#1. 다양한 차원축소 기법들

차원 축소란 방대한 차원을 가진 데이터 세트를 더 적은 수의 차원을 가진 데이서 세트로 변환하는 과정을 말한다. 이 프로세스는 원래 데이터 세트에 의해 전달된정보가 손실되지 않도록 보장함으로써 수행됨!

[1] : Dataset Preparation
필요한 데이터 세트를 로드하고 target 변수를 분리하여 원래 데이터 집합에서 제거한다.
그리고 모든 값이 동일한 범위에 있도록 데이터 집합을 표준화한다.

[2] : Feature Statistics
평균, 분산, 표준 편차와 같은 피쳐에 대한 기본 통계량을 계산하면 피쳐를 이해하는데 도움이 될 수 있다.

cf. 분산이란?
분산운 데이터가 평균으로부터 어떻게 퍼져있는지를 정의한다. 변수의 평균 값에서 모든 값의 차이의 제곱을 취하여 계산된다. 통계적 의도 중 하나는 형상 분산이 매우 작으면 형상이 모형에 더 적은 기여를 한다는 것이다. 그러나 대부분의 딥러닝 및 부스팅 모델은 그러한 문제에 강하기 때문에 이것을 맹목적으로 따르지는 않는다. 그러나 분산은 버릴 수 있는 features에 대한 아이디얼르 제공한다.

cf. Correlation with Target Variable?
Pearson의 상관 계수는 두 연속형 변수 사이의 통계적 관계또는 연관성을 측정하는 검정 통계량임. 상관 관계가 높은 feature는 모형에 적합하다.

[3] : Decomposition into EigenVectors and EigenValues
선형대수학에서 선형변환의 고유벡터는 선형변환이 적용될 때, 스칼라 계수만큼만 변하는 0이 아닌 벡터임.

[4] : PCA-Principal Component Analysis
주성분 분석은 고차원 데이터 세트의 대부분의 유용한 벡터를 찾는 기술이다. 즉, PCA는 많은 수의 기능을 포함하는 데이터 세트에서 구성 요소의 형태로 중요한 변수를 추출한다. 중요한 기능은 데이터 세트에서 가능한 최대 정보를 캡처하는 것을 목표로 추출한다.

첫 번째 주성분은 최대 분산을 갖는 데이터 세트 기능의 선형 결합이다. 데이터의 변동성이 가장 높은 방향을 결정한다. 성분의 상관관계가 없으면 방향이 직교해야 한다. 이것은 성분들의 상관관계가 0이라는 것을 시사한다. 모든 후속 주성분은 유사한 개념을 따른다. 즉, 이전 성분과 상관없이 나머지 변동을 포착한다.

*PCA Variants
: sklearn은 도움이 될 수 있는 다양한 PCA 변형을 제공한다.

(1) Kernel PCA : 커널 PCA는 커널을 사용하여 비선형 차원 축소를 달성하는 PCA의 확장이다. 노이즈 제거, 압축 및 구조화 예측(커널 종속성 추정)을 포함한 많은 응용 프로그램을 가지고 있다.

(2) Incremental PCA : Incremental PCA는 PCA와 유사하게 작동하지만 입력 데이터의 크기에 따라 Incremental PCA가 훨씬 더 메모리 효율적입니다. 이 기술은 미니 배치 방식으로 데이터를 처리하는 동안 PCA의 결과와 거의 정확하게 일치하는 부분 연산을 허용합니다.

(3) Sparse PCA : Sparse PCA는 데이터를 최적으로 재구성할 수 있는 희소 구성 요소 집합을 찾는다. 희소성의 양은 조정 가능한 매개 변수입니다.

(4) Mini Batch Sparse PCA : Mini Batch Sparse PCA는 Sparse PCA와 유사하지만 데이터에서 한 번에 미니 배치를 취하여 구성 요소를 계산합니다. 그것은 더 빠르지만 정확성은 떨어진다.

[5] Truncated SVD
특이값 분해(SVD)는 행렬을 구성 부분으로 축소하는 행렬 분해 방법이다. Truncated SVD는 치수 감수에도 사용되는 SVD의 변형이다. PCA와는 달리 이 estimator는 특이치 분해를 계산하기 전에 데이터를 중심에 두지 않는다. 즉, sparse 행렬과 함께 매우 효율적으로 작동할 수 있다.

[6] Independent Component Analysis - ICA
Independent component analysis은 다변량 특징을 포함하는 데이터 세트를 최대 독립성이 있는 추가 하위 성분으로 분리한다. 일반적으로 ICA는 차수 감소를 위해 사용되는 것이 아니라 개별 구성 요소를 분리하기 위해 사용된다.

[7] Factor Analysis
가우스 잠재 변수를 가진 간단한 선형 생성 모델이다.

[8] Non Negative Matrix Factorization
NMF는 두 개의 음이 아닌 행렬을 찾는 데 사용되는 기법이다.

[9] Gaussian Random Projection

[10] Sparse Random Projection

[11] t-SNE
t-SNE) t-분산 확률적 이웃 임베딩은 고차원 데이터를 탐색하는 데 사용되는 비선형 차원 감소 알고리즘이다. 다차원 데이터를 인간의 관찰에 적합한 2개 이상의 차원에 매핑한다. t-SNE는 데이터 내의 구조를 찾기 위해 이웃 그래프에서 무작위 보행으로 확률 분포를 기반으로 한다. 목표는 고차원 공간에서 점들의 집합을 취하여 저차원 공간, 전형적으로 2D 평면

에서의 점들의 표현을 찾는 것이다.

