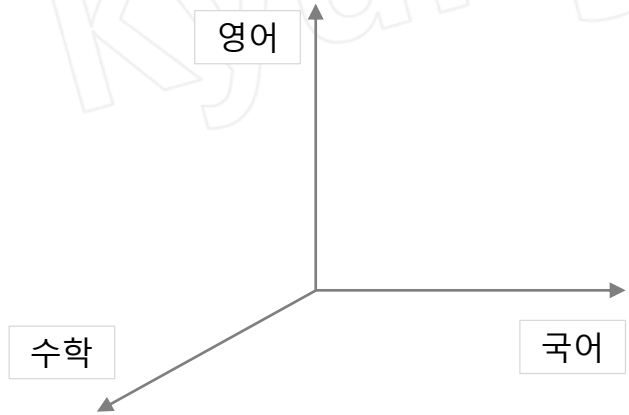
가장 중요하지 않은 주축을 따르는 정보는 삭제하고 가장 높은 분산을 갖는
데이터의 성분만 남김

주성분 분석의 개요

- 데이터 분석시 무수히 많은 변수를 가진 데이터들 만남
- 변수의 개수를 차원으로 표현하기도 함



2차원 그래프



3차원 그래프

주성분 분석의 필요성

- 국, 영, 수 처럼 1차원, 2차원, 3차원까지는 그래프로 시각화 가능
  - 데이터의 경향 파악이 가능

- 과목수가 많아질 경우 차원의 수가 4차원을 넘길 경우 그래프로 표현 불가
  - 그래프로 그리지 못 할 경우 데이터의 분포를 시각적으로 파악하기 힘듬

주성분 분석의 필요성

- 예시
  - 조건이 다섯 개인 변수가 하나 있을 경우 분석을 위해 필요한 최소한의 데이터
    - n=5인 경우
    - 변수가 두 개가 된다면 최소한의 데이터
      - n x n = n^2 = 5 x 5 = 25
    - 만일 변수가 세개가 된다면?
      - $n^3 = 125,\ n^4 = 625\ , n^5 = 3125\ , \ldots\ldots$
  - 데이터 차원이 커질수록 분석에 필요한 데이터 개수가 기하급수적으로 증가

주성분 분석의 필요성

- 데이터가 많아질수록 계산 비용도 기하급수적으로 증가
- 필요한 데이터를 구하기도 어려워 짐
  - 부족한 데이터로 적당히 분석 시도
  - 부족한 데이터를 사용해 분석하게 되면 모델을 구축하기 힘겨워 짐
    - 차원의 저주 라고 함

- 그렇다면 "차원의 저주 " 를 해소하기 위한 데이터 분석 방법은?

주성분 분석(Principal Component Analysis)

고차원의 데이터를 데이터의 손실을 최소화 하면서 저차원 데이터로 압축함으로써 차원을 축소하는 방법

주성분 분석

- 딥러닝 전의 기술
  - 예시
    - 사람의 얼굴을 인식시키기 위해서는 사진으로 부터 중요한 특징을 상당히 많이 추출
    - 특징이 많은 만큼 인식률은 매우 좋았으나 데이터 처리에 시간이 많이 걸림
  - 해결책
    - 주성분 분석을 사용하게 함으로써 특징을 1/10로 줄여서 인식률을 조금 떨어뜨리는 대신 처리 시간을 크게 줄임

