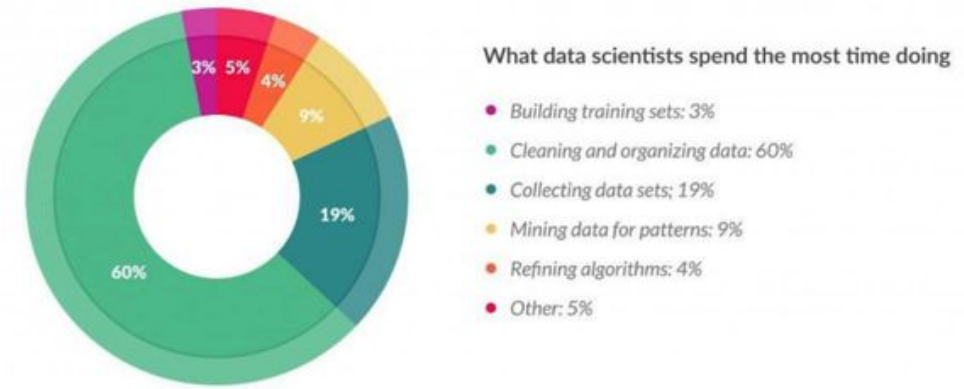
# Pre-Processing
# 2020년 봄 학기

2020년 02월 28일
https://www.github.com/KU-BIG/KUBIG_2019_Autumn

## 1. 전처리란?

"관측된 패턴에서 본질적인 정보를 쉽게 추출할 수 있도록 현재 주목하고자 하는 부분 패턴을 선정하거나 이 부분 패턴을 정형하여 불필요한 정보를 분리 제거하기 위한 예비적인 조작. 전처리에는 분할 잡음 제거, 정규화 등과 같은 처리가 포함된다. "  - 한국정보통신기술협회
모든 데이터 분석 프로젝트에서 데이터 전처리는 꼭 거쳐야 하는 과정이다. 대부분 분석가가 좋아하지 않는 과정이지만 분석 결과에 직접적인 영향을 미치는 과정이기 때문에 중요하게 다루어지고 있다. 한 설문조사에서는 분석가의 80% 시간을 데이터 수집 및 전처리에 사용한다고 한다.




전처리가 필요한 이유는 모든 데이터가 정형 데이터가 아니기 때문이다. 정형 데이터란 형식이 정해져 있는 데이터로, 데이터베이스에 이미  정리가 되어 있어 바로 통계적 분석에 이용될 수 있는 데이터를 의미한다. 그러나 실제로 우리가  대부분의 데이터는 비정형 데이터로 존재한다. 비정형 데이터란 정형 데이터와 달리 형식이 정해지지 않은 raw data이다. 사진, 음성, sns나 채팅 텍스트 등이 대표적인 비정형 데이터인데, 이러한 비정형 데이터의 형태는  분석 자료로 쓰일 수 없기 때문에 전처리 과정을 통해 이를 정제 및 가공하여 정형화 해야 한다. 따라서 머신러닝, 딥러닝 과정에서 전처리는 필수적인 단계라고 할 수 있다.

# 2. 데이터 분석을 위한 전처리

데이터 분석은 지표,표,그래프 작성 (시각화) ②지도학습 ③비지도학습으로 나눌 수 있는데, 각 분석의 준비에 필요한 각 전처리 과정의 목적과 내용은 조금씩 다르다.

## 2.1. 지표, 표, 그래프 작성용 전처리

지표의 계산, 표나 그래프로 쉽게 변환할 수 있는 데이터를 준비하기 위한 전처리 과정. 필요한 열이 모두 있고, 다루기 쉬운 단위로 집약된 행이 필요한 범위만큼 존재하는 데이터를 작성한다. 예를 들어 많은 가게를 운영하는 회사라면 관리 중인 각 지점의 월평균 매출이나 최대 매출을 알고 싶을 것이다. 이러한 집계 분석의 전처리로 가게마다 월 매출을 기록해두면 알고 싶은 정보를 쉽게 얻을 수 있다. 또한 매상에 관한 월별, 연령별 손님 수 등을 지표로 준비해두면 매상과 지표의 관계성을 쉽게 집계할 수 있다.

