아이리포 지식창고

02 주성분분석 PCA

(주)씨에스리 조경미 기술사
(kmicho@cslee.co.kr)

다차원 데이터에서 특징, 패턴을 추출하는 통계 기반 알고리즘

Concept	3개 이상의 다차원 데이터에서 Principal Component를 추출하여 그 특징(feature)을 추출하거나 패턴(Pattern)을 찾는 통계 기반 분석 알고리즘
Keyword	표준편차(Standard Deviation), 분산(Variance), 공분산(Covariance), 고유값(Eigenvalue), 고유벡터(Eigenvector), 차원의 축소

이론과 실제가 다를 수 있지만, 차원을 줄여서 얼굴인식에 쓸 수 있겠다고 막연히 생각했던 PCA(Principal Component Analysis) 알고리즘이 스마트팩토리에서 활용되는 알고리즘이란 걸 알게 되어, 엇! PCA 가 뭐지? 알아봐야겠다는 생각에 오늘은 차원의 저주를 해결한다는, 얼굴인식에 활용된다는 주성분 분석, PCA(Principal Component Analysis) 에 대해 알아보려고 한다.

주성분 분석으로 다차원 특징 벡터의 정보를 유지하며,
저차원으로 차원을 축소하는 다변량 데이터 분석 기법으로, 안면인식 등 영상인식에서 널리 활용 됨

(출처: 아이리포카페 http://café.naver.com/itlf)

PCA 의 개념에 소개한 **다변량 데이터 분석 기법**은 변수들 간의 인과관계를 규명하거나, 변수들 간의 상관관계를 이용하여 변수들을 축약하거나 개체들을 분류하는 분석 기법이다. 변수들 간의 인과관계를 규명하는 다변량 데이터 분석기법으로는 다중회귀분석(Multiple Regression Analysis), 다중 분산분석(Multiple ANOVA) 이 있으며, 변수들을 축약하거나 개체들을 분류하는 분석기법으로는 PCA 등을 활용하기도 한다.

다변량 데이터 분석 기법	**인과관계 분석**	- 다중회귀분석(Multiple Regression Analysis)
		- 다중분산분석(Multiple ANalysis Of VAriance)
	변수들의 축약	- PCA(Principal Component Analysis)

이와 같이 다변량 데이터를 대상으로 하는 데이터 분석에서는 **여러 변수들 간의 상관관계를 소수의 주성분으로 차원을 축소하여 데이터를 쉽게 이해**할 수 있고, IoT 의 활성화로 다수의 센서를 통해 수집되는 센서데이터를 주성분분석으로 **차원을 축소한 후에 시계열 분포나 추세의 변화를 분석하면 스마트팩토리에서 기계의 고장을 예측, 사전 감지**하는데 활용할 수 도 있다고 한다. 그래서 스마트팩토리의 데이터 분석에 PCA 알고리즘이 활용된다.

PCA 를 이해하기 위해 기본적인 수학지식으로 공분산(covariance), 고유값(eigen value), 고유벡터(eigen vector)등의 개념 이해가 필요하다. 각각을 알아보고, PCA 알고리즘 동작 절차 및 얼굴인식에 활용된 사례를 알아본다.

1. PCA 이해에 필요한 기본적인 수학 지식

가. 평균(Mean), 분산(Variance), 표준편차(Standard Deviation)

- 평균(Mean)은 중심 위치를 측정하는 기법이고, 분산(Variance), 표준편차(Standard Deviation)는 중심으로 부터 얼마나 데이터가 퍼져 있는지 측정하는 변동의 측정 기법이다.

항목	개념	수식
평균 Mean	-데이터의 총합을 표본의 크기로 나눈 값. -대표적인 중심위치 측정 기법	$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
분산 Variance	-관찰값과 평균의 차이의 제곱의 평균 -평균으로부터 얼마나 떨저여서 분산되어 있는가를 가늠하는 변동을 측정하는 하나의 척도	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$ $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
표준편차 Standard Deviation	-분산의 제곱근 -평균으로부터 관찰값까지의 평균 거리	$\sqrt{\frac{\sum(X-Mean)^2}{n-1}}$

나. 공분산(Covariance)

- **분산, 표준편차는 1 차원(1 Dimension)의 변동 측정 기법**이다. 즉 input 변수 x 가 한 개 일 때 사용한다. 그러나 다변량 데이터 분석, 다차원의 데이터에서의 변동을 측정하기 위해서는 **2 차원(2 Dimensions)의 변동 측정 기법으로 공분산**을 이용하며, 3 차원 이상의 데이터(예, data set(x, y, z))를 분석하고자 할 때는, data set(x, y), data set (y, z), data set(x, z)의 공분산을 이용하여 분석 가능하다.