주성분 분석

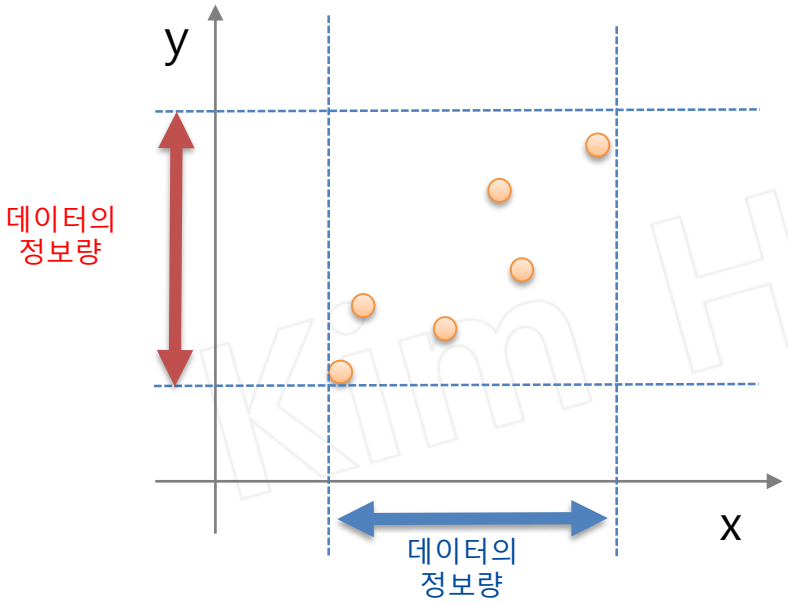
- 일반적인 차원 축소의 예시
  - 국어, 영어, 수학, 과학 등 교과 과목의 평균점수
  - 4개 차원을 과목 평균이라는 1개 차원으로 감소

$$\text{과목 평균} = \frac{(80+70+90+90)}{4} = 82.5 \text{ 점}$$

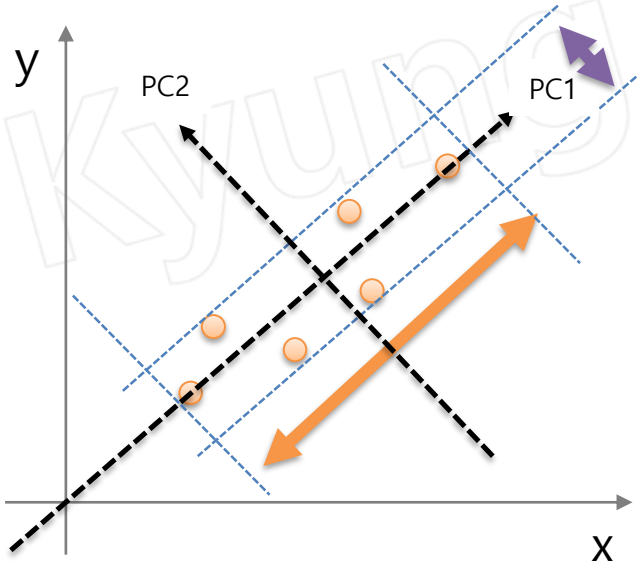
  - 축소된 점수를 반에서 몇 등, 전교에서 몇 등과 같이 활용하게 됨
  - 단, 주성분 분석은 위와 같이 차원을 축소하지 않고 다른 방법 사용

