dimension reduction기법

생성일	@2022년 3월 3일 오후 4:32
속성	머신러닝
태그	면접질문

차원의 저주 - Curse of dimension

- 입력된 데이터의 수보다 데이터의 차원이 더 큰 경우 발생하는 문제를 차원의 저주라 합니다.
- 우리는 *데이터(벡터)가 뿌려진 벡터 공간에서 분류 또는 예측하는 가장 적합한 함수를 찾는게 기계학습, 딥러닝 학습의 목표*인데요. 입력한 데이터의 양은 적고, 데이터의 차원이 커지게 된다면
- 이때 벡터 공간의 차원이 무수히 커지고 데이터는 여기저기 흩뿌려져 있는 상황입니다. 이 흩어진 벡터들을 분류 예측하는 함수의 모형은 복잡해지게 됩니다. *즉, 모델의 복잡도가 증가하고 예측 성능이 낮아지게 됩니다.*
- 선형대수로 표현하자면 0으로 가득한 벡터로 채워진 분산 행렬(sparse matrix)의 형태일 것입니다. *이를 위해서 사용하는 방법이* ***Feature selection (피쳐 선택), Feature Extraction(피쳐 추출)*** *입니다*.

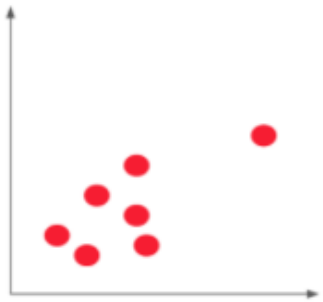
Feature selection & Feature Extraction

- 피쳐 선택은 가령 500개의 차원중 N개의 차원을 골라 내는 것이고, 반면 피쳐 추출은 500개의 차원을 N차원으로 압축하는 겁니다.
- 피쳐 선택의 경우에는 다중공산성을 고려해 상관성이 높은 피쳐들을 소거해가는 방식으로 진행할 수 있고, 피쳐 추출의 경우는 PCA, SVD, MF 와 같은 차원 축소 기법들을 활용해 벡터의 차원을 줄여나갈 수 있습니다.

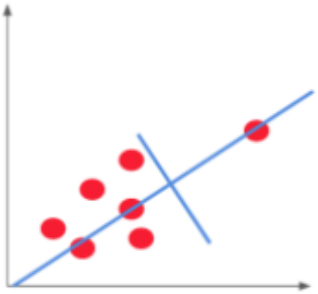
PCA는 차원 축소 기법이면서, 데이터 압축 기법이기도 하고, 노이즈 제거 기법이기도 합니다.

PCA 분석

다음과 같은 데이타가 있다고 하자.



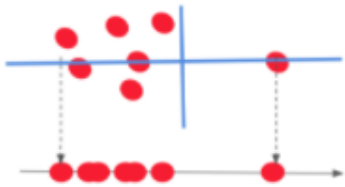
PCA 분석에서는 데이타의 변화의 폭이 가장 큰 축을 정하고, 그 다음 그와 직교하는 축을 구한다



그리고 데이타의 중심점에 축을 위치 시켜서 0,0을 중심으로 데이타가 양쪽으로 균등하게 퍼지도록 분포를 시켜서 축을 뒤틀어서 아래와 같이 원래의 데이타를 변화 시킨다.

이렇게 PCA 분석을 하면, 데이타의 중심축을 0,0으로 위치 시킬 수 있고, 가장 데이타의 변화의 폭이 큰 순으로 X,Y축등을 지정하여 데이타를 볼 수 있다.

앞의 예제 데이타에서 2차원 데이타를 PCA 분석을 해서 첫번째 피쳐가 Variance가 가장 높다고 했을 때 이를 변환하면 다음과 같이 PCA 변환된 데이타의 X축의 값만을 사용하도록 해서 2차원을 1차원으로 줄일 수 있다.



