chapter 3. 비지도 학습과 데이터 전처리

비지도 변환 : 데이터를 새롭게 표현하여 사람이나 다른 머신러닝 알고리즘이 원래 데이터보다 쉽게 해석할 수 있도록 만드는 알고리즘.

비지도변환이 사용되는 분야 – 차원 축소(특성이 많은 고차원 데이터를 특성의 수를 줄이면서 꼭 필요한 특징을 포함한 데이터로 표현하는 방법) ex) 시각화를 위해 데이터셋을 2차원으로 변경, 주제를 찾기.

군집 알고리즘 : 데이터를 비슷한 것끼리 그룹으로 묶는 것.

3.3 데이터 전처리와 스케일 조정

데이터의 특성 값 조정 –특성마다 스케일을 조정해서 데이터 변경

StandardScaler : 각 특성의 평균 – 0, 분산 – 1

RobustScaler : 특성들이 같은 스케일을 갖게 된다는 통계적 측면에서 StandardScaler와 비슷, 평균과 분산 대신 중간값과 사분위 값 사용 (전체 데이터와 동떨어진 데이터 포인트에 영향받지x )->이상치

MinMaxScaler : 모든 특성이 정확히 0과1 사이에 위치하도록 데이터 변경

Nomalizer : 매우 다른 스케일 조정 기법 / 특성 벡터의 유클리디안 길이가 1이 되도록 데이터 포인트 조정(지름이 1인 원에 데이터 포인트 투영) 각 데이터 포인트가 다른 비율로 스케일이 조정된다. – 벡터의 길이는 상관없고 데이터 방향만 중요할 때 사용

데이터 전처리 – MinMaxScaler 사용

Transform 메서드(훈련 데이터의 스케일 조정) – 테스트 세트의 최소값과 범위를 사용x, 항상 훈련 세트의 최솟값을 빼고 훈련 세트의 범위로 나눈다.

지도학습 모델에서 테스트 세트 사용시, 훈련 세트와 테스트 세트에 같은 변환 적용해야함.

3.4 차원 축소, 특성 추출, 매니폴드 학습

주성분 분석(PCA) : 특성들이 통계적으로 상관관계가 없도록 데이터셋을 회전시키는 기술

먼저, 분산이 가장 큰 방향 찾기(데이터에서 가장 많은 정보를 담고있다. / 특성들의 상관관계가 가장 크다.), 그 다음 첫 번째 방향과 직각인 방향 중 가장 많은 정보를 담은 방향 찾기

이 과정을 거쳐 찾은 방향 -> 데이터에 있는 주된 분산의 방향(주성분) / 일반적으로 원본 특성 개수만큼 주성분o

PCA가 널리 사용되는 분야 – 고차원 데이터셋의 시각화

PCA 적용 – PCA 객체 생성, fit 매서드 호출해 주성분 찾기, transform 메서드 호출해 데이터 회전시키고 차원 축소

기본값일 때 PCA는 데이터를 회전만 시키고 모든 주성분 유지

PCA 단점 – 그래프의 두 축을 해석하기 쉽지 않다.

특성 추출 – 원본 데이터 표현보다 분석하기 더 적합한 표현 찾기

PCA의 화이트닝 옵션 – 주성분의 스케일이 같아지도록 조정(변환 후 StandardScaler를 적용한 것과 같다)

PCA 변환 – 테스트 포인트를 주성분의 가중치 합으로 나타내는데 필요한 수치를 찾는 것.

원래 특성 공간으로 되돌리는 작업 – inverse\_transform

비음수 행렬 분해(NMF) : 유용한 특성을 뽑아내기 위한 또 다른 비지도 학습 알고리즘, PCA와 비슷, 차원 축소에도 사용o, 음수가 아닌 성분과 계수 값을 찾는다.(주성분과 계수가 모두 0보다 크거나 같아야함.)

여러 사람의 목소리가 담긴 오디오 트랙이나 여러 악기로 이뤄진 음악처럼 독립된 소스를 추가하여 만들어진 데이터에 특히 유용 – 섞여 있는 데이터에서 원본 성분 구분

NMF는 무작위로 초기화하기 때문에 난수 생성 초깃값에 따라 결과가 달라진다./ 데이터에 있는 유용한 패턴을 찾는데 활용

성분들이 모두 양수 값이어서 PCA 성분보다 원형처럼 보인다.

t-SNE를 이용한 매니폴드 학습

매니폴드 학습 알고리즘은 훨씬 복잡한 매핑을 만들어 더 나은 시각화를 제공한다.

