비지도학습 - PCA

목차

1-1 비지도 학습

- 비지도 변환
- 군집(Clustering)

1-2 차원 축소

▶ 차원 축소

- 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소하여 새로운 차원의 데이터 세트를 생성하는 것.
- 차원이 증가할수록 데이터 포인터 간의 거리가 기하급수적으로 멀어지며, 희소한(sparse)한 구조를 갖게 된다.
- 피처가 많을 경우 개별 피처 간에 상관 관계가 높을 가능성이 큽니다.
- 선형 회귀와 같은 모델은 입력 변수 간의 상관관계가 높을 경우, 이로 인한 다중 공선성 문제로 모델의 예측 성능이 저하된다.

1-2 차원 축소

▶ 차원 축소 효과

- 많은 다차원의 피처를 차원 축소하여 피처 수를 줄이면 더 직관적으로 데이터 해석이 가능해짐.
- 노이즈 제거(Reduce Noise) – 쓸모없는 피처를 제거함으로 노이즈 제거가 가능해짐.
- 메모리 절약 – 쓸모없는 피처를 제거하여 메모리 절약이 가능해짐.
- 퍼포먼스 향상 – 불필요한 피처들을 제거하여 모델 성능 향상에 기여.

1-2 차원 축소

▶ 차원 축소 방법

- **피처 선택**(feature selection)과 **피처 추출**(feature extraction)으로 나눌 수 있음.

- **피처 선택**(feature selection) : 특정 피처에 종속성이 강한 불필요한 특징은 제거하고, 데이터를 잘 나타내는 특징만 선택하기

- **피처 추출**(feature extraction) : 기존 피처를 저차원의 중요 피처로 압축해서 추출.

1-2 차원 축소

▶ 특징 추출(feature extraction)

- 기존 피처의 단순 압축이 아닌, 특징을 함축적으로 더 잘 설명할 수 있는 다른 공간으로 매핑해 추출하기

1-2 차원 축소

▶ 대표적인 알고리즘

- PCA, SVD, NMF 등이 있음

1-2 차원 축소

▶ 사용되는 분야

(A) 매우 많은 픽셀로 이뤄진 이미지 데이터에서 잠재된 특징을 피처로 도출.

(B) 텍스트 문서의 숨겨진 의미를 추출.

(C) 추천 시스템에서의 차원 축소 활용

1-3 PCA

▶ PCA(Principal Component Analysis)

(A) 대표적인 차원 축소 기법

(B) 여러 변수 간에 존재하는 상관관계를 이용하여 대표하는 주성분 추출하여 차원 축소

(C) 가장 높은 분산을 갖는 데이터의 축을 찾아 이 축으로 차원을 축소.
이 차원이 주성분이 됨.

1-4 PCA – 주성분 찾기

▶ PCA(Principal Component Analysis)

(A) 데이터를 가장 잘 표현하는 분산이 가장 큰 성분을 찾습니다.(직선, 평면 등)

(B) 우리가 찾은 이것이 첫번째 주성분이 됩니다.

(C) 두번째 주성분은 동일한 방법으로 찾습니다.
단 조건은 PC1에 수직인 직선이어야 합니다.

(D) 우리가 찾은 주성분에 데이터를 각각 사상-Projection 시킵니다.

(E) PC1이 x축에 수평이 되도록 데이터를 회전시킵니다.

(F) 마지막으로 사상된 점을 중심으로 데이터를 복원시킵니다.

1-5 PCA – Variation

▶ 주성분의 설명력

(A) SS를 n-1로 나누어 variation으로 구할 수 있다.

(B) PC1, PC2의 분산

$$\frac{SS(distances\ for\ PC1)}{n-1} = \text{Variation for PC1}$$

$$\frac{SS(distances\ for\ PC2)}{n-1} = \text{Variation for PC2}$$

1-5 PCA – Variation

▶ 주성분의 설명력

$$\frac{SS(distances\ for\ PC1)}{n-1} = \text{Variation for PC1}$$

$$\frac{SS(distances\ for\ PC2)}{n-1} = \text{Variation for PC2}$$

(A) PC1의 분산이 15, PC2의 분산이 3일 때, PC1은 $\frac{15}{18} = 0.83 = 83\%$

(B) PC1의 분산이 15, PC2의 분산이 3일 때, PC2은 $\frac{3}{18} = 0.17 = 17\%$

1-5 PCA – Variation

▶ 주성분의 설명력

(A) PC1~PC4 분산이 15, 5, 3, 1일 때, PC1은 $\frac{15}{24} = 0.625 = 62.5\%$

(B) PC1~PC4 분산이 15, 5, 3, 1일 때, PC2은 $\frac{5}{24} = 0.21 = 21\%$

(C) PC1~PC4 분산이 15, 5, 3, 1일 때, PC3은 $\frac{3}{24} = 0.125 = 12.5\%$

(D) PC1~PC4 분산이 15, 5, 3, 1일 때, PC4은 $\frac{1}{24} = 0.04 = 4\%$

우리는 PC1과 PC2를 주성분을 가지고 전체 데이터의 83.5%를 설명할 수 있다.