MathWorks®

비지도 학습 적용

MathWorks

비지도 학습을 고려하는 경우

비지도 학습은 데이터를 탐색하려고 하지만 아직 구체적인 목표가 없거나 데이터에 포함된 정보가 무엇인지 확실하지 않은 경우 유용합니다. 데이터의 차원을 줄이는 것도 좋은 방법입니다.

비지도 학습 기법

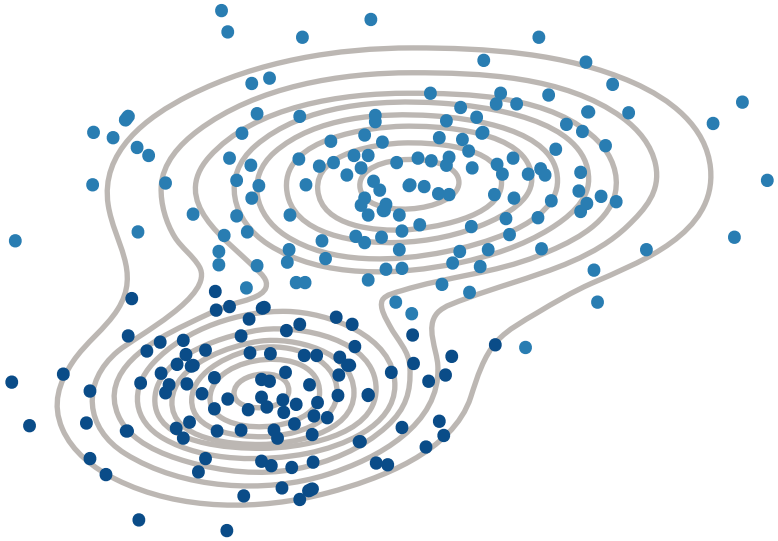
섹션 1에서 확인한 대로 대부분의 비지도 학습 기법은 클러스터 분석의 형태입니다.

클러스터 분석에서 데이터는 유사성 또는 공유 특성의 측정값을 기반으로 그룹으로 분할됩니다. 클러스터는 같은 클러스터의 객체는 매우 비슷하고 다른 클러스터의 객체는 뚜렷이 구별되도록 구성됩니다.

클러스터링 알고리즘은 다음 두 가지 큰 범주로 구분됩니다.

- 하드 클러스터링 - 각 데이터 포인트가 하나의 클러스터에만 속함
- 소프트 클러스터링 - 각 데이터 포인트가 두 개 이상의 클러스터에 속함

가능한 데이터 그룹을 이미 알고 있다면 하드 또는 소프트 클러스터링 기법을 사용할 수 있습니다.



데이터를 두 개의 클러스터로 분할하는 데 사용되는 가우시안 혼합 모델.

데이터 그룹화 방법을 잘 모르는 경우:

- 자기 조직화 특징 맵 또는 계층 클러스터링을 사용하여 데이터에서 가능한 구조체를 찾습니다.
- 클러스터 평가를 사용하여 지정된 클러스터링 알고리즘에 대한 “최적”의 그룹 수를 찾습니다.

일반적인 하드 클러스터링 알고리즘

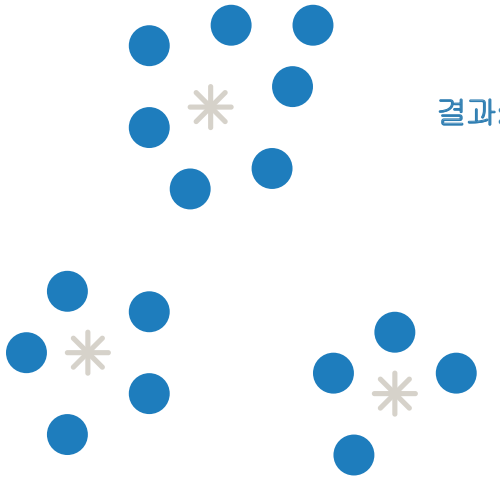
k-평균

작동 방식

데이터를 k개의 상호 배타적인 클러스터로 분할합니다. 포인트가 클러스터에 얼마나 잘 피팅되는지는 해당 포인트에서 클러스터 센터까지의 거리에 따라 결정됩니다.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 대용량 데이터 세트의 빠른 클러스터링을 위해