일부 매니폴드 알고리즘들은 훈련 데이터를 새로운 표현으로 변환시키지만 새로운 데이터에는 적용하지 못한다. 그래서 지도학습용으로는 거의 사용하지 않는다.

t-SNE는 데이터 포인트 사이의 거리를 가장 잘 보존하는 2차원 표현을 찾는 것

각 데이터 포인터를 2차원에 무작위로 표현한 후 원본 특성 공간에서 가까운 포인트는 가깝게, 멀리 있는 포인트는 멀어지게 만든다. 멀리 떨어진 포인트와 거리를 보존하는 것보다 가까이에 있는 포인트에 더 비중을 둔다.

처음 두 개의 주성분으로 숫자 데이터를 변환하고 숫자 텍스트를 이용해 산점도를 그린다.

t-SNE로 찾은 두개의 성분을 사용한 숫자 데이터의 산점도는 PCA를 사용한 것 보다 모든 클래스가 잘 구분되었다.

t-SNE는 클래스 레이블 정보를 사용하지 않으므로 비지도학습이다.

군집 : 데이터 셋을 클러스터라는 그룹으로 나누는 작업이다.

한 클러스터 안의 데이터 포인트끼리는 매우 비슷하고 다른 클러스터의 데이터 포인트와는 구분되도록 데이터를 나눈다.

k-평균 군집은 1)데이터의 어떤 영역을 대표하는 클러스터 중심을 찾는다.

2) 데이터 포인트를 가장 가까운 클러스터 중심에 할당하고 3) 클러스터에 할당된 데이터 포인트의 평균으로 클러스터 중심을 다시 지정한다./ 클러스터에 할당되는 데이터 포인트에 변화가 없을 때 알고리즘이 종료된다.

클러스터 중심에 할당되는 포인트에 변화가 없으면 알고리즘이 멈춘다.

코드로 나타내기 -> KMeans(K평균 알고리즘)의 객체를 생성하고 찾고자 하는 클러스터의 수를 지정한다. 그 다음 fit 메서드를 호출한다. 알고리즘을 적용하면 X에 담긴 각 훈련 데이터 포인트에 클러스터 레이블이 할당된다.

Predict 메서드를 사용해 새로운 데이터의 클러스터 레이블을 예측할 수 있다.

군집은 각 데이터 포인트가 레이블을 가진다는 면에서 분류와 비슷하다.

차이점은 군집은 정답을 모르며 레이블 자체에 의미가 없다.(알고리즘이 찾은 클러스터 3이라고 레이블된 얼굴은 비슷)

벡터 양자화 또는 분해 메서드로서의 k-평균

k-평균 – 클러스터 중심으로 각 데이터 포인트를 표현 즉, 하나의 성분으로 표현

벡터 양자화 : k-평균을 각 포인트가 하나의 성분으로 분해되는 관점으로 보는 것

k-평균을 사용한 벡터 양자화는 입력 데이터의 차원보다 더 많은 클러스터를 사용해 데이터를 인코딩할 수 있다.

k-평균은 이해하기 쉽고 구현도 쉽다. 비교적 빠르다. 단점은 무작위 초기화를 사용하여 알고리즘의 출력이 난수 초깃값에 따라 달라진다, 클러스터의 모양을 가정하고 있어 활용 범위가 제한적이다, 찾으려 하는 클러스터의 개수를 지정해야 한다.

병합군집

병합 군집 알고리즘은 시작할 때 각 포인트를 하나의 클러스터로 지정하고, 그 다음 어떤 종료 조건을 만족할 때까지 가장 비슷한 두 클러스터를 합쳐나간다. 이러한 군집 알고리즘의 모음

Scikit-learn에서 사용하는 종료조건은 클러스터 개수, 클러스터가 남을때까지 비슷한 클러스터를 합친다.

Scikit-learn에서 구현된 옵션

Ward : 모든 클러스터 내의 분산을 가장 작게 증가시키는 두 클러스터를 합친다.

Average : 클러스터 포인트 사이의 평균 거리가 가장 짧은 두 클러스터를 합친다.

Complete : 연결은 클러스터 포인트 사이의 최대 거리가 가장 짧은 두 클러스터를 합친다.