## 2.2. 지도학습용 전처리

지도학습 모델이 이용할 학습 데이터, 테스트 데이터, 적용 데이터 세 종류를 준비하는 것이 목적이다. 학습 데이터와 테스트 데이터는 적용 데이터에 예측 대상의 데이터를 추가하면 된다, 또한 학습 데이터와 테스트 데이터는 본질적으로 같은 데이터고 준비한 데이터를 분할해 학습 데이터와 테스트 데이터로 사용할 수 있다.

이 과정 역시 머신러닝 모델의 종류에 따라 머신러닝 모델이 다루기 쉬운 데이터로 변환하는 전처리가 필요하다. 예를 들어 입력값이 성별 및 연령이고, 출력값이 캠페인 반응 예측인 로지스틱 회귀 모델이 있다고 가정한다. 이 로지스틱 회귀 모델에 성별과 연령을 입력값으로 선택하면 '연령이 높을수록 반응한다'라든가 '여성이 더 반응한다'라는 식의 경향은 표현될 수 있다. 하지만 '연령이 30대에 가까운 사람일수록 잘 반응한다 ' 또는 '30대 여성이 특별히 더 잘 반응한다' 와 같은 경향은 표현되지 않는다. 이런 문제를 해결하기 위해서 연령을 범주형으로 변환하는 전처리나 연령과 성별을 조합한 새로운 범주형의 값으로 변환하는 전처리가 필요하다.

## 2.2. 지도학습용 전처리

비지도학습 모델이 이용할 설명 변수를 가진 데이터를 준비하는 전처리 과정이다. 충분한 열과 행이 필요할 뿐만 아니라 머신러닝 모델의 종류에 따라 다루기 쉬운 자료형으로 변환해야 한다. 예를 들어 성별 열을 문자열에서 범주형으로 변환하는 것으로 머신러닝 모델은 성별이 남성과 여성뿐이라는 것을 이해할 수 있다. 또한 클러스터링은 데이터 간 거리를 계산해야 하는데,

이용하는 열에 따라 수치의 크기가 크게 다르면 특정 열의 값이 결과에 과도한 영향을 미친다. (정규화 필요)

# 3. 전처리의 흐름

데이터의 무엇을 대상으로 변환하는가에 따라 전처리를 다음과 같이 두 종류로 분류할 수 있다.

## 3.1. 데이터 구조 대상의 전처리

데이터 구조 대상의 전처리는 여러 행에 걸쳐 데이터 전체를 아우르는 처리이다. 대부분 많은 양의 데이터를 다루며, 데이터 전처리 과정 중에서도 비교적 빠른 단계에서 구현된다. 특정 데이터를 뽑아내기도 하고 (추출), 데이터끼리 결합하거나 특정 규칙에 따라 여러 행을 하나의 행으로 묶기도 한다.

<예시>
-랜덤 샘플링으로 행을 추출
-매출 레코드와 상품 마스터를 상품 ID와 결합해 상품 정보를 첨부한 매출 레코드 생성
-지도학습 모델을 위한 학습 데이터와 테스트 데이터를 분할
-캠페인에 반응하는 데이터가 적을 때 오버샘플링 으로 데이터 늘리기

## 3.2. 데이터 내용 대상의 전처리

데이터 내용대상의 전처리는 데이터 행에 대한 처리이다. 데이터 구조 대상의 전처리와 달리 행 단위로 독립적인 처리가 가능한 소규모 데이터를 다룬다. 따라서 데이터 전처리 과정 중 후반에 주로 실행하며, 조건을 변경하며 반복 검증하는 일이 많다. 일별 데이터를 월별 데이터로 변환하거나 숫자 데이터의 열을 조합해 새로운 숫자 열을 생성한다.

<예시>
-연령을 나타내는 수치를 열개의 범주형 데이터 (연대로 표시)로 반환
-날짜 데이터를 요일 데이터로 변환
-두 지점의 경도/위도에서 각 지점의 거리를 계산

## 3.3 전처리 순서

분석 내용에 따라 필요한 전처리는 크게 달라진다. 물론 모든 경우에 적용되는 것은 아니나 전형적인 패턴은 존재한다.