결과: 클러스터 센터

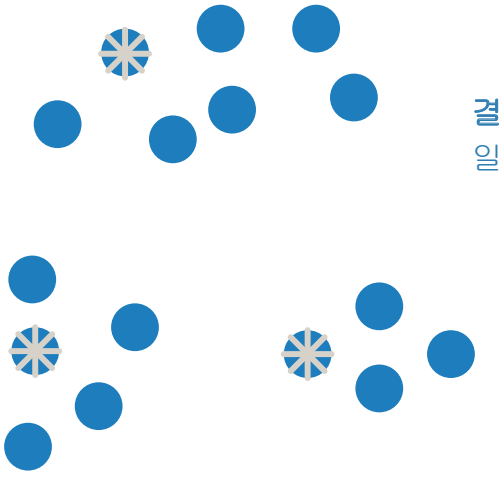
k-중간점

작동 방식

k-평균과 비슷하지만 클러스터 센터가 데이터의 여러 포인트와 일치해야 하는 요구 사항이 있습니다.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 범주형 데이터의 빠른 클러스터링을 위해
- 대용량 데이터 세트로 확장하기 위해



결과: 데이터 포인트와 일치하는 클러스터 센터

일반적인 하드 클러스터링 알고리즘 *계속*

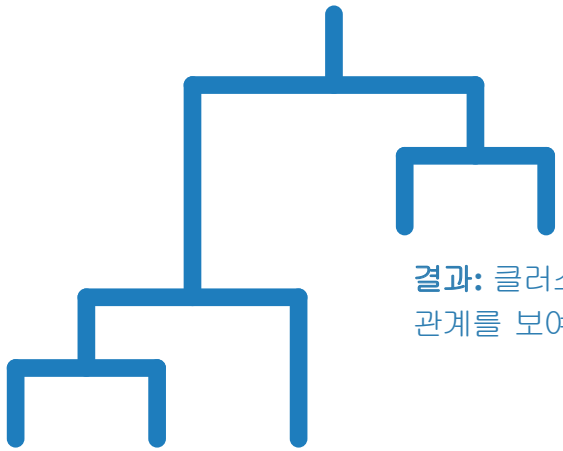
계층 클러스터링

작동 방식

포인트 쌍 간의 유사성을 분석하고 객체를 이진, 계층적 트리로 그룹화하여 중첩된 클러스터 세트를 생성합니다.

최적 사용...

- 데이터에 있는 클러스터 수를 미리 알지 못하는 경우
- 선택에 도움이 되도록 시각화를 원할 경우



결과: 클러스터 간 계층적 관계를 보여주는 덴드로그램

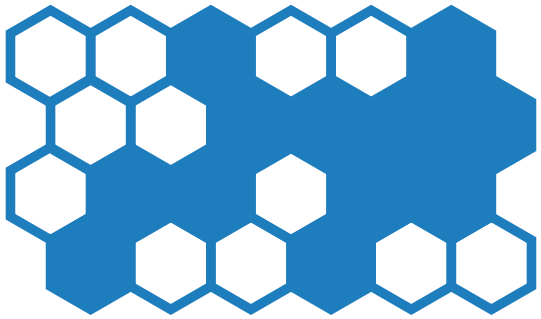
자기 조직화 맵

작동 방식

데이터셋을 토폴로지 보존 2차원 맵으로 변환하는 신경망 기반 클러스터링.

최적 사용...

