#1. [와인 품질 예측 노트북]

대회 목표 : 인증 및 품질 평가와 보증 과정이 보다 통제될 수 있도록 와인의 여러 화학적 특성을 바탕으로 와인의 품질을 예측하고 분류 모델을 만드는 것이 목표

PCA : 기존 데이터를 거의 유지하면서 large 변수 세트를 smaller 변수 세트로 변환시켜 large data sets의 차원을 줄이는 기법.

step1 : Standardization
step2 : Covariance Matrix computation
step3 : Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal componenets
step4 : Feature vector
step5 : Recast the data along the principal components axes

#2. [이미지 데이터 차원 축소]
차원축소?
: 매우 많은 특징들로 구성된 다차원 데이터 세트의 차원을 감소시킴으로써 새로운 차원의 데이터 세트를 생성하는 것을 말한다. 일반적으로 차원이 증가함에 따라 데이터 점들 사이의 거리는 기하급수적으로 멀어지고 희박한 구조를 가지게 된다.
-> 수백 개 이상의 기능으로 구성된 데이터 세트의 경우 예측 신뢰도는 상대적으로 적은 차원에 대해 훈련된 모델보다 낮다.
또한 특징이 많으면 개별 feature 간의 산관 관계가 높을 가능성이 커진다.
-> 선형 회귀와 같은 선형 모형에서는 입력 변수 간의 상관관계가 높을 때 다중 공선성 문제로 인해 모형의 예측 성능이 저하된다.

차원의 저주?
주로 수치해석, 샘플링, 조합론, 머신러닝, 데이터 마이닝 및 데이터베이스와 같은 영역에서 발생한다. 이러한 문제의 공통 주제는 차원이 증가할수록 점들 사이의 공간이 너무 빠르게 증가하여 사용 가능한 데이터가 희소해진다는 것이다.
-> 예측 성능 저하로 이어질 수 있다.
신뢰할 수 있는 결과를 얻기 위해 필요한 데이터의 양은 종종 차원에 따라 기하급수적으로 증가한다. 또한 데이터를 구성하고 검색하는 것은 종종 객체가 유사한 속성을 가진 그룹을 형성하는 영역을 탐지하는 데 의존한다.
그러나 고차원 데이터에서는 모든 객체가 희박하고 여러 면에서 서로 다른 것으로 나타나기 때문에 일반적인 데이터 구성 전략이 효율적이지 못하다.

[대회]
다차원 featre의 차원 축소를 통해 featre의 개수를 줄임으로ㅆ 데이터를 보다 직관적으로 해석할 수 있는 여러 차원 축소 방법들을 요약한 노트북

MNIST 데이터 이용.(손으로 쓴 0~9까지의 숫자 이미지로 구성된 데이터셋. 784개의 pixel로 구성됨. 0에 가까울수록 어두운 색이며 255에 가까울수록 밝은 색이다.)
-> 차원 축소를 통해 이 데이터 세트를 살펴볼 때, 데이터 간 규칙성이 있는지? 각 차원 축소 방법에 따라 label이 어뚷게 clustering 되는지에 중점을 둘 예정.

PCA?
차원 축소의 가장 대표적인 방법으로 다차원 데이터에서 분산이 최대한 보조노디는 방향으로 축을 계속해서 재설정하는 방법.
I번째 축 : I번째 주성분
(i+1)번째 주성분 : I번째 주성분에 수직이고 분산이 가장 큰 방향

주성분 자체는 원본 차원과 동일하고, 주성분을 통해 변경된 데이터는 차원이 감소함.
일반적으로 주성분은 원본 특성의 개수만큼 찾을 수 있음.
변수 간의 의존성이 클수록 주성분은 원래 데이터를 나타낼 수 있음. 그러나 각 feature는 정규 분포를 따른다고 가정.
-> 왜곡된 분포를 갖는 변수를 PCA에 적용하는 것은 적절하지 않음.