출처: 데이터 전처리 대전 (모토하시 도모미쓰)

전형적인 패턴에서는 우선 대상 데이터를 추출해 데이터 양을 줄인다. 이때 마스터 데이터를 결합하거나 데이터 내용의 변환을 추출하기 전에 실행하면 실제로 이용하지 않는 데이터에 대해서도 전처리를 적용하게 되므로 불필요한 계산 비용이 커진다. (다만 결합 후에 값 조건 크기에따라 추출을 나중에 미룰수도 있다) 추출 후에는 결합 및 집약하여 분석에 필요한 데이터를 갖춘다. 그 후 데이터 내용 대상의 전처리를 진행한다. 집계나 BI 도구의 입력 정보로 이용할 경우엔 자료형 변환 등 최소한의 데이터 내용을 변환한다. 머신러닝 모델 변환은 모델 특성에 따른 적합한 전처리를 한후에, 결과를 받아 추가/변경을 반복한다. 지도학습 모델을 위한 전처리는 학습데이터와 학습 데이터로 분할해야 한다.

# 4. 전처리 방법

데이터 전처리에는 몇가지 방법이 있다. '데이터 정제' 단계에서는 데이터 내의 Noise를 제거하고 일관성 결여를 교정한다. '데이터 축소' 단계에서는 집계, 중복제거, 군집화와 같은 과정을 통하여 데이터의 크기를 축소한다. '데이터 변환' (예: 정규화)은 데이터를 0.0에서 1.0까지의 보다 작은 범위로 크기를 조정하여 정확도와 효율을 개선하는 과정이다.

## 4.1 데이터 정제

현실 세계의 데이터는 불완전하고 Noise가 있으며 일관성이 없다. 데이터 정제는 결측치를 채워 넣고 이상치를 다루며 Noise를 제거하여 데이터의 비일관성을 수정하는 시도이다.

### 4.1.1 결측치 대치 (Missing Imputation)

결측값은 모델의 결과를 왜곡하는 등 정확한 분석을 방해한다. 따라서 결측값이 있는 데이터는 적절한 조치를 통해 분석 가능한 형태로 만들어 줄 필요가 있다. 결측값을 포함한 obs(관측치, Observation)을 삭제하는 것이 가장 간단하겠지만 이러한 obs가 많을 경우 이를 모두 삭제하면 데이터 수가 과도하게 줄어드는 문제가 생긴다. 따라서 결측값의 유형, 규모, 다른 변수와의 관련성 등을 고려하여 결측값을 처리할 방법을 정할 필요가 있다.

1. 행을 무시하기
전체 삭제는 간편한 반면 앞에서 관측치가 과도하게 줄어들어 모델의 유효성이 낮아질 수 있다. 그리고 삭제는 결측값이 무작위(random)하게 발생한 경우 사용한다. 결측값이 무작위로 발생한 것이 아닌데 관측치를 삭제한 데이터를 사용할 경우 왜곡된 모델이 생성될 위험이 있다.

2. 열을 무시하기
이 역시 매우 간편하다. 또한 분석에 포함되는 feature가 줄어들기에 좋지 않을 수 있다. 하지만 한 열에 결측치의 비율이 너무 많을 때는 어쩔 수 없이 이 방법을 선택하기도 한다.

3. 수작업으로 결측치 채우기

4. 글로벌 상수값으로 결측치 채우기
모든 결측치 속성값을 "Unknown"이나 - ∞ 와 같은 라벨로 대체한다. 결측값을 "Unknown"으로 대체할 경우 프로그램이 유의미한 개념으로 이 값을 고려할 수도 있다.

5. 해당 속성의 결측치에 중심 경향 측정값을 사용하기 (예: 평균, 중위수)

