7/6 강의노트

스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

k-means 군집 알고리즘 : 임의의 클러스터 중심을 기준으로 군집화하여 그 각각의 군집의 평균으로 새로운 클러스터 중심을 만듦, 군집이 바뀌지 않을 때까지 반복

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

Kmeans의 단점 : 무작위로 찍히는 클러스터의 위치에 따라 성능이 달라지는 경우가 있음

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

각 레이블에는 어떠한 의미도 없기 때문에 그 레이블이 어떤 의미인지 알려면 직접 확인해야 함

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

각 샘플의 개수 확인

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

클러스터 시각화 : 넘파이 ceil = 올림 함수

Subplot의 Squeeze = False는 자동전처리를 막는 패러미터, axis("off")는 축 표기 X (이미지이므로)

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

클러스터 중심의 시각화

텍스트, 과일, 스크린샷, 베리이(가) 표시된 사진

자동 생성된 설명

특정 값의 각 클러스터 중심까지의 거리 계산 및 군집 예측

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

클러스터 계산 횟수 조회

텍스트, 폰트, 화이트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

일반적으로 적절한 k값은 알 수 없음 적합한 k 값을 찾기 위해서는 elbow 기법을 이용하여 이너셔 값 개선 효율이 감소하기 시작하는 지점을 찾아냄, 코드 자체는 반복 작업에 가까움

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

라인, 그래프, 도표, 텍스트이(가) 표시된 사진

자동 생성된 설명

주성분분석

분석할 컬럼(차원)이 늘어나면 속도가 느려질 뿐 아니라 최적의 솔루션을 찾을 수 있는 범위가 줄어들게 됨 (차원의 저주)

텍스트, 폰트, 영수증, 화이트이(가) 표시된 사진

자동 생성된 설명

텍스트, 도표, 라인, 지도이(가) 표시된 사진

자동 생성된 설명

데이터에 있는 분산이 큰 방향을 찾으면 그것이 데이터의 분포를 가장 잘 설명하는 선이 됨

하나의 컬럼(축)을 무시해버리면 데이터를 올바르게 요약하지 않기 때문에, 데이터의 분포를 잘 설명하는 선을 기준으로 (새로운 축) 데이터를 모아주면 효율적으로 차원을 축소할 수 있음

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

주성분을 찾고, 특정 개수의 피처로 줄이는 모델 (50개)

스크린샷, 패턴이(가) 표시된 사진

자동 생성된 설명

주성분의 시각화 : 1번째로 갈수록 데이터의 특성을 가장 잘 설명하는 형태임

스크린샷, 텍스트, 디자인이(가) 표시된 사진

자동 생성된 설명

주성분 분석으로 줄였던 차원을 다시 원본에 가깝게 재구성 가능, 이 경우 어느 정도의 데이터 손실이 있음

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

총 분산 비율 : 50개의 주성분으로 원본 데이터의 약 92% 정도를 설명할 수 있다

텍스트, 스크린샷, 디스플레이, 라인이(가) 표시된 사진

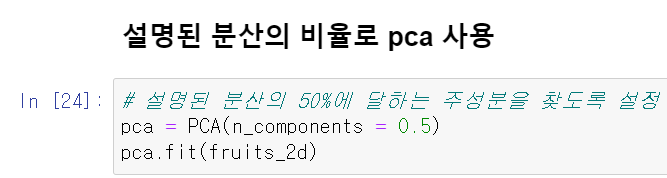
자동 생성된 설명

주성분 10개까지는 원본데이터에 기여하는 정도가 큼

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

로지스틱 회귀로 차원축소 검증, 계산 시간이 크게 감소하고 일반화 성능도 보존됨



n\_components를 0~1의 실수로 지정하면 해당 비율만큼의 데이터를 설명할 수 있는 주성분 개수로 자동 지정됨

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

주성분 개수는 n\_components\_로 조회 가능

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

성능 검증

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

차원 축소데이터로 kmeans 사용시 성능