주성분 분석

- 주성분 분석의 차원 축소 방법 이해를 위한 예시
  - 데이터의 정보량을 데이터의 분산(데이터가 흐트러진 범위)이라 가정



x, y 축에서의 정보량



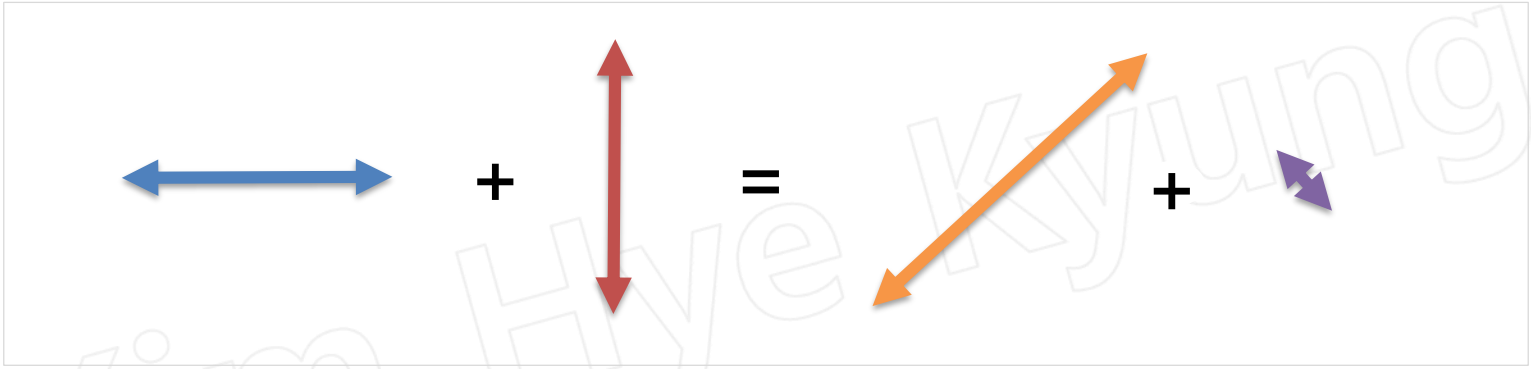
PC1, PC2축에서의 정보량

x, y 각 축에서 볼 수 있는 데이터의 정보량을 파란색, 빨간색 화살표로 표시

동알한 데이터에 임의의 새로운 축인 PC1(principal component)과 PC2를 만든 후 축에서 볼 수 있는 정보량 표현

주성분 분석

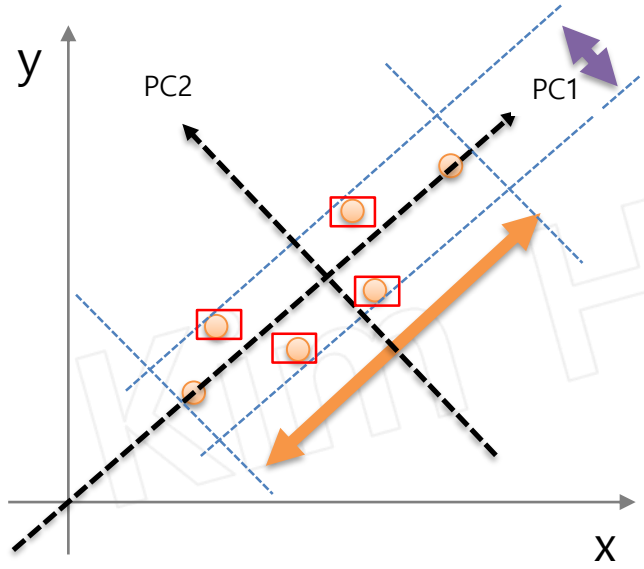
- 주성분 분석의 차원 축소 방법
  - 두 그래프의 데이터 정보량의 합



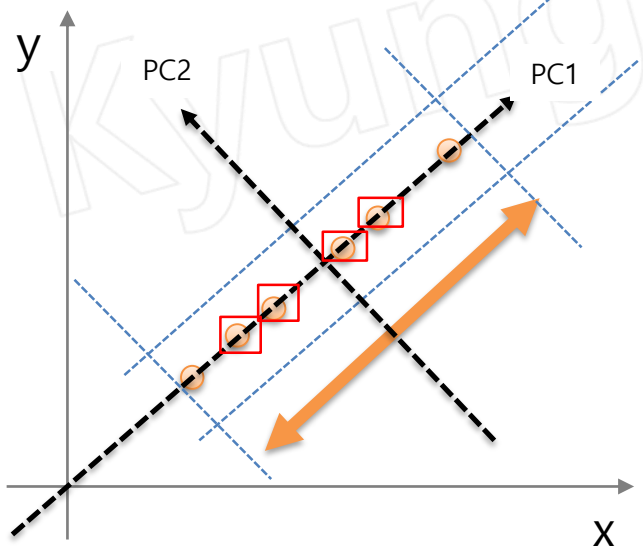
  - 두 그래프의 데이터가 같으므로 x, y 축에서 볼 수 있는 데이터의 정보량의 합과 PC1, PC2 축에서 볼 수 있는 데이터의 정보량의 합은 동일
  - 결론
    - 2차원 데이터를 1차원으로 차원 축소
    - 다차원 데이터의 정보를 가능한 손실 없이 저차원 공간에 압축하는 것

