기계학습[11w] ch9.비지도 학습

군집

-비슷한 샘플을 클러스터로 모음

이상치 탐지

-'정상' 데이터가 어떻게 보이는지를 학습하고, 비정상 샘플을 감지하는 데 사용

밀도 추정

-데이터셋 생성 확률 과정의 확률 밀도 함수를 추정함

-밀도 추정은 이상치 탐지에 널리 사용됨

-밀도가 매우 낮은 영역에 놓인 샘플이 이상치일 가능성 높음

9.1 군집

각 샘플은 하나의 그룹에 할당

비슷한 샘플을 구별해 하나의 클러스터 또는 비슷한 샘플의 그룹으로 할당하는 작업

-군집을 사용하는 다양한 어플리케이션

고객분류(추천시스템), 데이터 분석, 차원 축소 기법, 이상치 탐지,준지도 학습, 검색 엔진, 이미지 분할..

9.1.1 k-평균

(로이드-포지 알고리즘)

:반복 몇번으로 레이블이 없는 데이터셋을 빠르고 효율적으로 클러스터로 묶는 간단한 알고리즘

-예제) 샘플 덩어리 다섯 개로 이루어진 데이터셋

-(사이킷런) k-평균 알고리즘 적용

ㄴ각 군집의 중심을 찾고 가장 가까운 군집에 샘플 할당

ㄴ군집수 지정해야 함

결정 경계

ㄴ결과: 보로노이 다이어그램

(평면을 특정 점까지의 거리가 가장 가까운 점의 집합으로 분할한 그림)

ㄴ경계 부분의 일부 샘플을 제외하고 기본적으로 군집이 잘 구성됨

k-평균 알고리즘의 단점

ㄴ군집의 크기가 서로 많이 다르면 잘 작동되지 않음

(샘플과 센트로이드까지의 거리만 고려되기 때문)

하드 군집: 각 샘플에 대해 가장 가까운 클러스터를 선택

소프트군집: 클러스터마다 샘플에 점수를 부여함 샘플별로 각 군집 센트로이드와의 거리를 측정

k-평균 알고리즘

ㄴ처음에는 센트로이드를 랜덤하게 선정

ㄴ수렴할 때까지 다음 과정 반복

(각 샘플을 가장 가까운 센트로이드에 할당)

(군집별로 샘플의 평균을 계산하여 새로운 센트로이드 지정)

k-평균 - 센트로이드 초기화 방법

-관성

ㄴk-mean 모델 평가 방법

ㄴ정의: 샘플과 가장 가까운 센트로이드와의 거리의 제곱의 합

ㄴ각 군집이 센트로이드에 얼마나 가까이 모여있는가를 측정

ㄴscore() 메서드가 측정 (음수 기준)

-좋은 모델 선택법

-K-평균++

ㄴ센트로이드를 무작위로 초기화하는 대신 특정 확률분포를 이용하여 선택

ㄴ센트로이드들 사이의 거리를 크게할 가능성이 높아짐

ㄴKMeans 모델의 기본값으로 사용됨

k-평균 (K-평균 속도 개선과 미니배치 k-평균)

-elkan 알고리즘

ㄴalgorithm=elkan: 불필요한 거리 계산을 많이 피함으로, 학습 속도 향상됨

ㄴ삼각 부등식 사용 algorithm=full:

-미니배치 K-평균

ㄴ전체 데이터셋 대신 각 반복마다 미니배치를 사용해 센트로이드를 조금씩 이동함

ㄴ미니배치를 지원하는 K-평균 알고리즘: MiniBatchMeans

ㄴmemmap 활용

(대용량 훈련 세트 활용하고자 할 경우)

(점진적 PCA에서 사용했던 기법과 동일)

ㄴmemmap 활용이 불가능할 정도로 큰 데이터셋을 다뤄야 하는 경우

(MiniBatchKMeans의 partial\_fit() 메서드 활용)

(미니배치로 쪼개어 학습)

ㄴ미니배치 K-평균의 특징

(K-평균 알고리즘 보다 훨씬 빠름)

(이너셔는 일반적으로 조금 더 나쁨)

(군집수가 증가할 때 이너셔는 더 나쁨)

k-평균 (최적의 클러스터 개수 찾기)

-> 최적의 군집수를 사용하지 않으면 적절하지 못한 모델을 학습할 수 있음

-관성과 클러스터

ㄴ클러스터 개수 k가 증가할 수록 관성(inertia)이 작아지므로, 좋은 성능 지표가 아님

ㄴ관성만으로 모델을 평가할 수 없음

ㄴ관성이 더 이상 획기적으로 줄어들지 않는 지점의 클러스터 개수 선택( k=4 선택 가능)

-실루엣 점수와 클러스터 개수

실루엣 점수: 모든 샘플에 대한 실루엣 계수의 평균

실루엣 계수 : -1과 +1사이의 값

(+1에 가까운 값: 자신의 클러스터 안에 포함되고, 다른 클러스터와는 멀리 떨어짐)