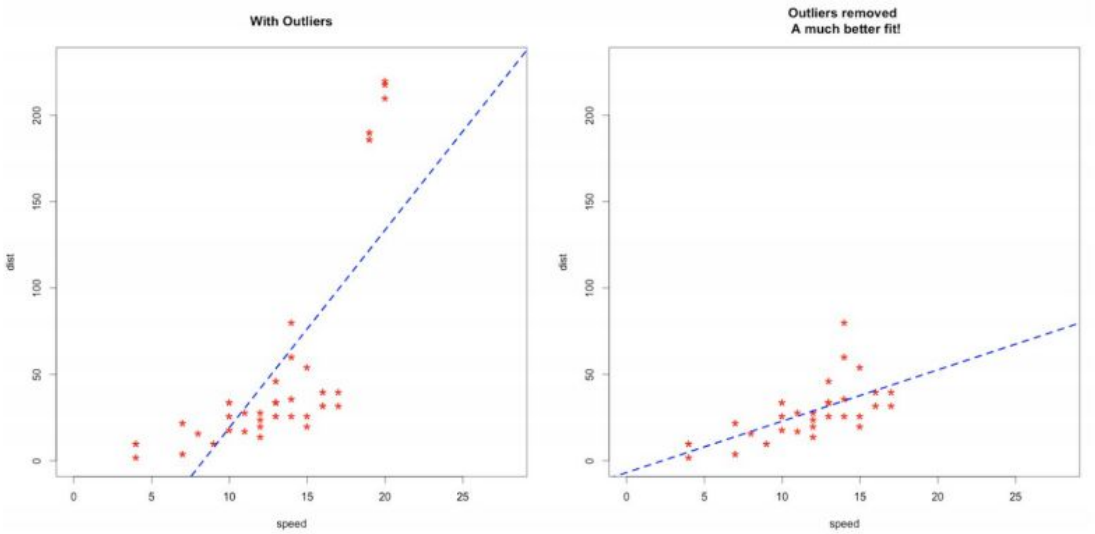
6. 결측값을 가진 샘플과 동일한 범주에 속하는 샘플들의 평균이나 중위수를 활용하기
예를 들어 고객을 Credit_risk에 따라 분류할 때, 결측치를 동일 신용위험 카테고리에 속하는 고객에 대한 평균수입으로 대체할 수 있다. 만약 주어진 클래스에 대한 데이터 분포가 편향되어 있다면 중위수가 좀 더 바람직한 값이 된다.

7. 가장 가능성이 높은 값으로 결측치 채우기
이 방법은 회귀분석이나 베이지안 공식, 의사결정나무, KNN 등 다른 모델의 도입을 이용한 추론 기반 도구로 결측치의 대체 값을 결정한다. 예를 들어 데이터 집합에서 다른 고객의 속성값을 사용하면 의사결정나무를 사용하여 수입에 대한 결측값을 예측할 수 있다.

### 4.1.2 이상치(Outlier)

이상치란 마치 다른 메커니즘으로 생성된 것처럼 나머지 다른 오브젝트로부터 멀리 뚝 떨어져 있는 데이터 오브젝트를 말한다. Outlier는 데이터의 Noise(잡음)와 다르다. Noise는 데이터 분석에서 중요한 요소가 아니지만, Outlier는 나머지 데이터와 생성 메커니즘이 다를 수 있다는 의혹을 제공한다. 따라서 이상치 탐색에서는 찾아낸 Outlier가 왜 다른 메커니즘을 따르는지 밝혀내는 것이 중요하다. 대개 나머지 데이터에 여러 가설을 적용한 다음 Outlier가 명백하게 주어진 가설에 위배된다는 점을 보임으로써 입증하는 방법을 사용한다.

- 검출 : 내면 스튜던트화 잔차 확인, Leverage, Cook's Distance
- 처리 : 삭제, 상·하한선 제한, 케이스 분리 분석



https://datascienceplus.com/outlier-detection-and-treatment-with-r/

머신러닝의 한 분야인 Anomaly(Outlier, Novelty) Detection은 위와 같은 이상치 검출을 다룬다.

### 4.1.3 노이즈(Noise)

Noise는 측정 변수의 랜덤 오류나 분산에 해당한다. 가격(price)과 같은 숫자 속성에 Noise를 제거하기 위해 데이터를 평활화(Smooth)하는 방법이 있다.

1. 비닝

비닝은 근접한 다른 값(Neighborhood)를 참고하여 정렬한 데이터 값을 평활화한다. 아래의 표에서는 몇 가지 기술을 보여주고 있다. 우선 가격 데이터를 정렬한 후 각 빈(Bin)이 3개 데이터를 갖도록 나눈다. 빈 평균으로 평활화할 때, 해당 빈에 속하는 개별 값은 빈의 평균값으로 대체한다.