- 고차원 데이터를 2차원 또는 3차원으로 시각화하기 위해
- 토폴로지(모양)를 보존하여 데이터의 차원을 추론하기 위해



결과:
저차원(일반적으로 2차원) 표현

일반적인 하드 클러스터링 알고리즘 *계속*

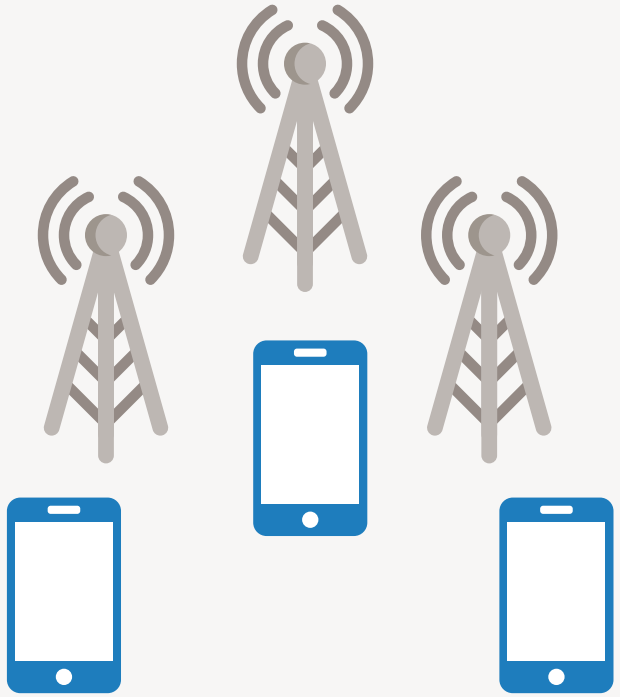
사례: 사이트 휴대 전화 기지국에 대한 k-평균 클러스터링 사용

한 휴대 전화 회사에서 가장 안정적인 서비스를 제공할 휴대 전화 기지국의 수와 배치를 알고자 합니다. 신호 수신을 최적화하기 위해 기지국은 사람들의 클러스터 내에 있어야 합니다.

워크플로우는 나중에 필요하게 될 클러스터 수에 대한 초기 추측(Initial Guess) 으로 시작됩니다. 이 추측을 평가하기 위해 엔지니어는 서비스를 기지국 3개 및 기지국 4개와 비교하여 각 시나리오에 대해 얼마나 잘 클러스터링할 수 있는지(즉, 기지국이 서비스를 얼마나 잘 제공하는지) 확인합니다.

전화기 한 대는 한 번에 한 곳의 기지국에만 신호를 보낼 수 있으므로 이러한 클러스터링은 매우 어렵습니다. k-평균은 데이터의 각 관측값을 공간 정위를 포함하는 객체로 처리하므로 팀에서는 k-평균 클러스터링을 사용합니다. k-평균 클러스터링은 각 클러스터 내의 객체가 가능한 한 서로에게 가깝고 가능한 한 다른 클러스터의 객체와 멀리 있는 파티션을 찾습니다.

알고리즘을 실행한 후 팀에서는 데이터를 클러스터 3개 및 4개로 분할한 결과를 정확히 확인할 수 있습니다.



일반적인 소프트 클러스터링 알고리즘

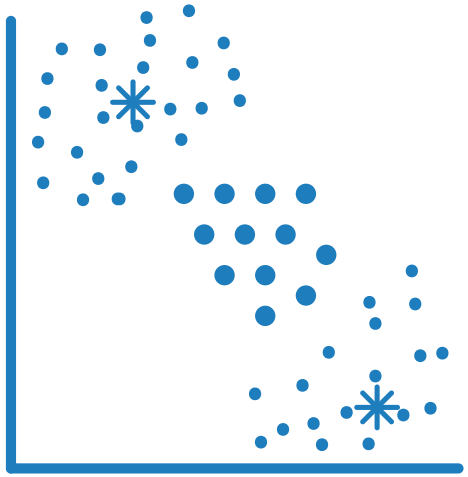
퍼지 c-평균

작동 방식

데이터 포인트가 둘 이상의 클러스터에 속할 수 있는 경우 파티션 기반 클러스터링.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 패턴 인식을 위해
- 클러스터가 겹칠 경우



결과: 클러스터 센터 (k-평균과 유사), 그러나 포인트가 둘 이상의 클러스터에 속할 수 있도록 퍼지니스 포함

가우시안 혼합 모델

작동 방식

데이터 포인트가 특정 확률을 포함한 다양한 다변량 정규 분포에서 나오는 파티션 기반 클러스터링.

최적 사용...

