1 자율 학습

1 자율 학습

- 학습이라는 것은 계산을 반복하면서 가중치 계수를 업데이트하는 작업
- 가중치 계수 업데이트를 통해 모델이 되는 기저 함수와 분포에 접근하는 것을 의미
- 자율 학습
  - 정답 정보가 없는 상태에서 학습을 통해 모델을 만드는 것
  - 데이터가 어떻게 구성되었는지를 알아내는 문제 범주에 속함
- 특징
  - 통계의 밀도 추정과 연관
  - 데이터의 주요 특징을 요약하고 설명

2 데이터 마이닝

- 정의
  - 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 것
  - KDD(Knowledge Discovery in Databases)
- 방법론
  - 통계학적 방법론: 통계학 쪽에서 발전
  - 기술적인 방법론: 인공지능 진영에서 발전
- 신용 평가 모델, 사기탐지시스템, 장바구니 분석, 최적 포트폴리오 구축 등과 같은 산업 분야에 광범위하게 사용 됨

2 데이터 마이닝

연관성 분석

- 거래나 사건을 포함하는 일련의 데이터로부터 연관 규칙을 발견하고 둘 이상의 품목들 간의 상호 연관성을 파악
- 장바구니 분석 또는 친화성 분석 방법 등

예: 맥주와 기저귀의 연관성

대형마트의 구매 패턴을 분석한 결과 주말에 아기용 기저귀를 구매하러 온 아버지들이 동시에 맥주를 구매, 이 패턴을 분석하여 월마트에서 기저귀와 맥주 패키지 판매를 통한 3배 매출 증가 효과 달성

2 데이터 마이닝

군집 분석

- 이질적인 집단을 몇 개의 동질적인 소집단으로 세분화하는 작업
- 특징: 사전에 정의된 집단이 없음
- 여러 집단의 데이터들이 섞여 있고 각 데이터 소속 집단을 모르는 경우 유사한 속성을 갖는 데이터의 군집을 찾는 분석 방법
- 주어진 데이터 중 유사한 것들만 몇몇의 집단으로 그룹화하여 각 집단의 성격을 파악하고 데이터 전체의 구조에 대한 이해를 돕도록 함

2 데이터 마이닝

의사결정 나무

- 분류 문제를 해결하기 위해 매우 강력하고 유용한 데이터 마이닝 알고리즘
- 분류를 하기 위한 목표 변수에 영향을 줄 수 있는 입력 변수들을 이용하여 최적의 분류를 위한 의사결정 규칙 생성

2 데이터 마이닝

- 클러스터 분석, 차원 압축 등을 주로 이용하며 그림을 결과를 표현하여 사람이 특징을 파악할 수 있도록 함

장점	단점
정보의 연관성을 파악하고 가치 있는 정보를 만들어 의사결정을 하고 이익을 극대화 시킴	자료에 의존하여 현상 해석 및 개선으로 인해 자료가 충실하지 못할 경우 오류 존재

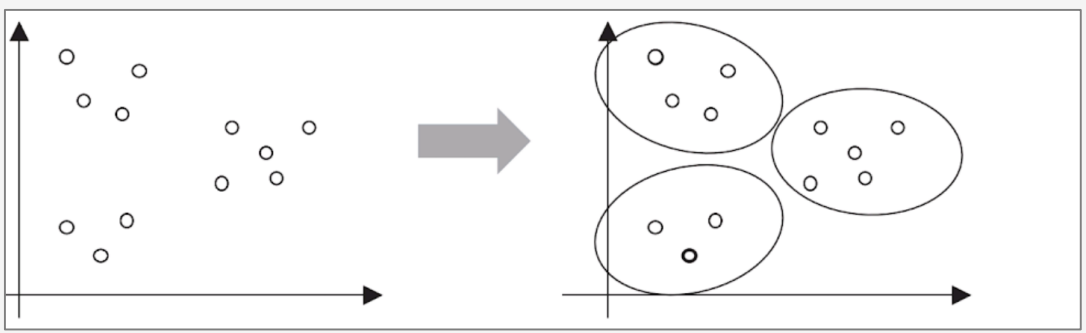
2 클러스터 분석과 K-평균 알고리즘

1 클러스터 분석

- 자율 학습의 대표적인 접근 방법
- 평면상에 그려져 있는 점들을 그룹으로 분류
- 그룹으로 만들 때 점들 사이가 어느 정도 떨어져 있는 지 측정 지표
- 클러스터란 비슷한 특성을 가진 데이터들의 집단
- 효율성
  - 수백 만의 데이터를 직접 확인하지 않고 각각 클러스터의 대표값만 확인하여 전체 데이터의 특성 파악 가능