[12] Baseline Model with Decomposed Features

2. 이미지 데이터 차원축소 mnist 예제

<The Curse of Dimensionality>
수치 분석, 샘플링, 조합론, 머신러닝, 데이터 마이닝 및 데이터베이스와 같은 영역에서 차원 저주받은 현상이 발생한다. 이러한 문제의 공통 주제는 차원이 증가하면 공간의 부피가 너무 빠르게 증가하여 사용가능한 데이터가 희박해진다는 것이다. 신뢰할 수 있는 결과를 얻기 위해 데이터의 양은 종종 차원에 따라 기하급수적으로 증가한다. 또한 데이터를 구성하고 검색하는 것은 종종 객체가 ㅇ사한 속성을 가진 그룹을 형성하는 영역을 탐지하는 데 의존한다. 그러나 고차원 데이터에서는 모든 rorpc가 희박하고 여러 면에서 서로 다른 것으로 나타나기 때문에 일반적인 데이터 구성 전략이 효율적이지 못하다.

<차원 축소>
치수 축소는 매우 많은 형상으로 구성된 다차원 데이터 세트의 치수를 축소하여 새로운 치수 데이터 세트를 생성하는 것이다. 일반적으로 차원이 증가함에 따라 데이터 점 사이의 거리는 기하급수적으로 멀어지고 희박한 구조를 갖는다. 수백 개 이상의 기능으로 구성된 데이터 세트의 경우, 예측 신뢰도는 상대적으로 적은 차원에 대해 훈련된 모델보다 낮다. 또한 특징이 많으면 개별 특징 간의 상관관계가 높을 가능성이 높다. 선형 회귀와 같은 선형 모형에서는 입력 변수 간의 상관 관계가 높을 때 다중 공선성 문제로 인해 모형의 예측 성능이 감소합니다.

<PCA>
PCA는 치수 감소의 가장 대표적인 방법이다. 이것은 다차원 데이터를 큰 분산 방향으로 재축하는 방법입니다. 변수 간의 의존성이 클수록 주성분은 원래 데이터를 나타낼 수 있습니다. 그러나 각 형상은 정규 분포를 따른다고 가정하므로 분포가 왜곡된 변수를 PCA에 적용하는 것은 적절하지 않다.

<Truncated SVD>
Truncated SVD는 시그마 행렬에서 대각선 원소의 상부, 즉 특이값의 상부만을 추출하여 분해하는 방법이다. 이러한 분해로 인해 원래 행렬은 더 작은 차원으로 인위적으로 분해하기 때문에 정확하게 복원할 수 없다. 그러나 데이터 정보가 압축 및 분해되고 있음에도 불구하고 원본 행렬을 상당한 수준으로 근사하는 것은 가능하다.

<NMF>
NMF는 SVD와 같은 Low-Rank 근사법의 변형이다. 그러나 소스 행렬에 있는 모든 요소의 값이 양수라는 것을 보장해야 한다. NMF는 행렬을 W 행렬과 H 행렬로 분해한다. W 행렬은 잠재 요소의 값이 소스 행렬에 얼마나 잘 해당하는지를 나타냅니다. H 행렬은 이 잠재 원소가 sour features로 어떻게 구성되어 있는지를 나타낸다.

<LDA>
LDA는 지도 학습의 분류 문제에서 차원을 줄이는 방법이다. 훈련 데이터를 잘 분류할 수 있는 저차원 형상 공간을 찾고, 해당 공간에 원래 형상을 투영해 차원을 줄인다.

<t-SNE>
t-SNE는 종종 2차원 평면에서 데이터를 압축하여 시각화 목적으로 사용된다. 원래 형상 공간에 가까운 점도 압축 후 2차원 평면으로 표현된다. 비선형 관계를 식별할 수 있기 때문에 t-SNE에 의해 표현된 압축 결과를 원래 특징에 추가하여 모델 성능을 향상시킬 수 있다. 그러나, 연산 비용이 높기 때문에, 2차원 또는 3차원을 초과하는 압축에는 적합하지 않다.

<UMAP>
비선형 차원 축소를 위해 t-SNE보다 빠르고 데이터 공간을 잘 분리한다. 즉, 매우 큰 데이터 세트를 빠르게 처리할 수 있으며 희소 행렬 데이터에 적합하다. 또한, t-SNE에 비해 다른 머신러닝 모델에서 새로운 데이터가 들어오면 즉시 임베딩이 가능하다는 장점이 있다.

<UMAP connectivity plot>
UMAP는 데이터가 샘플링되었을 수 있는 대략적인 매니폴드의 중간 위상 표현을 구성함으로써 작동한다. 실제로 이 구조는 가중 그래프로 단순화할 수 있다. 때때로 결과 임베딩과 관련하여 그래프(다양체에서 연결성을 나타냄)가 어떻게 보이는지 보는 것이 유익할 수 있다. 임베딩을 더 잘 이해하고 진단 목적으로 사용할 수 있다.