항목	개념	수식
공분산 Covariance	-두 확률 변수의 상관관계를 나타내는 값 - C > 0, 양의 상관관계 - C < 0, 음의 상관관계 - C = 0, 두 변수는 독립임	$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$

다. 고유값(EigenValue) & 고유벡터(EigneVector)

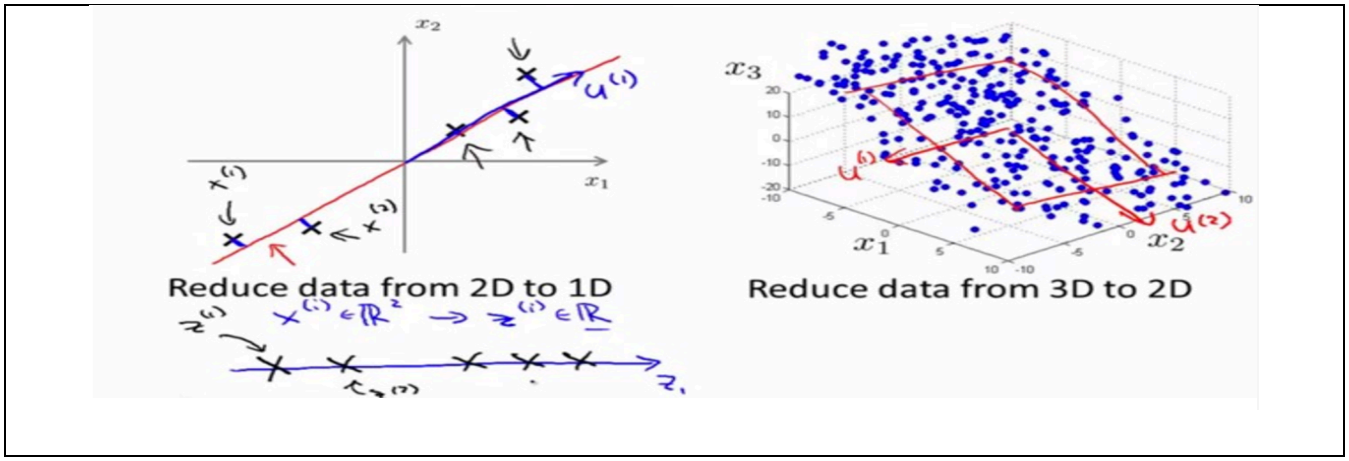
항목	개념	사례
고유값 EigenValue	행렬 A 를 선형변환한 결과가 자기 자신의 상수배가 되게 하는 값	$2 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$ $\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix}$ 위 행렬에서 고유값은 4 고유벡터는 $\begin{pmatrix} 6 \\ 4 \end{pmatrix}$
고유벡터 EigenVector	행렬 A 를 선형변환한 결과가 자기 자신의 상수배가 되게 하는 벡터	

*선형변환(Linear Transformation)은 벡터 공간에서 벡터 공간으로 가는 함수로, 그것들 중 벡터 공간의 성질을 보존하는, 즉 선형성을 갖는 함수로 선형맵(Linear Map), 일차변환(First Order Transformation)이라고도 한다. 행렬을 확장(Scaling), 회전(Rotation) 등의 변환이 가능하다.

그렇다면, 이제 PCA 에서 공분산과 고유값, 고유벡터를 이용하는 PCA 에 대해 알아본다.

2. PCA (Principal Component Analysis)

가. PCA 알고리즘 개념도



(출처 Machine Learning Lectures by Prof. Andrew NG at Stanford University)

-PCA 는 데이터 하나하나에 대한 성분을 분석하는 것이 아닌, 여러 데이터들이 모여 하나의 분포를 이 때 그 분포의 주 성분을 분석하는 기법이다. 그림에서는 2 차원 데이터를 1 차원으로 변환, 3 차원 데이터를 PCA 를 통해 2 차원데이터로 변환하는 예를 보여주고 있다.

나. PCA 알고리즘 수행 절차

단계	수행 절차	절차 설명
1	데이터셋 로드	-PCA 분석을 위한 데이터셋을 준비한다.
2	평균, 공분산 계산	-평균값과 편차를 구하고, 공분산을 계산한다.
3	고유값, 고유벡터 계산	-해당 데이터집합을 가장 잘 표현하는 고유값, 고유벡터를 구한다. (2.가 에서 PC1, PC2 가 고유값, 고유벡터가 된다)
4	변환 (Transform) 수행	-고유값, 고유벡터를 이용하여 회전, 확장을 하여 새로운 기존 데이터셋을 설명하거나, 새로운 데이터셋을 예측한다.

- 스마트팩토리에서 다양한 센서들을 통해 수집되는 수많은 데이터 분석시, PCA 를 이용하여 주성분을 파악하고, 장애나 고장을 예측하는 분석 모델을 설계할 수 있다.

3. PCA 를 이용한 얼굴인식 eigenface 사례

데이터 준비	45*40 얼굴 이미지 20 장 = 1800 차원의 벡터 (즉, 1800 차원 공간에서 한 점에 대응)
PCA 수행	-평균, 공분산 계산 -고유값, 고유벡터 계산 -주성분 도출
이미지 해석	-도출된 주성분을 통해 이미지 센싱으로 eigenface 도출

-컴퓨터 비전에서도 얼굴인식, 얼굴검출을 위해 PCA 알고리즘을 사용한다.

빅데이터분석이 다양한 산업분야에 활용되며 그 가치를 인정받기 시작하며, 여러가지 분석 알고리즘과 그 활용에 대한 관심도 높아졌다. 빅데이터분석을 잘 하기 위해서 해당 도메인에 대한 깊은 이해를 기반으로 인사이트를 가지고, PCA 등과 같은 알고리즘 기법을 활용한다면 빅데이터분석을 통해 보다 나은 가치를 창출할 수 있을거라 생각한다.

[참조]

A tutorial on Principal Components Analysis / Lindsay I Smith

http://wolfpack.hnu.ac.kr/2014_fall/LN_MDA_SAS%202014f.pdf

http://www.nlpca.org/pca_principal_component_analysis.html

http://rfriend.tistory.com/61

http://darkpgmr.tistory.com/105