Truncated SVD?
SVD : 특이값 분해 기법
Truncated SVD : sigma 행렬에서 대각 원소의 윗부분. 즉, 특이값의 상위 몇 개만을 추출하여 분해하는 방법.
simgan 행렬의 비대각 부분과 대각원소 중 특이값이 0인 부분을 모두 제거하고 제거된 sigma 행렬에 대응되는 U 행렬과 V행렬의 원소도 함께 제거해 차원을 줄인 형태로 분해
이러한 분해로 인해 A = UsigmaV.t는 더 작은 차원으로 이누이적으로 분해됨.
-> 원소의 손실로 인해 원본 행렬을 정확하게 복원할 수 없음.
그러나 데이터 정보가 압축되고 분해되었음에도 원래의 행렬을 상당한 정도로 근사하는 것은 가능.

NMF?
음수 미포함 행렬 분해 기법 -> 원본 행렬에 있는 모든 요소의 값이 양수라는 것이 보장되어야 함.
SVD와 같은 low-rank 근사법의 변형.
음수를 포함하지 않는 행렬V를 음수를 포함하지 않는 W행렬과 H 행렬의 곱으로 분해.

LDA?
지도 학습의 분류 문제에서 차원을 줄이는 방법.
-> 개별 클래스를 분별할 수 잇는 기준을 최대한 유지하며 차원을 축소.

학습 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 저차원 feature 공간을 찾고, 해당 공간에 원래 feature를 투영해 차원 축소.
-> 이후 feature들을 잘 구분할 수 있는 직선을 찾는 걸 목표로 함.
클래스 간 분산은 최대한 크게, 클래스 내부 분산은 최대한 작게 되도록 클래스를 분리함.

t-SNE?
고차원의 데이터를 압축하여 2차원 평면에 시각화하기 위해 사용.
높으 ㄴ차원 공간에서 비슷한 데이터 구조는 낮은 차원 공간에서 가깝게 대응하며 비슷하지 않은 데이터 구조는 멀리 떨어져 대응한다.
거리 측정시 t분포를 활용하고 압축 후 원래의 feature 공간에 가ᄁᆞ운 점도 2차원 평면으로 표현된다.
비선형 관계를 식별할 수 있다. -> t-SNE에 의해 표현된 압축 결과를 원래 특징에 추가하여 모델 성능을 향상시킬 수 있다.

UMAP?
비선형 차원 축소를 위해 제안된다.(t-SNE보다 빠르고 데이터 공간을 잘 분리한다.)
고차원에서 데이터를 graph로 만들고, 저차원으로의 graph projection을 수행.
1) data point에서 simplex 복합체로 만들어서 graph 구성
2) 각 node에서의 길이 k의 raidus를 그린다. 이때 k가 작으면 local structure, 크면 global structure를 가져올 수 있다.
3) 이 strength를 그대로 저차원으로 이동시키면 UMAAP이 완성된다.

=> 매우 큰 데이터 세트를 빠르게 처리할 수 있으며, 희소 행렬 데이터 처리에 적합하다.
embedding 차원 크기에 대한 제한이 X + 다른 머신러닝 모델에서 새로운 데이터가 들어오면 즉시 embedding이 가능하다. 일반적인 차원 축소 알고리즘으로 적용 가능하다
global structure를 더 잘 보존한다는 장점이 있다.

UMAP connectivity plot?
데이터가 샘플링 되었을 수 있는 대략적인 매니폴드의 중간 위상 표현을 구성함으로써 작동한다. 해당 구조를 가중 그래프로 단순화할 수 있다.
때때로 결과 embedding과 관련하여 그래프가 어뚷게 보이는지 보는 것이 유익할 수 있다.
embedding을 더 잘 이해하고 진단 목적으로 사용할 수 있다.

AutoEncoder?
신경망을 이용한 차원 축소 방식.
입력 데이터의 차원보다 작은 중간 레이어를 사용하여 입력과 동일한 값을 출력하는 신경망을 학습한다. 원본 데이터를 재현할 수 있는 저차원 표현을 학습한다.
항상 Encoder와 Decoder의 두 파트로 구현된다.
1) Encoder(인식 네트워크):입력 -> 내부 표현
2) Decoder(생성 네트워크):내부표현 -> 출력

#2. [차원축소 기법들]

PCA

- 가장 대표적인 차원축소 기법 중 하나
- 여러 변수 간에 존재하는 상관 관계를 이용해 이를 대표하는 주성분을 추출해서 차원을 축소하는 기법
- 분산이 데이터의 특성을 가장 잘 나타내는 것으로 간주하여 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축호함.
- 입력 데이터의 공분한 행렬은 고유벡터와 고유값으로 분해될 수 있으며, 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환한느 방식.