주성분 분석

- 주성분 분석의 차원 축소 방법 이해를 위한 예시
  - 데이터의 정보량을 데이터의 분산(데이터가 흐트러진 범위)이라 가정



PC1, PC2축에서의 정보량

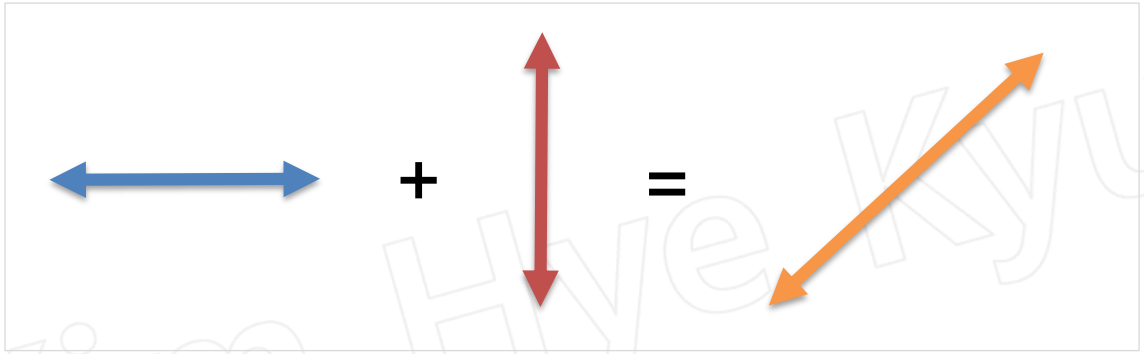


PC1, PC2축에서의 정보량

동알한 데이터에 임의의 새로운 축인 PC1(principal component)과 PC2를 만든 후
그 축에서 볼 수 있는 정보량 표현

주성분 분석

- 주성분 분석의 차원 축소 방법
  - 두 그래프의 데이터 정보량의 합



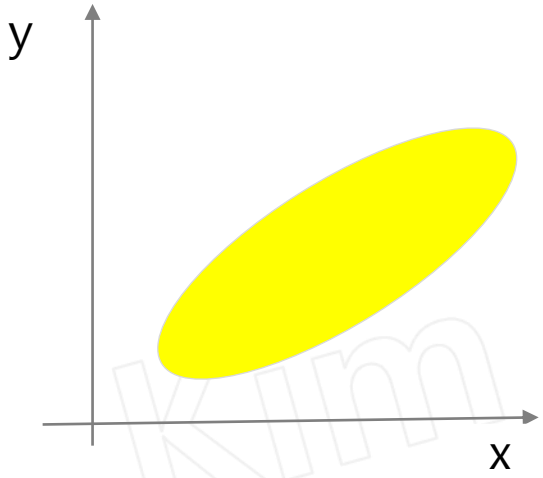
  - 두 그래프의 데이터가 같으므로 x, y 축에서 볼 수 있는 데이터의 정보량의 합과 PC1, PC2 축에서 볼 수 있는 데이터의 정보량의 합은 동일
  - 결론
    - 2차원 데이터를 1차원으로 차원 축소
    - 다차원 데이터의 정보를 가능한 손실 없이 저차원 공간에 압축하는 것