- 데이터 포인트가 둘 이상의 클러스터에 속할 수 있는 경우
- 클러스터의 크기와 클러스터 내의 상관관계 구조체가 다양할 경우

결과: 포인트가 클러스터에 있을 확률을 제공하는 가우시안 분포 모델

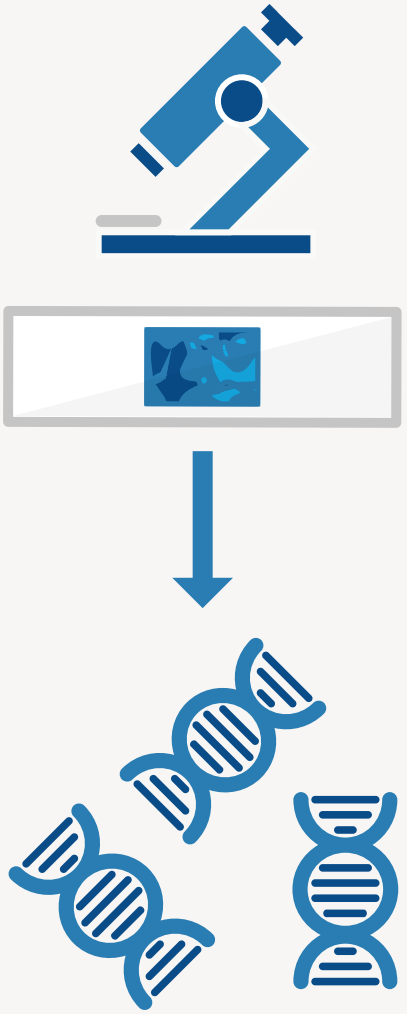
일반적인 소프트 클러스터링 알고리즘 *계속*

사례: 퍼지 c-평균 클러스터링을 사용하여 유전자 발현 데이터 분석

한 생물학자 팀은 정상 및 비정상 세포 분열에 관여한 유전자를 더 잘 이해하기 위해 마이크로어레이에서 유전자 발현 데이터를 분석하고 있습니다. (유전자가 단백질 생성과 같은 세포성 기능에 적극적으로 관여한 경우 유전자가 "발현"된다고 말합니다.)

마이크로어레이에는 두 개의 조직 표본에 기반을 둔 발현 데이터가 포함됩니다. 연구자들은 표본을 비교하여 유전자 발현의 특정 패턴이 암 확산에 관련되는지 확인하고자 합니다.

데이터를 전처리하여 노이즈를 제거한 후 데이터를 클러스터링합니다. 같은 유전자가 여러 생물학적 과정에 관여될 수 있으므로 단일 유전자는 하나의 클러스터에만 속하지 않을 가능성이 높습니다. 연구자들은 퍼지 c-평균 알고리즘을 데이터에 적용합니다. 그러고 나서 클러스터를 시각화하여 비슷한 방식으로 동작하는 유전자 그룹을 식별합니다.



차원성 감소를 통한 모델 개선

머신 러닝은 큰 데이터셋에서 패턴을 찾을 수 있는 효과적인 방법입니다. 하지만 데이터가 크면 더 복잡해집니다.

데이터셋이 커질수록 특징 수 또는 *차원*을 줄여야 하는 경우가 많아집니다.

사례: EEG 데이터 감소

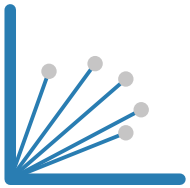
뇌의 전기 활동을 캡처하는 EEG(뇌전도) 데이터가 있다고 가정하고 이 데이터를 사용하여 미래에 있을 수 있는 발작을 예측하려고 합니다. 데이터는 각각 원래 데이터셋의 변수에 해당하는 수십 개의 리드를 사용하여 캡처되었습니다. 이러한 각 변수에는 노이즈가 포함됩니다. 예측 알고리즘을 더 강력하게 만들려면 차원성 감소 기법을 사용하여 더 적은 수의 특징을 도출합니다. 이러한 특징은 여러 센서에서 계산되므로 원시 데이터를 직접 사용한 경우보다는 개별 센서의 노이즈에 덜 민감합니다.

일반적인 차원성 감소 기법

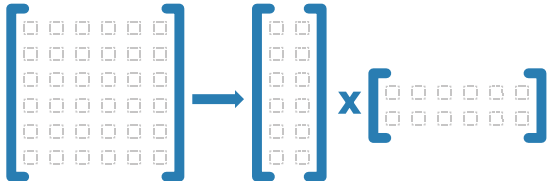
가장 일반적으로 사용되는 세 가지 차원성 감소 기법은 다음과 같습니다.