(0에 가까운 값: 클러스터 경계에 위치)

(-1에 가까운 값: 샘플이 잘못된 클러스터에 할당됨)

실루엣 다이어그램

ㄴ클러스터별 실루엣 계수 모음. 칼 모양의 그래프

ㄴ칼 두께 : 클러스터에 포함된 샘플의 개수

ㄴ칼 길이: 클러스터에 포함된 샘플의 실루엣 계수 (길 수록 좋음)

빨간 파선: 클러스트 계수에 해당하는 실루엣 점수

ㄴ대부분의 칼이 빨간 파선보다 길어야 함 (낮으면 다른 클러스터랑 너무 가까움)

ㄴ칼의 두께가 서로 비슷해야, 즉, 클러스터별 크기가 비슷해야 좋은 모델임

-k-평균의 한계

최적이 아닌 솔루션을 피하려면 알고리즘을 여러 번 실행해야 함

클러스터 개수를 미리 지정해야 함

클러스터의 크기나 밀집도가 다르거나, 원형이 아닐 경우 잘 작동하지 않음

-군집을 사용한 이미지 분할

이미지 분할 (image segmentation)

ㄴ이미지를 세그먼트 여러 개로 분할하는 작업

ㄴ다양한 클러스터 개수로 k-평균을 사용해 만든 이미지 분할

시맨틱 분할 (semantic segmentation)

ㄴ동일한 종류의 물체에 속한 모든 픽셀은 같은 세그먼트에 할당

ㄴ자율주행: 보행자들을 모두 하나의 영역, 또는 각각의 영역으로 할당 가능

색상 분할(color segmentation)

ㄴK-평균을 이용하여 색상분할 실행

-군집을 사용한 전처리

미니 MNIST 데이터셋 전처리

K-평균을 전처리 단계로 사용한 후 로지스틱회귀 학습

클러스터 개수 k를 임의로 정하지 않고, GridSearchCV를 사용해 최적의 클러스터

개수를 찾음

-군집을 사용한 준지도 학습

레이블이 없는 샘플이 많고 레이블이 있는 샘플이 적을 때 사용

예제: 미니 Mnist

레이블 전파

ㄴ레이블을 동일한 클러스터에 있는 모든 샘플로 전파

-DBSCAN

밀집된 연속적 지역을 클러스터로 정의

반달모양 데이터 활용

DBSCAN과 예측

ㄴ predict() 메서드 지원하지 않음. fit\_predict() 메서드 제공함

DBSCAN의 장단점

ㄴ매우 간단하면서 매우 강력한 알고리즘

ㄴ군집의 모양과 개수에 상관없음.

ㄴ이상치에 안정적임. 군집 간의 밀집도가 크게 다르면 모든 군집 파악 불가능

-가우시안 혼합

가우시안 혼합 모델( GMM(Gaussian Mixture Model))

ㄴ샘플이 파라미터가 알려지지 않은 여러 개의 혼합된 가우시안 분포에서 생성되었다고 가정하는 확률 모델

ㄴ가우시안 분포 = 정규분포

(종 모양의 확률밀도함수를 갖는 확률분포)

클러스터

ㄴ하나의 가우시안 분포에서 생생된 모든 샘플들의 그룹

ㄴ일반적으로 타원형 모양.

ㄴ아래 그림에서처럼 일반적으로 모양, 크기, 밀집도, 방향이 다름.

ㄴ각 샘플이 어떤 정규분포를 따르는지를 파악하는 것이 핵심.

GMM 활용

GMM 훈련과 EM(Expectation-Maximization) 으로 파라미터 추정

GaussianMixture 모델 적용

n\_components: 군집수 지정

n\_init: 모델 학습 반복 횟수

(파라미터(평균값, 공분산 등)를 무작위로 추정한 후 수렴할 때까지 학습시킴)

EM 알고리즘이 추정한 파라미터를 확인

가우시안 혼합 - EM 알고리즘

초기값: 랜덤하게 확률값을 넣음.

E-Step: Hidden variable의 resposibility를 계산하는 단계

M-Step: 추정값 θ를 업데이트 하는 단계

학습된 모델을 보여줌

ㄴ군집 평균, 결정 경계, 밀도 등고선

GMM 모델 규제

ㄴ특성수가 크거나, 군집수가 많거나, 샘플이 적은 경우 최적 모델 학습 어려움

ㄴ공분산(covariance)에 규제를 가해서 학습을 도와줄 수 있음

가우시안 혼합을 사용한 이상치 탐지

ㄴ밀도가 임곗값보다 낮은 지역에 있는 샘플을 이상치로 간주 가능

클러스터 개수 선택하기

K-평균에서 사용했던 관성 또는 실루엣 점수 사용 불가

베이즈 가우시안 혼합 모델

BayesianGaussianMixture 모델

베이즈 확률통계론 활용

모델의 장단점

ㄴ장점: 타원형 클러스터에 잘 작동

ㄴ단점: 다른 모양을 가진 데이터셋에서는 성능 좋지 않음