주성분 분석 진행 과정

- 데이터 표준화
  - 모든 변수들의 단위를 무효화 한 후 비교 가능
  - 데이터 분포가 평균은 0, 표준편차가 1인 데이터 분포로 변환
- 주성분 축 생성
  - 1-1. 첫 번째 축 구성
    - 분산이 가장 큰(데이터가 가장 넓게 퍼져있는) 방향을 구함
    - 그 방향으로 첫 번째 축을 구성
  - 1-2. 두 번째 축 구성
    - 첫 번째 축과 90도 직교하며, 분산이 두 번째로 큰 방향을 구함
    - 그 방향으로 두 번째 축을 구성
  - 1-3. 세 번째 축 구성
  - ……
- 구해진 새로운 공간으로 원래 데이터의 좌표를 이동시킴

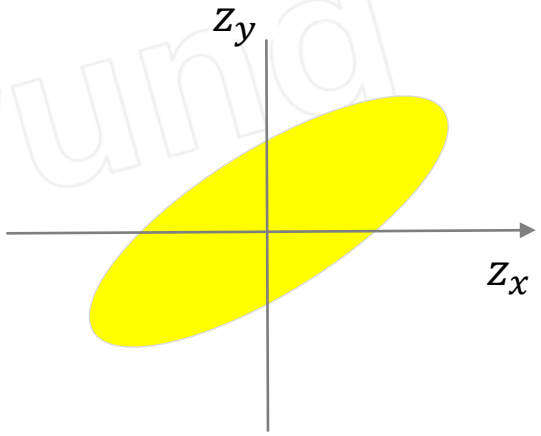
주성분 분석 진행 과정

- 1단계 – 데이터 표준화



원래의 데이터 분포

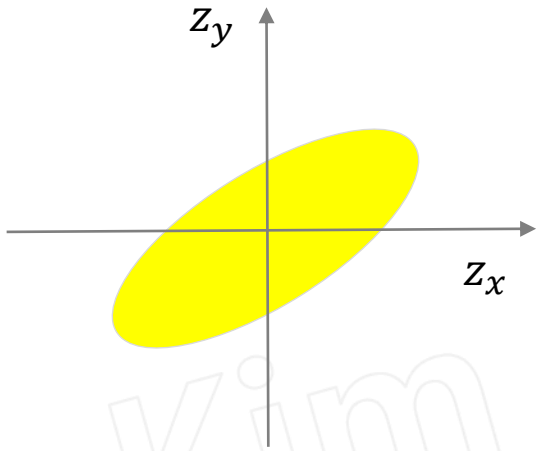
표준화



새로운 축 생성

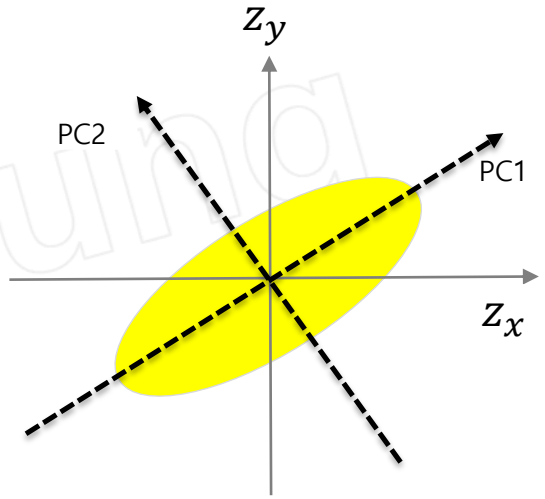
주성분 분석 진행 과정

- 2단계 – 주성분 축 생성



표준화 후 데이터 분포

주성분 생성

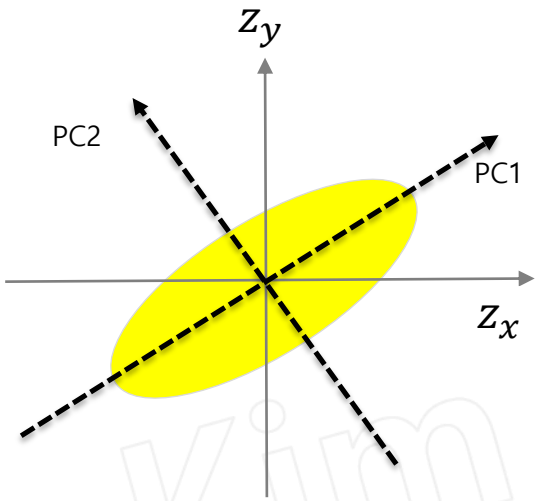


새로운 축 생성

PC1 분산이 가장 크고, PC2가 두번째로 큰 축
PC1 생성 후에 직교해야 하는 주성분 분석의 조건으로 인해
PC2가 새로운 축으로 구성됨

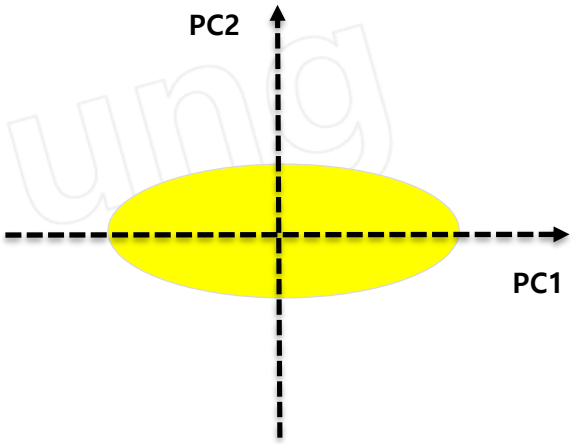
주성분 분석 진행 과정