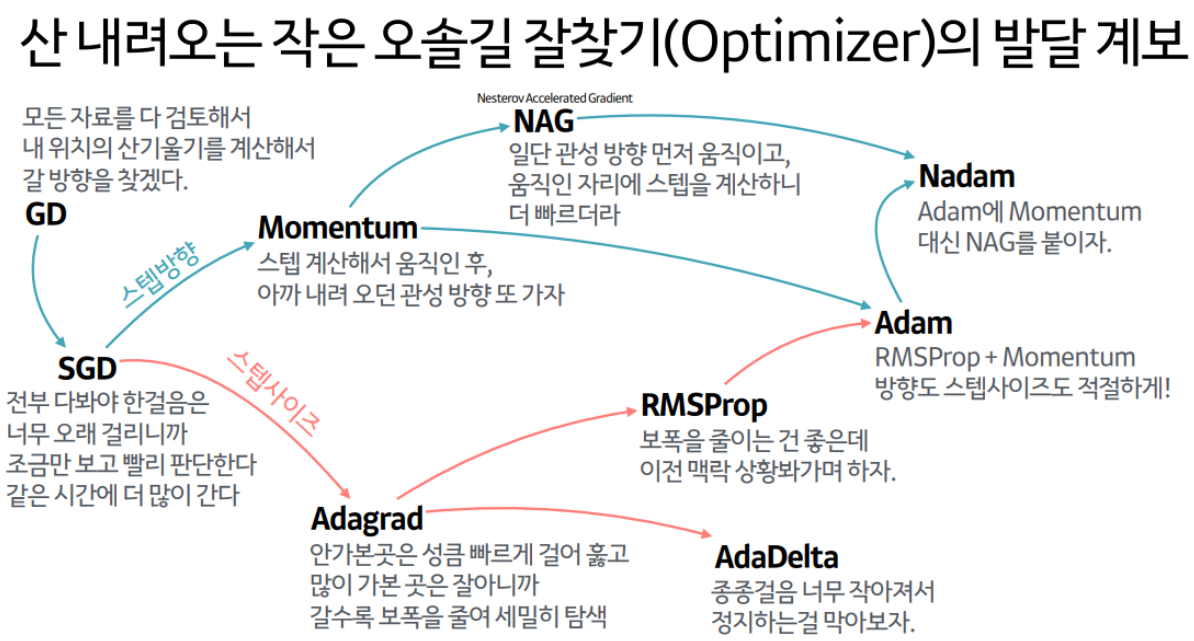
물론 차원을 줄이면, 원래 데이타가 가지고 있는 특징이 다소 사라지는 단점이 있지만, 전체적인 데이타의 특성을 파악하는 대세에는 큰 영향이 없기 때문에, 더 장점이 많다.

- 주성분 분석의 기본적인 개념은 차원이 큰 벡터에서 선형 독립하는 고유 벡터만을 남겨두고 차원 축소를 하게 됩니다. 이때 상관성이 높은 독립 변수들을 N개의 선형 조합으로 만들며 변수의 개수를 요약, 압축해 내는 기법입니다. 그리고 이 압축된 각각의 독립 변수들은 선형 독립, 즉 직교하며 낮은 상관성을 보이게 됩니다. 높은 주성분들만 선택하면서 정보 설명력이 낮은, 노이즈로 구성된 칼럼들은 배제하기 때문에
- 노이즈 제거 기법이라고 불리기도 합니다.
- 선형대수학에서는 공분산행렬을 이용해 종합점수를 잘 내자!라는 의미를 가지고 있다

PCA, LDA, SVD 등의 약자들이 어떤 뜻이고 서로 어떤 관계를 가지는지 설명할 수 있나요?