1 클러스터 분석

- 평면 상에 점을 그림
- 데이터 특성 고려한 클러스터 정의
- 데이터 집단의 대표점 검색
- 데이터 마이닝의 대표적 예



[클러스터 분석 예]

1 클러스터 분석

기법

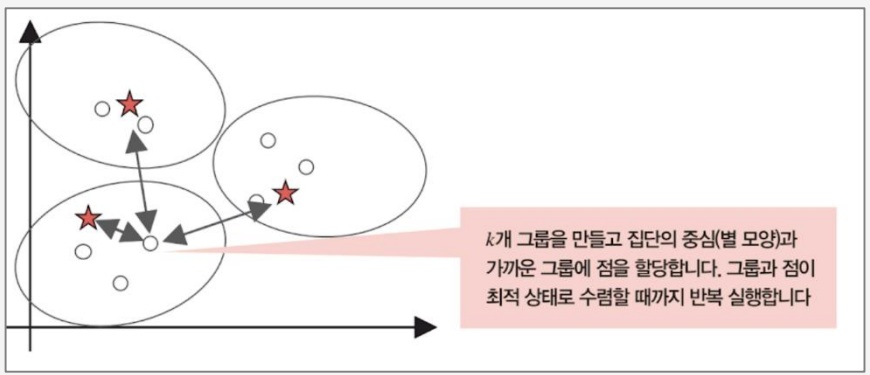
파티셔닝	data를 구간구간으로 나눠서 그들의 중심을 계산
k-평균	각 구간을 나눈 다음 중심을 찾고 중심을 기준으로 구간을 다시 나누고 변경 사항이 있을 경우 다시 중심을 찾아가는 방식
k-중간점	k-mean의 경계에 약한 단점을 보완. 중심 대신 중심객체를 사용
clara	샘플링 k-medoids(중앙 객체)
clarans	근처에 있는 데이터들끼리 샘플링

2 k-평균 알고리즘

- 클러스터 분석 시 자주 사용하는 방법
- 과정
  - 우선 전체를 k개의 그룹으로 나눔
  - 각 점에 무작위로 그룹을 할당한 다음 그룹 각각의 중심과의 거리를 계산
  - 어떤 그룹에 속해 있는 점이 다른 그룹과 더 가깝다면 해당 점을 거리가 가까운 그룹으로 변경
  - 이러한 작업을 반복해서 가까운 점끼리 묶어 k개의 그룹을 분류

2 k-평균 알고리즘

- 임의의 점들의 분포를 이용하여 k개의 그룹으로 분리
- 중심의 대표점을 선택
- 각각의 점들과 그룹의 대표점들 간의 거리 측정
- 재 그룹화



[K-평균 알고리즘 예]

2 k-평균 알고리즘

- 단점
  - 그룹의 중심점만 기준으로 삼다가 잘못된 그룹으로 할당할 수 있음
  - 계산 시간이 길어짐
- 해결 방법
  - k-평균 알고리즘으로 여러 번 그룹을 생성하여 가장 좋은 결과를 채용
  - 그룹 생성 전 중심점을 되도록 떨어져 있게 설정하는 k-평균++ 알고리즘 사용