- 3단계 – 새로운 축으로 좌표 변환



새로운 축 생성

좌표 변환



새로운 축으로 좌표 변환

주성분 분석 진행 과정[수학적 관점에서 보기]

- 데이터 표준화
  - 모든 변수들의 단위가 없어지면서 같이 비교 가능
  - 데이터 분포가 평균은 0, 표준편차가 1인 데이터 분포로 변환
- 주성분 축 생성
  - 1-1. 첫 번째 축 구성
    - 분산이 가장 큰(데이터가 가장 넓게 퍼져있는) 방향을 구함
    - 그 방향으로 첫 번째 축을 구성
  - 1-2. 두 번째 축 구성
    - 첫 번째 축과 90도 직교하며, 분산이 두 번째로 큰 방향을 구함
    - 그 방향으로 두 번째 축을 구성
  - 1-3. 세 번째 축 구성
  - ……
  - 구해진 새로운 공간으로 원래 데이터의 좌표를 이동시킴

수학적 관점에서 보기

데이터 표준화

상관행렬 구하기

상관행렬의 고윳값, 고유벡터 구하기

고유 벡터를 이용하여 표준화된 데이터를 주성분 공간으로 이동시키기

주성분 분석 진행 과정

- Scikit-Learn의 데이터 전처리
- 스케일링은 자료 집합에 적용되는 전처리 과정
  - 안정성 및 수렴 속도를 향상시킴
  - 모든 자료에 선형 변환을 적용하여 전체 자료의 분포를 평균 0, 분산 1이 되도록 만드는 과정
  - 스케일링은 자료의 오버플로우(overflow)나 언더플로우(underflow)를 방지
  - 독립 변수의 공분산 행렬의 조건수(condition number)를 감소시키는 최적화 과정