- LSA(잠재의미분석) : DTM을 차원 축소해서 토픽을 묶는다
- LDA 는 토픽모델링(Topic Modeling) 기법 중 하나인 잠재디리클레할당(Latent Dirichlet Allocation, LDA)와 이니셜이 같아서 헷갈리는데 SVD (행렬분해)를 자연어처리 토픽 모델링에 적용한게 잠재디리클레할당(Latent Dirichlet Allocation, LDA)고, 또 다른 LDA (Linear Discriminant Analysis)는 PCA 에서 확장된 차원 축소 기법입니다.
- PCA(principal Component Analysis) 주성분분석: 데이터의 차원을 줄이기 위해, 공분산 행렬에서 고유 벡터/고유값을 구하고 가장 분산이 큰 방향을 가진 고유벡터에 입력데이터를 선형변환한다.
- LDA는 지도 학습(supervised - learning)에서 적용하는 차원 축소 기법이자, 입력 데이터의 클래스(정답) 를 최대한 분리할 수 있는 축을 찾는 기법입니다.
- SVD(singular Value Decomposion)특이값분해: SVD는 정사각행렬이 아닌 m*n 형태의 다양한 행렬을 분해하며, 이를 특이값 분해라 말합니다. 이때 분해되는 행렬은 두 개의 직교 행렬과 하나의 대각행렬이며, 두 직교행렬에 담긴 벡터가 특이벡터입니다.

정리



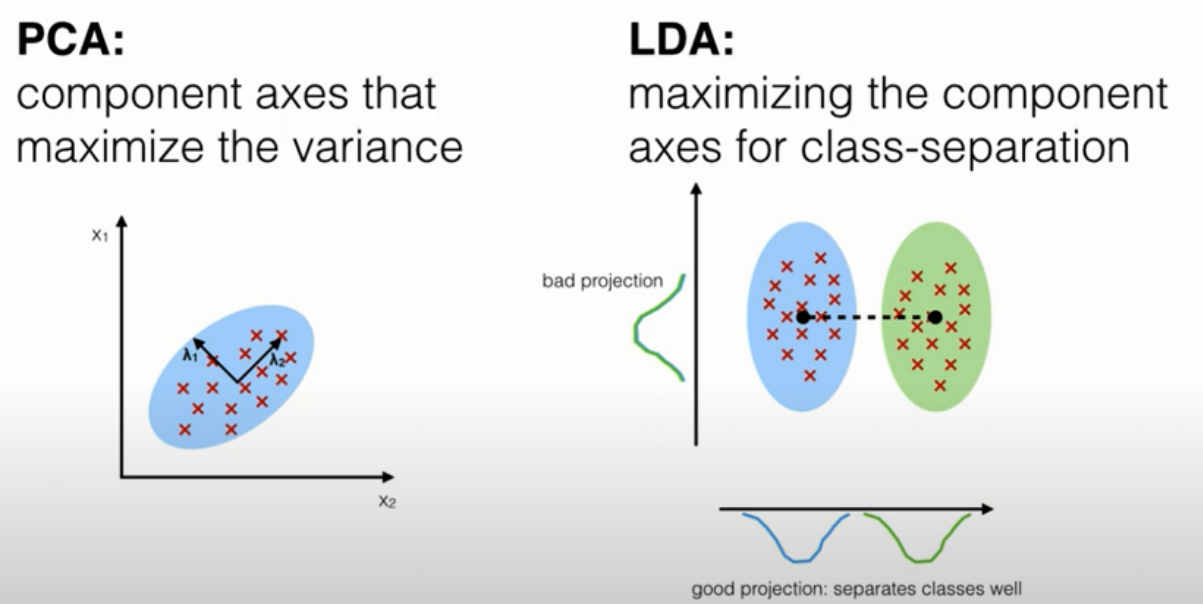
1. PCA (Unsupervised) : 분산이 최대가 되는 방향 Projection(정사영)

```
# PCA
from sklearn.decomposition import PCA
pca = PCA(n_component=2)
pca_result = pca.fit_transform(x)
```

2. LDA (Supervised) : 집단을 분류(classification)가 잘 되는 방향Projection(정사영)

- 클래스 내부, 클래스간 분산행렬 계산

```
from sklearn.discriminant_analyis import LinearDiscriminantAnalysis as LDA
# LDA
lda = LDA(n_components=2)
# LDA는 input으로 class가 들어감
lda_result = lda.fit_transform(x, y)
# 수치형 설명
lda.explained_variance_ratio
```



1. SVD (특이 값 분해)

SVD는 또한 선형 차원 감소 기술입니다.

PCA와 매우 유사하지만 특이 값 분해를 계산하기 전에 데이터를 중앙에 배치하지 않습니다. 이는 희소 행렬을 효율적으로 사용할 수 있음을 의미합니다.