2. Preprocessing
2019년 가을 학기

2019년 08월 28일
https://www.github.com/KU-BIG/KUBIG_2019_Autumn

이 강의는 '*지아웨이 한, 미셸린 캠버, 지안 페이, 데이터 마이닝 개념과 기법, 에이콘, 2017*'에 바탕을 두고 있습니다.

1. Introduction

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. 먼저 데이터 특성을 조사하기 위한 기술통계를 알아보고, 데이터 정제, 데이터 축소, 데이터 변환에 대해 살펴보도록 한다.

2. 데이터에 대한 기술통계

데이터 전처리를 성공하려면 데이터에 대한 전체 그림을 파악할 필요가 있다. 기술통계값은 데이터의 특성을 식별하고 잡음이나 이상치로 처리해야 하는 데이터를 찾아낼 때 사용한다.

2.1 중심 경향 측정 : 평균(mean), 중위수(median), 최빈값(mode)

중심경향의 측정값에는 평균, 중위수, 최빈값이 있다. 데이터 집합의 중심에 대한 가장 일반적이고 효과적인 수치는 '**평균**'이다. $X_1,X_2,...,X_n$을 n개의 값의 집합이라고 하자. 값의 집합에 대한 평균은 아래와 같다.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

하지만 평균이 언제나 데이터의 중심을 측정하는 바람직한 방법인 것은 아니다. 평균은 이상치에 민감하기 때문이다. 적은 양의 이상치도 평균을 오염시킬 수 있다. 예를 들어, 회사의 평균 연봉은 고액 연봉을 받는 소수의 관리자에 의해 값이 올라갈 수 있다. 비대칭 데이터(이상치를

포함하거나 밀집된 지역이 치우쳐진 데이터)에서 보다 바람직한 중심 측정단위는 대소에 따라 정렬된 데이터에서의 중간값인 '중위수'이다. 이 값은 데이터를 상위 50%와 하위 50%로 구분하는 값이다. 중심 경향에 대한 또 다른 측정 값으로 '최빈값'이 있다. 최빈값은 데이터 집합에서 가장 빈번하게 발생하는 값이다.

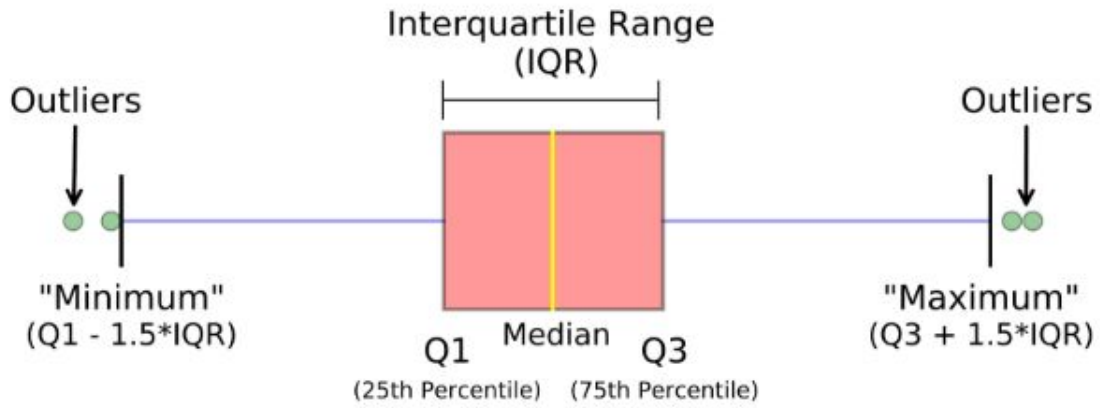
2.2 데이터 산포 측정 : 범위, 사분위수, 분산, 표준편차, 사분위 범위

'범위(range)'는 데이터의 최대값과 최소값 간의 차이를 말한다. '사분위수(quartile)'는 측정값을 오름차순으로 정렬한 후 4등분 했을 때 각 등위에 해당하는 값을 의미한다. 25%의 데이터가 제 1사분위수(Q1)보다 작거나 같으며, 중앙값(median)은 제 2사분위수(Q2)에 해당한다. Q1과 Q3사이의 거리는 데이터가 흩어져 있는 정도를 측정하는 통계량으로도 쓰이는데 이를 '사분위 범위(IQR,interquartile range)'라고 한다. 사분위 범위의 식은 다음과 같다.

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$$

2.3 다섯 숫자 요약, 박스 플롯, 이상치

IQR과 같이 산포에 대한 단일 측정치는 분포의 대칭성과 비대칭성을 설명하는 데 큰 도움이 되지는 않는다. 중위수와 함께 Q1과 Q3를 제시하는 것이 더 많은 정보를 제공한다. 하지만 Q1, 중위수, Q3는 데이터 분포의 꼬리에 대한 정보를 포함하고 있지 않다. 이는 다섯 숫자 요약을 통해 제공된다. 다섯 숫자 요약은 최소값, Q1, 중위수, Q3, 최대값으로 구성된다. 박스플롯은 분포를 시각화하는데 자주 사용하는 방법이다. 박스플롯은 다음과 같이 다섯 숫자 요약을 포함한다.



https://www.simplypsychology.org/boxplots.html

3. 데이터 정제

현실 세계의 데이터는 불완전하고 Noise가 있으며 일관성이 없다. 데이터 정제는 결측치를 채워 넣고 이상치를 다루며 Noise를 제거하여 데이터의 비일관성을 수정하는 시도이다. 이번 절에서는 데이터 정제를 위한 기초적인 방법에 대해 알아보도록 한다.