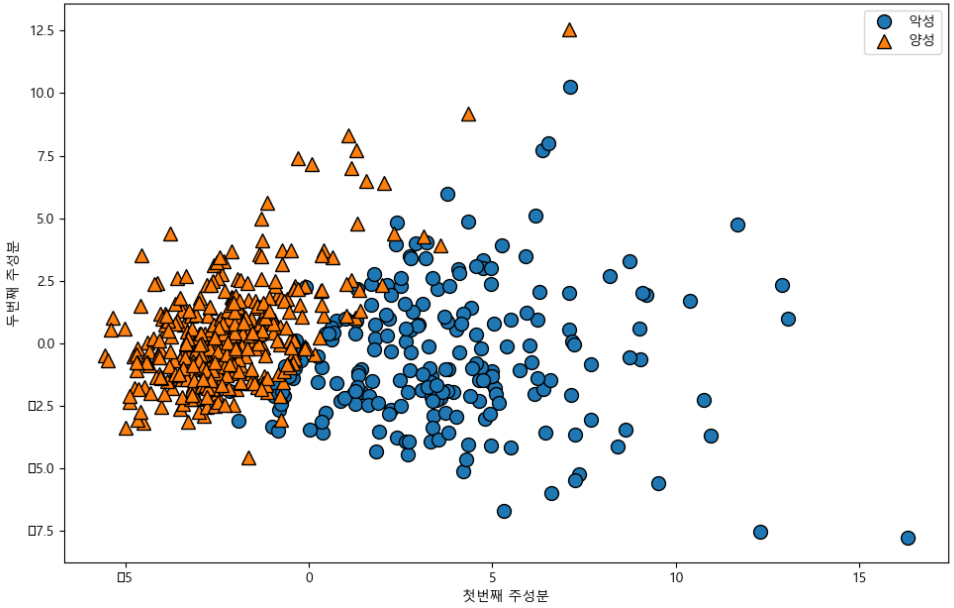
주성분 분석 진행 과정

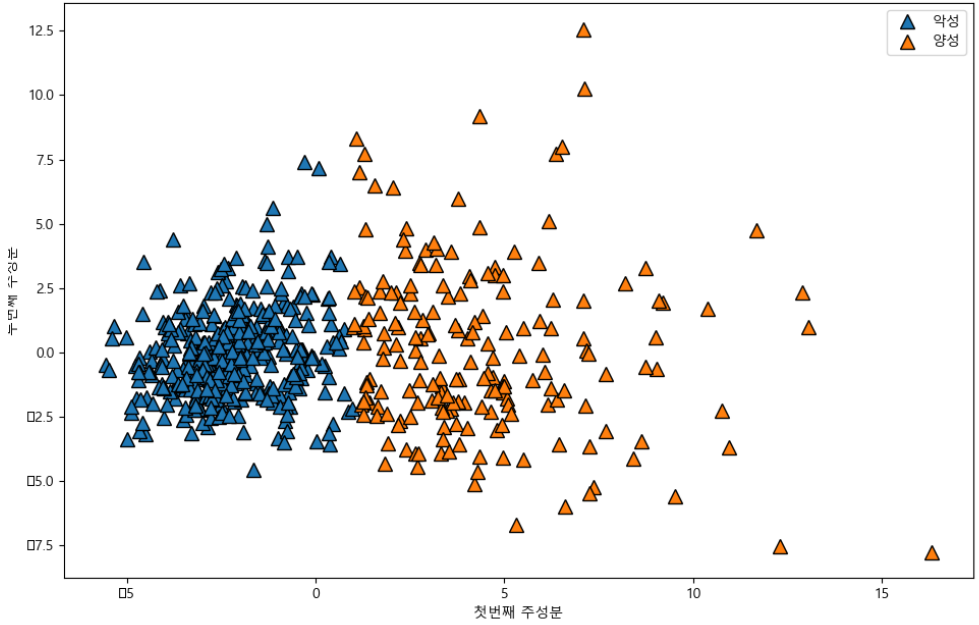
- 구현 방법
  - 클래스로 사용시 StandardScaler 클래스 이용
    - 클래스 객체 생성
    - fit() : 트레이닝 데이터를 사용하여 변환 계수 추정
    - transform() : 메서드를 사용하여 실제로 자료를 변환
    - 또는 fit_transform() 메서드를 사용하여 계수 추정과 자료 변환을 동시에 실행할 수도 있음

```
1  from sklearn.preprocessing import StandardScaler
2
3  scaler = StandardScaler()
4  scaler.fit(data)
5  X_scaled = scaler.transform(data)
```