예를 들어 3개의 값 (4, 8, 15)이 있는 경우 평균이 9이므로 원래 값은 9로 대체할 수 있다. 빈 중위수에 의한 평활화도 사용할 수 있는데 이 경우 각 빈의 값은 해당 빈의 중위수로 대체한다. 빈 경계값에 의한 평활화는 해당 빈에 대한 최소와 최대값을 빈 경계값으로 계산한다. 각 빈의 값은 2개의 값 중에서 가장 근접한 경계값으로 대체한다.

가격으로 정렬 (달러 기준) : 4, 8, 15, 21, 21, 24, 25, 28, 34

동일빈도 빈으로 **분할** :
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

**빈 평균**으로 평활화 :
Bin 1: 9,9,9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

**빈 경계**로 평활화 :
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

2. 회귀분석

데이터 평활화는 데이터 값을 함수에 적용한 기술인 회귀분석으로도 가능하다. 선형회귀는 2개 속성에 대하여 최적합하는 라인을 알아내는 방법이다. 두 속성 중 한 개의 속성은 다른 속성을 예측하는 데 사용한다. 다중회귀식은 선형회귀식의 확장으로 2개 이상의 속성을 대상으로 다차원 평면에 해당 데이터를 적합시키는 방법이다.

## 4.2 데이터 축소

데이터 축소는 샘플링 등으로 관측치를 줄이거나 차원(변수,속성)을 줄이는 작업이다. 분석하려는 데이터에 너무 많은 변수가 존재한다면, 데이터 분석의 시간 효율이 떨어질 수 밖에 없다. 또한 데이터 분석에 영향을 미치지 않거나 타 변수와 중복적 성격을 띠는 것도 많이 존재할 수 있다. 이러한 변수들을 충분히 제거하지 않고 분석작업을 시행하면, 알고리즘에 혼동을 줄 수 있다.

따라서 연관성이 낮은 변수를 제거하고, 중복된 데이터 차원(변수)를 제거하거나, 통합하여 데이터 집합의 크기를 줄이는 노력이 필요하다. 가장 쉬운 방법은 변수들 간의 상관계수를 확인하는 것이다. 상관계수가 클수록 굳이 두 변수 모두 넣을 필요 없이 하나만 선택하여 넣어도 충분하다. 또 하나의 방법으로는 주성분분석(PCA)이 있다. 이는 변수들의 선형결합을 통해 새로운 변수를 만드는 것으로서 차원을 축소시키는 대표적인 방법이다.

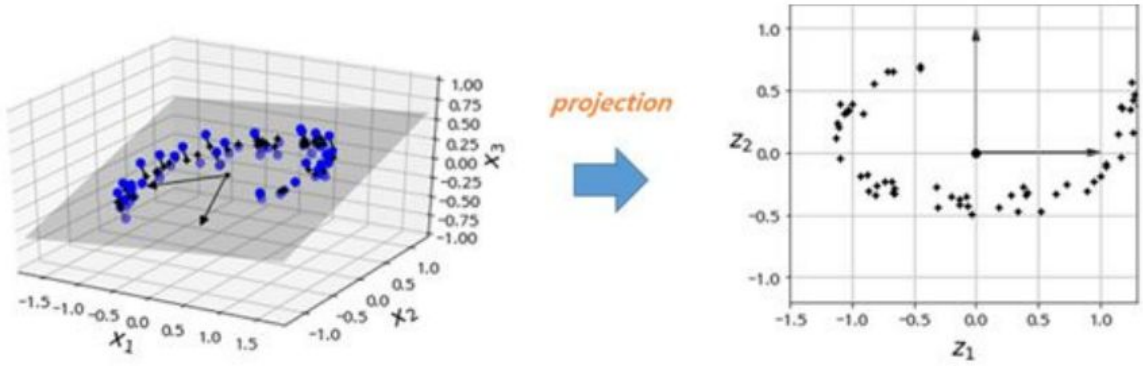
### 4.2.1 차원축소

많은 경우 훈련 샘플 각각이 수천 심지어 수백만 개의 특성을 가지고 있다. 이는 학습을 느리게 하며 좋은 솔루션을 찾는 데에도 방해가 된다. 이를 '차원의 저주(Curse of Dimensionality)'라고 한다. '차원축소'란 이 저주를 해결하기 위해 샘플링 등으로 데이터의 양을 줄이거나 차원(변수)을 줄이는 작업이다. 너무 많은 변수로 인해 야기되는 비효율성을 개선하는 것이 주요 목적이다.