병합 군집은 새로운 데이터 포인트에 대해 예측을 할 수 없다. Predict 메서드가 없다.

계층적 군집과 덴드로그램

병합 군집은 계층적군집을 만든다. 군집이 반복하여 진행되면 모든 포인트는 하나의 포인트를 가진 클러스터에서 시작하여 마지막 클러스터까지 이동하게 된다. 중간단계는 데이터에 대한 클러스터를 생성한다. 그림 3-35 – 그림 3-33에 나타난 모든 클러스터를 겹쳐서 표현 , 계층적 군집의 모습 / 특성이 셋 이상인 데이터셋에는 사용 불가능

->이를 보완 : 다차원 데이터 셋을 처리할 수 있는 덴드로그램

덴드로그램은 SciPy(사이파이)를 사용해 만들 수 있다.

Scipy(사이파이)는 데이터 배열 X를 받아 계층 군집의 유사도가 들어있는 연결 배열을 반환하는 함수를 제공한다.

연결 배열을 Scipy의 덴드로그램 함수에 넣어 덴드로그램 그래프를 그릴 수 있다.

3.5.3 DBSCAN – 군집 알고리즘

장점 : 클러스터의 개수를 미리 지정할 필요x / 비교적 큰 데이터셋에도 적용o

밀집 지역 가까이 있는 데이터가 많아 붐비는 지역의 포인트를 찾는다.

데이터의 밀집 지역이 한 클러스터를 구성하며 비교적 비어있는 지역을 경계로 다른 클러스터와 구분된다.

밀집 지역에 있는 포인트 – 핵심 샘플

한 데이터 포인트에서 eps 거리 안에 데이터가 min\_samples 개수만큼 들어 있으면 이 데이터 포인트를 핵심 샘플로 분류

eps보다 가까운 핵심 샘플은 DBSCAN에 의해 동일 클러스터로 합쳐진다.

무작위 포인트 선택, eps 거리 안의 모든 포인트 찾기

잡음 – eps 거리 안의 포인트 수가 min\_samples 보다 적다면 어디에도 속하지 x

Eps 거리 안의 min\_samples 보다 포인트 수가 많다면 그 포인트는 핵심 샘플

Eps 거리 안에 더 이상 핵심 샘플이 없을 때까지 자라남.

핵심포인트, 경계포인트(핵심 포인트에서 eps 거리 안에 있는 포인트), 잡음 포인트

경계포인트가 어떤 클러스터에 속할 지 포인트 방문 순서에 따라 달라짐.

Min\_samples를 키우면 핵심 포인트 수가 줄어들며 잡음 포인트 증가

클러스터의 최소 크기 결정 – min\_samples

타깃값으로 군집 평가하기 -> 1과 0사이의 값을 제공 – ARI(Adjusted Rand Index), NMI

ARI 사용시 문제점 : 군집 알고리즘을 적용할 때 그 결과와 비교할 타깃값이 없다.

타깃값이 필요 없는 군집용 지표 – 실루엣 계수

실루엣 계수는 클러스터의 밀집 정도를 계산, 높을수록 좋으며 최대 점수는 1이다. / 모양이 복잡할때는 잘 맞지 않다.

실루엣 점수를 사용할 경우 결과가 더 낫다., k-평균의 실루엣 점수가 높다.

얼굴 데이터셋으로 군집 알고리즘 비교

PCA로 생성한 100개의 주성분, 즉 고유얼굴을 입력데이터로 사용

LFW 데이터에서 고유얼굴을 찾은 다음 데이터 변환

레이블이 -1뿐인 것 – 모든 데이터가 DBSCAN에 의해 잡음 포인트로 레이블 되었다.

Eps 값을 크게 하여 각 포인트의 이웃을 늘릴 수 있고, min\_samples 값을 낮추어 클러스터에 모을 포인트 수를 줄일 수 있다.

Eps가 작으면 모든 포인트가 잡음으로 레이블이 된다. Eps를 크게 하면 할수록 큰 클러스터 한 개한 여러 클러스터가 생긴다.

DBSCAN- 하나의 큰 클러스터만 만들 수 있다. 클러스터에 할당되지 않은 잡음 포인트를 인식할 수 있으며, 클러스터의 개수를 자동으로 결정한다.

병합 군집과 k-평균은 비슷한 크기의 클러스터를 만들 수 있지만 클러스터 개수를 지정해야한다.