3.1 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 $-\infty$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.

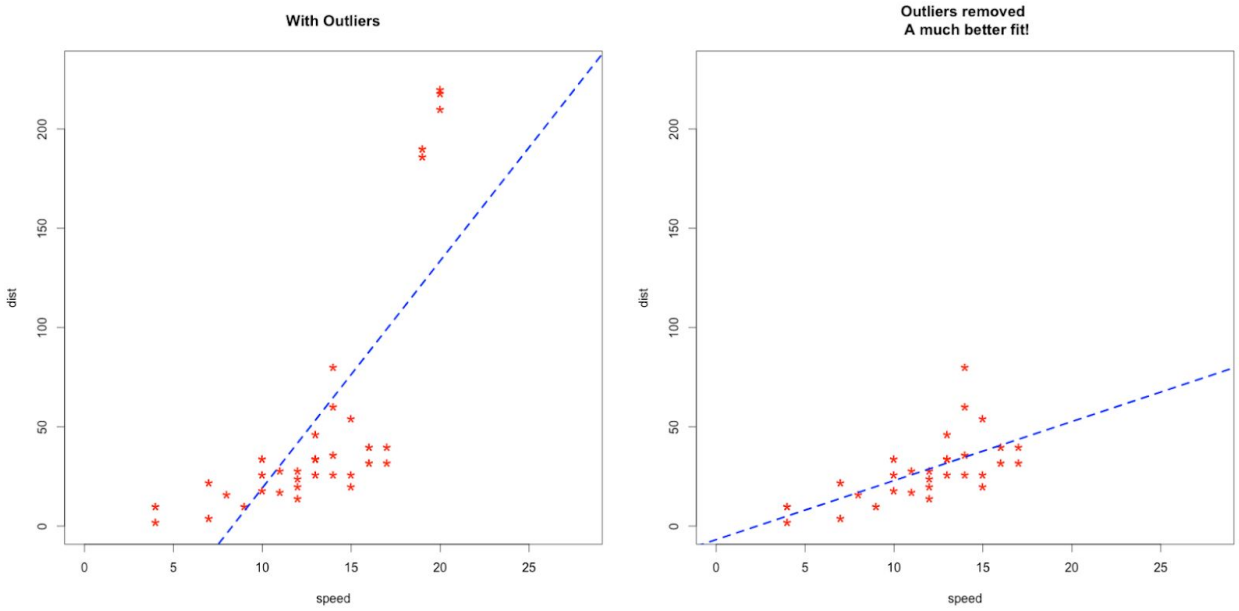
7. 가장 가능성이 높은 값으로 결측치 채우기
이 방법은 회귀분석이나 베이지안 공식, 의사결정나무, KNN 등 다른 모델의 도입을 이용한 추론 기반 도구로 결측치의 대체 값을 결정한다. 예를 들어 데이터 집합에서 다른 고객의 속성값을 사용하면 의사결정나무를 사용하여 수입에 대한 결측값을 예측할 수 있다.

3.2 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.

3.3 노이즈(Noise)

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

- 비닝(Binning)

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 평균으로 평활화 :
Bin 1: 9,9,9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

빈 경계로 평활화 :
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

- 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. 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.

4.1 차원축소

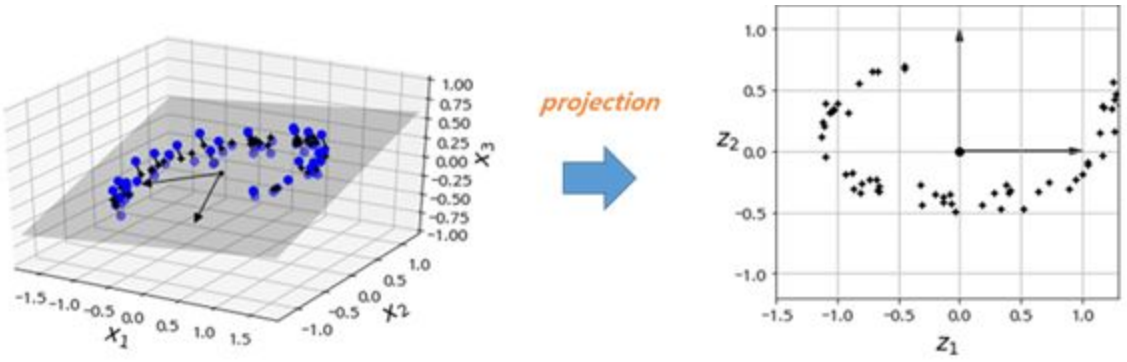
많은 경우 훈련 샘플 각각이 수천 심지어 수백만 개의 특성을 가지고 있다. 이는 학습을 느리게 하며 좋은 솔루션을 찾는 데에도 방해가 된다. 이를 '**차원의 저주(Curse of Dimensionality)**'라고 한다. '차원축소'란 이 저주를 해결하기 위해 샘플링 등으로 데이터의 양을 줄이거나 차원(변수)을 줄이는 작업이다. 너무 많은 변수로 인해 야기되는 비효율성을 개선하는 것이 주요 목적이다.