차원 축소는 주로 분석에 영향을 미치지 않는 데이터를 배제하거나 타 변수와 중복적 성격을 띠는 변수들을 통합 혹은 제거하여 효율성을 제고한다. 가장 간단한 예를 들자면 상관관계가 큰 두 변수를 모두 사용하는 것은 정보의 중첩으로 볼 수 있으므로 한 변수만을 채택하여 분석에 활용하는 것 또한 차원축소라 할 수 있다. 이 파트에서는 차원축소에 사용되는 주요 접근 방법 '투영(Projection)'과 가장 인기있는 기법인 PCA(주성분 분석)에 대해 다루도록 한다.

### 4.2.2 투영

대부분 실전문제는 훈련 샘플이 모든 차원에 걸쳐 균일하게 퍼져 있지 않다. 많은 변수의 관점에서는 거의 변화가 없는 반면 다른 특정 변수의 입장에서는 서로 강하게 연관되어 변화하고 움직인다. 결과적으로 모든 훈련 샘플이 고차원 공간 안의 저차원 부분공간(Subspace)에 놓여있다고 할 수 있다.

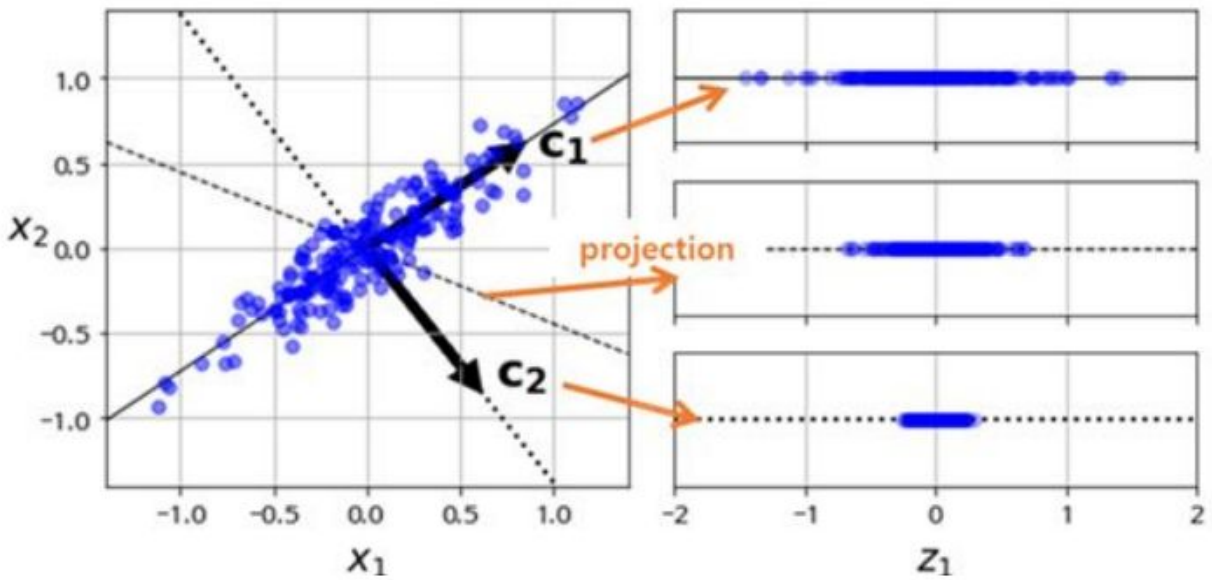



다음 그림에서 볼수 있듯이 비록 3차원 공간의 데이터이지만 거의 모든 샘플이 2차원 평면 위에 놓여있다. 이것이 바로 고차원 공간에 있는 저차원 부분 공간이다. 여기서 모든 샘플을 이 부분공간에 수직으로 투영하면 두 번째 그림과 같은 데이터 셋을 얻을 수 있다. 이렇게 우리는 정보의 손실을 최소화 하면서 3차원 데이터를 2차원으로 줄이는 데에 성공했다.

### 4.2.3 PCA(Principal Component Analysis | 주성분 분석)

주성분 분석은 가장 인기 있는 차원 축소 알고리즘이다. 먼저 앞서 Projection에서 본 바와 같이 데이터에 가장 가까운 초평면(Hyperplane)을 정의한 다음 데이터를 이 평면에 투영시킨다.