1) 입력 데이터 세트의 공분산 행렬을 생성한다.
2) 공분산 행렬의 고유벡터와 고유값을 계산한다.
3) 고유값이 가장 큰 순으로 k개만큼 고유벡터를 추출한다.
4) 고유값이 가장 큰 순으로 추출된 고유 벡터를 이용해 새롭게 입력 데이터를 변환한다.

Kernel PCA?

- 커널 트릭(선형 분류가 불가능한 데이터에 대한 처리를 위해 데이터의 차원을 증가시켜 하나의 초평면으로 분류할 수 있도록 도와주는 커널 함수)을 PCA에 적용해 차원축소를 위한 복잡한 비선형 투영을 수행한다.
- 커널 PCA로 비선형 매핑을 수행해서 고차원 공간으로 변환하고 표준 PCA로 샘플이 선형 분류기로 구분될 수 있는 저차원으로 데이터를 투영시킴. 이때 계산 비용이 매우 비싸다는 단점이 있는데, 이때 커널 트릭을 사용하여 원본 특성 공간에서 두 고차원 특성 벡터의 유사도를 계산할 수 있음.

Incremental PCA?

- SVD를 수행하기 위해선느 전체 학습 데이터셋을 메모리에 올려야 한다는 단점을 보완하기 위해 개발된 PCA 알고리즘.
- 학습 데이터셋을 미니배치로 나눈 뒤 IPCA 알고리즘에 하나의 미니배치를 입력으로 넣어준다.
- 학습 데이터셋이 클 때 유용

Sparse PCA?

데이터를 최적으로 재구성할 수 있는 희소 구성 요소 집합을 찾는 방법
희소성의 정도는 조정 가능한 파라미터이다.

ICA?

다변량 신호를 통계적으로 독립적인 하부 성분으로 분리하는 계산 방법.
가정 : source들이 서로 독립적이다.
차원 축소보다는 개별 요소를 분리할 때 사용된다.

ICA를 중심극한정리의 반대 과정으로 생각해본다면, ICA는 독립 랜덤 변수들의 조합으로 얻

어진 x에 적절한 행렬을 곱해 원래의 독립 랜덤 변수들인 source들 s를 찾는 과정이다.
-> source들이 서로 독릿벅이라는 가정을 최대한 만족할 수 있도록 하는 행렬을 찾는 것이 목적

PCA와 ICA 비교
공통점 주어진 데이터를 대표하는 기저벡터를 찾아줌.
차이점 : PCA는 속성 공간에서 직교하는 기저벡터 집합을 찾아준다.
데이터를 정사영했을 때, 최대 분산을 얻을 수 있는 벡터를 차례대로 기저벡터로 삼는다.
ICA를 통해 찾은 기저벡터들은 서로 직교하지 않을 수도 있다. ICA를 통해 얻은 기저벡터들의 데이터를 정사영했을 때 그 결과들이 최대한 독립적일 수 있도록 하는 벡터들을 기저벡터로 삼는다.

Factor Analysis?
수많은 변수들 중에서 잠재된 몇 개의 변수를 찾아내는 것.
가정1) 데이터에는 특이치가 없다.
2) 표본 크기는 요인보다 커야 한다.
3) 완벽한 다중공선성은 없어야 한다.
4) 변수들 사이에 동질성이 있어서는 안된다.

t-SNE?
비선형 관계를 이용한 데이터셋 분해 방법
차원이 감소되어 군집화된 데이터들이 뭉개져 제대로 구별할 수 없는 PCA의 문제점 해결
고차원 공간에서 점 세트를 가져와 저차원 공간에서 해당 점의 표현을 찾는 것이 목적
정규분포 대신 t분포를 이용함.
탁월한 성능을 가졌지만, 많은 시간을 소요한다는 단점이 존재한다.