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 투영 (projection)

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. 추상적이기 때문에 그림의 예와 함께 살펴보자

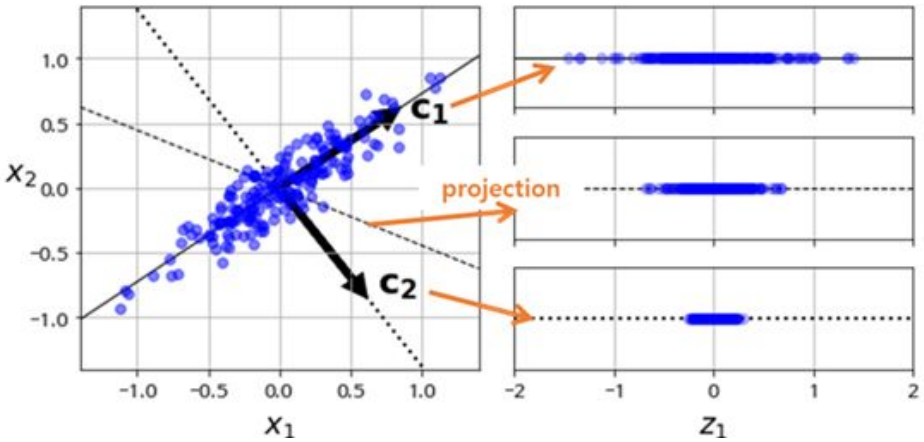


비록 3차원 공간의 데이터이지만 거의 모든 샘플이 2차원 평면 위에 놓여있다. 이것이 바로 고차원 공간에 있는 저차원 부분 공간이다. 여기서 모든 샘플을 이 부분공간에 수직으로 투영하면 두 번째 그림과 같은 데이터 셋을 얻을 수 있다. 이렇게 우리는 정보의 손실을 최소화 하면서 3차원 데이터를 2차원으로 줄이는 데에 성공했다.

4.3 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.

4.3.1 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_1과 X_2가 양의 상관관계를 가지고 있기 때문에 애초에 정보의 중첩이 일어난 상태였고 PCA는 이 중복된 정보를 하나의 초평면에 반영하여 하나의 변수로 줄이는 과정이라 할 수 있다.

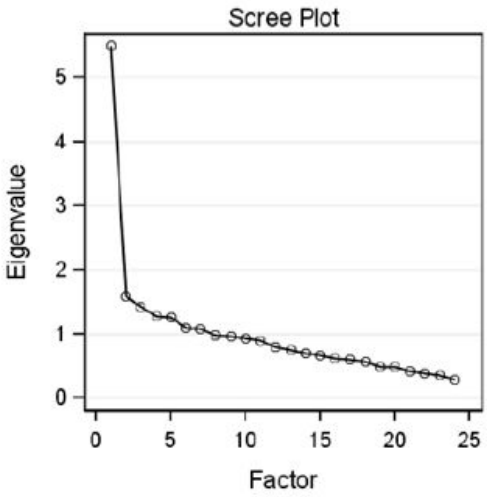
4.3.2 Principal Component

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 있지만 자세한 설명은 생략하고 Python 혹은 R의 힘을 빌리기로 한다.

4.3.3 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 변수들(X_1, X_2, ..., X_n)이 Scaling되어있다 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.

4.3.4 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_1과 X_2의 정보를 가장 잘 반영한 실선을 PC1이라 하면 PC1은 다음과 같은 형식으로 구한다. $PC1 = kX_1 + sX_2$. 이렇게 병합된 데이터는 X_1도 X_2도 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.

5. 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. 여기서는 최소-최대 정규화, Z스코어 정규화, 십진스케일 정규화를 소개한다.

5.1 최소-최대 정규화

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.

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

5.2 Z스코어 정규화

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 평균으로부터 각 값들의 분산을 나타낸다. 각 요소의 값에서 평균을 뺀 다음 표준편차로 나누어 준다.

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

5.3 십진스케일 정규화

십진스케일링에 의한 정규화는 속성 X의 값을 10의 배수 값으로 이동시켜 정규화한다. 십의 자리수는 X의 최대 절대값만큼 이동한다. 속성 X의 값 x_i는 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.

Reference

지아웨이 한, 미셸린 캠버, 지안 페이, 데이터 마이닝 개념과 기법, 에이콘, 2015

오렐리앙 제롱, 핸즈온 머신러닝, 한빛미디어, 2018

https://bit.ly/2GMKK8k

https://bit.ly/2ZBrXFc