1) 분산(정보)의 보존



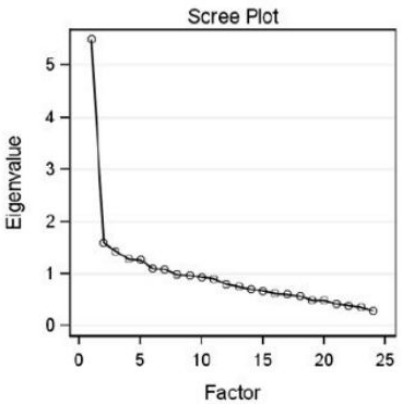
PCA에서는 올바른 초평면을 선택하는 것이 가장 중요하다. 위 그림은 세 개의 축에 투영된 원 데이터의 결과를 보여준다. 여기서 볼 수 있듯이 실선에 투영된 건은 분산을 최대로 보존하는 반면, 점선에 투영된 것은 분산을 매우 적게 유지하고 있다. 가운데의 파선에 투영된 것은 분산을 중간 정도로 유지하고 있다.
여기서 분산을 데이터의 변동(Variation)으로 받아들이면 분산의 유지가 곧 오리지널 데이터의 정보의 유지라는 것을 알 수 있다. 만약 실선에 투영된 결과가 원 데이터의 총 변동의 80% 정도를 반영한다면 2차원 데이터를 1차원 데이터로 축소하는 것이 가능할 것이다. 다른 관점에서 보면 $X_1$과 $X_2$가 양의 상관관계를 가지고 있기 때문에 애초에 정보의 중첩이 일어난 상태였고 PCA는 이 중복된 정보를 하나의 초평면에 반영하여 하나의 변수로 줄이는 과정이라 할 수 있다.

2) Principal Component

1. Train set에서 분산이 최대인 축을 찾는다. 위 그림에서는 실선이 그에 해당한다.
2. 첫 번째 축에 직교하고 남은 분산을 최대한 보존하는 두 번째 축을 찾는다. 위 그림에서는 선택의 여지가 없다(2D이기 때문에). 즉 점선이 된다.
3. 고차원 데이터 셋이라면 PCA는 이전의 두 축에 모두 직교하는 세 번째 축을 찾으며 데이터셋에 있는 차원의 수(변수의 수)만큼 네 번째, 다섯 번째, ..., n 번째 축을 찾는다.

이렇게 찾은 i 번째 축을 정의하는 Eigenvector를 i 번째 주성분(Principal Component)라고 한다. 축은 모두 상호 직교하므로 주성분들은 서로 독립이다(Independent: 내적 결과값이 0).
주성분을 찾는 방법으로는 Singular Vector Decomposition, Spectral Decomposition 등의 수학적 방법론이 있다. (R과 파이썬에 이에 대한 편리한 패키지들과 함수들이 존재한다)

3) Scree Plot



scree plot에서는 elbow point를 찾자.

Scree plot은 유의한 주성분 개수를 정하는 데에 유용하게 사용되는 도표이다. 가로축은 첫 번째 주성분부터 오름차순으로 주성분들이 나열되어 있고 세로축의 Eigenvalue는 각 주성분의 분산으로 이해하면 된다. 만약 모든 기존 변수들이 Scaling 되어있다 가정하면 각 변수들의 분산은 모두 1이 된다. 즉 첫 번째 주성분의 분산이 5를 넘는 값을 가진다는 것은 그만큼 많은 정보가 첫 번째 초평면에 투영되었다는 것을 의미한다.

유의한 주성분의 개수를 정하는 것에는 많은 기준이 있다. Kaiser's Rule에 따르면 기존 변수의 분산인 1보다 작은 분산을 가지는 주성분은 의미가 없다고 여겨 분산이 1보다 크거나 같은 주성분만이 유의하다 판단한다. Elbow Rule이라는 기준도 있는데 Scree plot의 그래프가 팔꿈치처럼 확 꺾이는 부분이 유의함의 척도가 된다는 이론이다. 그 외에도 연구자의 배경지식 혹은 연구 목적에 따른 다양한 자의적 판단 또한 가능하다.