PCA(주성분 분석) - 처음 몇 개의 주성분을 통해 고차원 데이터셋에서 대부분의 차이 또는 정보가 캡처되도록 데이터에 대한 선형 변환을 수행합니다. 첫 번째 주성분이 가장 큰 차이를 캡처하고, 이어서 두 번째 주성분이 그 다음 차이를 캡처합니다.

인자 분석 - 데이터셋에서 변수 간 기본 상관관계를 식별하여 더 적은 수의 눈에 띄지 않는 잠재적인 인자 또는 일반적인 인자 측면에서 표현을 제공합니다.



음수 미포함 행렬 분해 - 모델 텀(term)이 물리적 수량과 같은 음이 아닌 수량을 표현해야 할 경우 사용됩니다.



주성분 분석 사용

많은 변수를 포함하는 데이터셋에서 변수 그룹은 보통 함께 이동합니다. PCA는 적은 수의 새 변수가 대부분의 정보를 캡처하도록 원래 변수의 일차 결합을 통해 새 변수를 생성하여 이 정보 중복을 활용합니다.

각 주성분은 원래 변수의 일차 결합입니다. 모든 주성분은 서로 직각이므로 중복된 정보가 없습니다.

사례: 엔진 상태 모니터링

엔진의 여러 센서에 대한 측정값이 포함된 데이터셋이 있습니다 (온도, 압력, 배출 등). 많은 데이터가 정상 상태의 엔진에서 나오는 데이터이지만, 센서에서 유지관리가 필요할 경우의 엔진에서 나오는 데이터도 캡처했습니다.

개별 센서를 살펴봐도 분명한 이상 증상을 확인할 수 없습니다. 하지만 PCA를 적용하면 센서 측정값에 있는 대부분의 변형이 적은 수의 주성분을 통해 캡처되도록 이 데이터를 변환할 수 있습니다. 원시 센서 데이터를 확인하는 것보다 이러한 주성분을 검사하면 정상 및 비정상 상태 엔진을 더 쉽게 구분할 수 있습니다.

인자 분석 사용

데이터셋에 겹치는(서로 종속된) 측정 변수가 포함될 수 있습니다. 인자 분석을 사용하면 모델을 다변량 데이터에 피팅하여 이런 종류의 상호 종속성을 예측할 수 있습니다.

인자 분석 모델에서 측정 변수는 더 적은 수의 눈에 띄지 않는 (잠재적) 인자에 의존합니다. 각 인자는 여러 변수에 영향을 미칠 수 있으므로 공통 인자라고 알려져 있습니다. 각 변수는 공통 인자의 일차 결합에 종속된다고 가정합니다.

사례: 주가 변동 추적

100주 동안 10개 회사에 대한 주가의 퍼센트 변화가 기록되었습니다. 10개 회사 중 4개는 기술 회사, 3개는 금융 회사, 나머지 3개는 유통 회사입니다. 같은 부문의 회사들의 주가가 경제 조건이 바뀌면서 함께 달라진다는 가정은 합리적인 것 같습니다. 인자 분석은 전제를 뒷받침할 정량적 증거를 제공할 수 있습니다.

음수 미포함 행렬 분해

이 차원 감소 기법은 특징 공간의 저차수 근사법을 기반으로 합니다. 음수를 포함 하지 않는 행렬에 대한 특징 수를 줄일수 있습니다.

사례: 텍스트 마이닝

여러 웹 페이지의 어휘와 스타일 차이를 살펴보려 한다고 가정해 보겠습니다. 각 행이 개별 웹 페이지에 해당하고 각 열이 단어 (“the”,”a”,”we” 등)에 해당하는 행렬을 만듭니다. 데이터는 특정 페이지에서 특정 단어가 나타나는 횟수입니다.