주성분 분석

- 주성분 분석 후 주성분 분석으로 군집화 전 후의 산점도 비교





참고 용어

용어 정리 : 이산

- 수학 구조에 대해 연구하는 학문으로, 연속되지 않는 공간을 다룸

용어 정리 : 선형회귀(linear regression)

- 선형 회귀 회귀분석 기법
  - 종속 변수 y와 한 개 이상의 독립 변수 (또는 설명 변수) X와의 선형 상관 관계를 모델링하는 회귀분석 기법
  - 단순 선형 회귀 : 한 개의 설명 변수에 기반한 경우
  - 다중 선형 회귀 : 둘 이상의 설명 변수에 기반한 경우
  - 선형 모델 : 선형 예측 함수를 사용해 회귀식을 모델링하며, 알려지지 않은 파라미터는 데이터로부터 추정, 회귀식을 이라 함

- 선형 회귀의 여러 사용 사례
  - 값을 예측하는 것이 목적일 경우, 선형 회귀를 사용해 데이터에 적합한 예측 모형을 개발
  - 개발한 선형 회귀식을 사용해 y가 없는 x값에 대해 y를 예측하기 위해 사용할 수 있음

용어 정리 : 선형회귀(linear regression)

- 종속 변수 y와 이것과 연관된 독립 변수 X1, ..., Xp가 존재하는 경우에, 선형 회귀 분석을 사용해 Xj와 y의 관계를 정량화할 수 있음
- Xj는 y와 전혀 관계가 없을 수도 있고, 추가적인 정보를 제공하는 변수일 수도 있음

- 일반적으로 최소제곱법(least square method)을 사용해 선형 회귀 모델을 세움
  - 최소제곱법 외에 다른 기법으로도 선형 회귀 모델을 세울 수 있음
  - 손실 함수(loss fuction)를 최소화 하는 방식으로 선형 회귀 모델을 세울 수도 있음
  - 최소제곱법은 선형 회귀 모델 뿐 아니라, 비선형 회귀 모델에도 적용할 수 있음
  - 최소제곱법과 선형 회귀는 가깝게 연관되어 있지만, 그렇다고 해서 동의어는 아님

용어 : 식별과 클래스

- 식별
  - 무엇인가를 판단하거나 사물의 종류를 구분하는 것
  - 예시
    - 개와 고양이 이미지를 기반으로 구분시 개? 고양이? 인지 구분하는 것을 의미

- 클래스
  - 클래스 - 개와 고양이와 같은 식별 결과
  - N-클래스 식별 문제
    - 2-클래스 : 개와 고양이라는 두개의 종류
    - 4-클래스 : 개와 고양이 이외에도 말과 양이 추가될 경우

용어 : 특징량 또는 특징 벡터

- 가정 사항

두가지 입력 데이터를 비교해서 합격? 불합격? 판정을 사람에게 의존하지 않고도 판별 하고자 할 경우를 가정

1. 0 또는 1이 붙어 있는 레이블 y를 판별하고자 할 때, 클래스 수가 0과 1이라는 의미

2. x0과 x1 = 특징량 또는 특징 벡터

3. 레이블의 종류 = 클래스

입력 데이터 x0	입력 데이터 x1	레이블 y
값0	값1	1 (합격)
값0	값1	0 (불합격)
값0	값1	1 (합격)
값0	값1	0 (불합격)

용어 : 학습률(learning_rate)

- 학습률
  - Learning rate
  - 훈련 데이터를 기반으로 어느 정도 학습 할지를 나타내는 비율

- 하이퍼파라미터
  - Hyper Parameter
  - 학습률은 자동으로 구해지지 않으므로, 사림이 직접 설정해야 함
  - 이처럼 사람이 직접 설정해야 하는 매개 변수를 의미

- Epoch
  - 훈련 데이터 전체를 나타내는 단위
  - 가령 훈련 데이터가 20개일 경우 20개의 데이터를 모두 확인한 것을 1 에포크라고 함

참고 도서

- “데이터 분석을 떠받치는 수학” 손민규 지음[위키북스]