4) PCA의 장점과 단점

PCA는 간단하면서도 직관적인 방법으로 원 데이터의 정보를 최대한 보존하면서 변수의 수를 줄이는(차원을 축소하는) 해법을 제시한다. 따라서 특정 기준을 통해 유의한 PC의 개수를 구하면 자동으로 정보의 중첩이 없는 데이터의 순수한 차원을 구할 수 있다. 하지만 단점으로는 이렇게 구한 PC는 해석이 아주 어렵다는 점이다. 위의 $X_1$과 $X_2$의 정보를 가장 잘 반영한 실선을 PC1이라 하면 PC1은 $PC1 = kX_1 + sX_2$.으로 구한다. 이렇게 병합된 데이터는 $X_1$도 $X_2$도 아닌 새로운 정보이므로 해석하기가 매우 모호해진다. 물론 k와 s의 정보로 각 변수의 기여도를 해석하는 방법이 있지만 고차원 데이터가 될수록 이 또한 애매해진다. 따라서 주성분 분석을 통해 구한 주성분을 추가적인 분석에 활용함에는 명확한 한계가 있다. 다만, 유의한 주성분의 개수가 곧 어떠한 알려지지 않은 유의한 잠재적 변수(Latent Variable)의 개수일 것이라는 추측이 가능하다는 점에서 의의가 있다.

## 4.3 데이터 변환과 데이터 구분

측정단위는 데이터 분석에 영향을 줄 수 있다. 예를 들어 신장 측정단위를 미터에서 인치로 변환하거나 몸무게 측정 단위를 킬로그램에서 파운드로 변환하는 것은 다른 결과를 도출한다. 측정 단위에 종속된 문제점을 방지하기 위해서는 데이터는 정규화(Normalization) 또는 표준화(Standardization)해야 한다. 이러한 과정은 해당 데이터가 [-1, 1] 또는 [0, 1]과 같은 작은 범위 내에 위치하도록 한다.

데이터 정규화는 모든 속성에 동일한 가중치를 적용한다. 정규화는 최근접 분류와 신경망이나 거리측정을 포함한 분류알고리즘에 매우 유용한다. 거리기반 방법의 경우 정규화는 초기 큰 범위를 갖는 속성이 상대적으로 작은 범위를 갖는 속성보다 가중치가 높게 적용되지 않도록 한다.

### 4.3.1 최소-최대 정규화

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

최소-최대 정규화는 원 데이터에 대해 선형 변환을 한다. 속성 X에 대한 최소값과 최대값을 $X_{min}$, $X_{max}$이라고 할 때, 최소-최대 정규화의 식은 아래에 있는 식과 같다. 최소-최대 정규화는 원 데이터 값 간의 관계를 그대로 유지하면서 해당 속성을 0~1사이의 값으로 나타낸다.

### 4.3.2 Z스코어 정규화

$$x_{new} = \frac{x - \mu}{\sigma}$$

각 Observation이 평균을 기준으로 어느 정도 떨어져 있는지를 나타낼때 사용된다. 값의 스케일이 다른 두 변수가 있을 때, 이 변수들의 스케일 차이를 제거해 주는 효과가 있다. 제로 평균 으로부터 각 값들의 분산을 나타낸다. 각 요소의 값에서 평균을 뺀 다음 표준편차로 나누어 준다.

### 4.3.3 십진스케일 정규화

십진스케일링에 의한 정규화는 속성 X의 값을 10의 배수 값으로 이동시켜 정규화한다. 십의 자리수는 X의 최대 절대값만큼 이동한다. 속성 X의 값 xi는 다음과 같은 식에 의해 정규화한다.

$$x_{new} = \frac{x_i}{10^j}, \quad \text{j는 } \max(|x_{new}|) < 1 \text{를 만족하는 최소 정수}$$

ex) X의 값은 -986에서 917의 범위에 존재한다. 따라서 A의 최대 절대값은 986이다. 십진스케일링을 이용하려면 각 값은 1000(j=3)으로 나누며 -986은 -0.986이 되고 917은 0.917이 된다.

## 5. Reference

지아웨이 한, 미셸린 캠버, 지안 페이, 데이터 마이닝 개념과 기법, 에이콘, 2015
오렐리앙 제롱, 핸즈온 머신러닝, 한빛미디어, 2018
데이터 전처리 대전 모토하시 도모미쓰