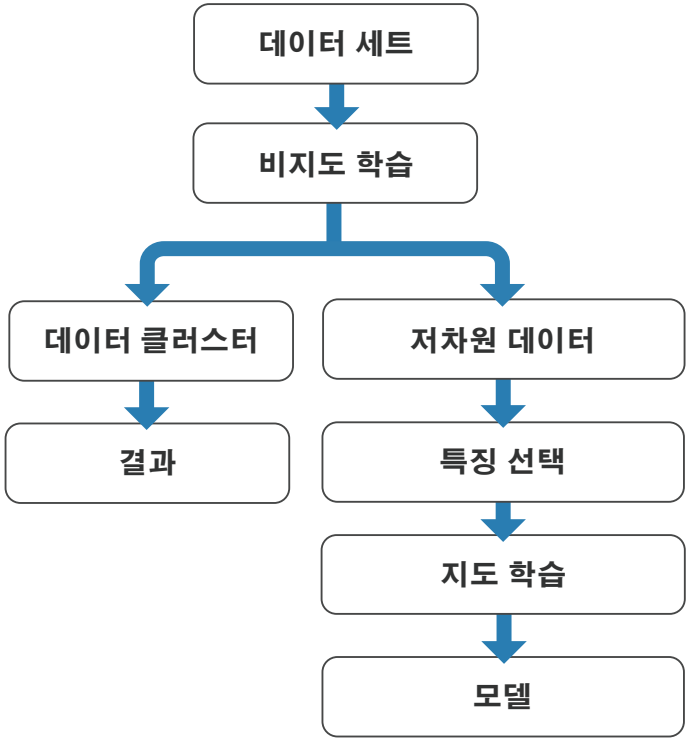
영어에는 수백만 개 이상의 단어가 있으므로 음수 미포함 행렬 분해를 적용하여 개별 단어가 아니라 하이 레벨 개념을 나타내는 임의 개수의 특징을 만듭니다. 이러한 개념을 사용하면 발언, 뉴스, 교육 콘텐츠, 온라인 소매 콘텐츠를 더 쉽게 구별할 수 있습니다.

다음 단계

이 섹션에서는 비지도 학습을 위한 하드 및 소프트 클러스터링 알고리즘을 더 자세히 살펴보고, 데이터에 적합한 알고리즘을 선택할 경우의 몇 가지 팁을 제공하고, 데이터셋에서 특징 수를 줄이면 모델 성능이 어떻게 향상되는지 설명했습니다. 다음 단계에서는 다음을 수행합니다.

- 비지도 학습이 최종 목표일 수 있습니다. 예를 들어 시장 조사를 수행하는 동안 웹 사이트 동작에 따라 소비자 그룹을 구분하여 대상으로 지정하고자 할 경우 거의 확실하게 클러스터링 알고리즘을 통해 찾고 있는 결과를 얻을 수 있을 것입니다.
- 반면, 비지도 학습을 지도 학습의 전처리 단계로 사용하려고 할 수 있습니다. 예를 들어 클러스터링 기법을 적용하여 더 적은 수의 특징을 도출하고 해당 특징을 분리기 훈련을 위한 입력으로 사용합니다.

섹션 4에서는 지도 학습 알고리즘 및 기법을 살펴보고 특징 선택, 특징 감소, 파라미터 튜닝을 통해 모델을 개선하는 방법을 알아보겠습니다.



추가 정보

자세히 살펴볼 준비가 되셨습니까? 다음 비지도 학습 리소스를 살펴보십시오.

클러스터링 알고리즘 및 기법

k-평균

K-평균 및 계층 클러스터링을 사용하여 데이터의 자연 패턴 찾기

K-평균 및 자기 조직화 맵을 사용하여 유전자 클러스터링

K-평균 클러스터링을 사용한 색상 기반 분할

계층 클러스터링

연결 기반 클러스터링

아이리스 클러스터링

자기 조직화 맵

자기 조직화 맵을 사용하여 데이터 클러스터링

퍼지 C-평균

퍼지 C-평균 클러스터링을 사용하여 의사 임의 데이터 클러스터링

가우시안 혼합 모델

가우시안 프로세스 회귀 모델

가우시안 분포 혼합으로부터 데이터 클러스터링

소프트 클러스터링을 사용하여 가우시안 혼합 데이터 클러스터링

가우시안 혼합 모델 튜닝

이미지 처리 예: 가우시안 혼합 모델을 사용한 자동차 감지

차원성 감소

PCA를 사용한 미국 도시 내 생활의 질 분석

인자 분석을 사용한 주가 분석

음이 아닌 행렬 분해

음이 아닌 행렬 분해 수행

차감 클러스터링을 사용한 교외 통근 모델링

93113v00