k-평균++ 알고리즘: k-means 알고리즘의 초기값을 선택하는 알고리즘

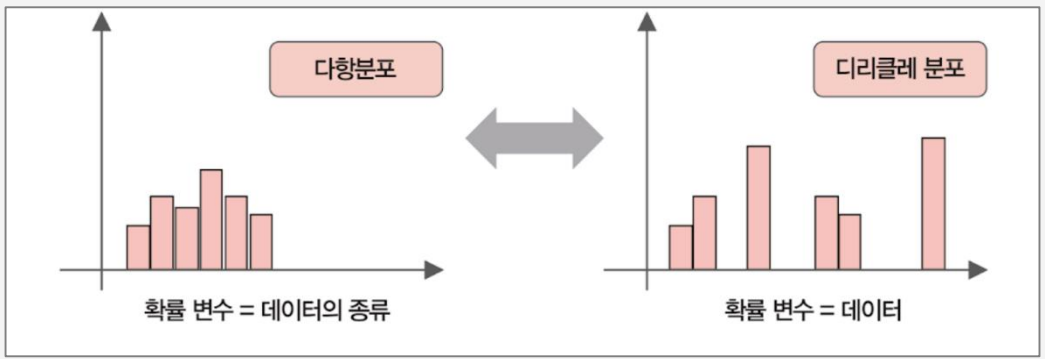
2 k-평균 알고리즘

- 초기 그룹의 수인 k는 경험적으로 결정할 수도 있고 계산으로 산출하여 결정할 수도 있음
- k 결정 시 디리클레 모델 사용
  - 혼합 디리클레 모델은 베이즈 모델 기반의 접근 방식
  - 디리클레 분포가 다항분포의 결합 사전 분포라는 점을 이용

다항분포는 각 사건의 확률을 나타내는 분포이며 디리클레 분포는 발생하는 사건의 개수를 나타내는 분포

2 k-평균 알고리즘

- 혼합 디리클레 모델로 데이터를 그룹에 할당할 때 보통 기존의 가까운 그룹에 할당
- 그룹 할당 과정을 EM 알고리즘으로 반복 실행하면 그룹 개수와 그룹 각각에 할당하는 데이터 분포 관찰 가능



[다항 분포와 디리클레 분포의 관계]

3 GMM 알고리즘

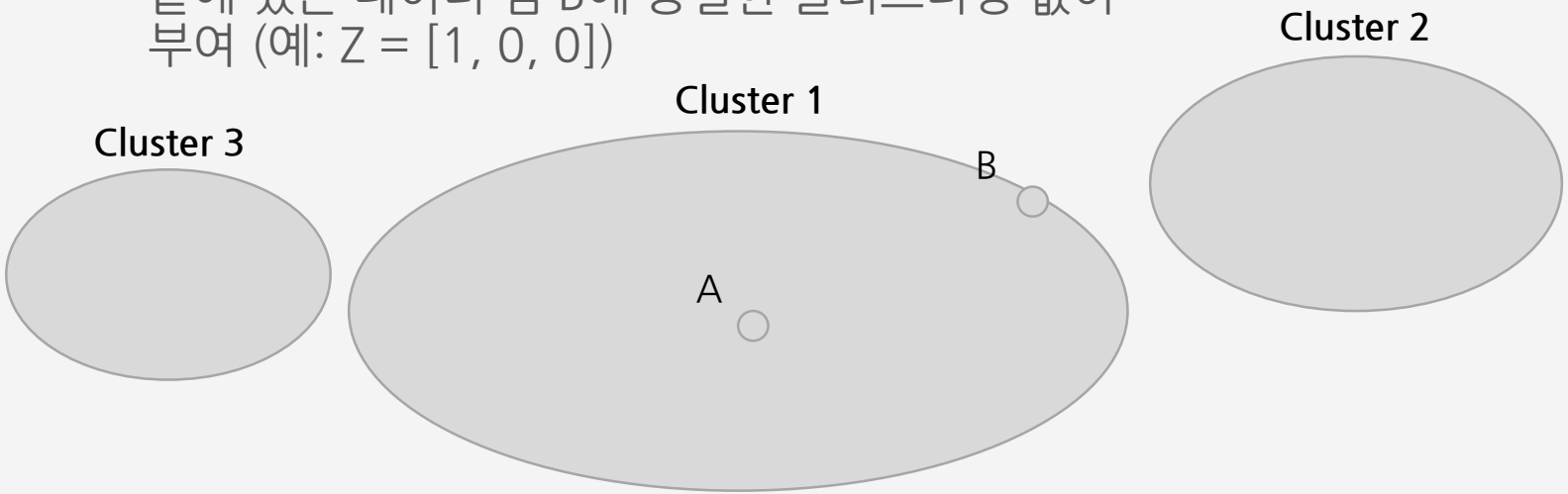
3 GMM 알고리즘

1 가우시안 혼합 모델 개념

- 가우시안 분포가 여러 개 혼합된 클러스터링 알고리즘
- 자연적인 현상 표현에 좋은 모델로 데이터 마이닝, 패턴 인식, 통계 분석 등에 광범위하게 적용
- 여러 개의 가우시안 분포로부터 생성(수집)된 데이터들을 학습하여 각 데이터들이 어떤 가우시안 분포에서 생성된 것인지 확률로 추정 가능
- 새로운 데이터가 입력될 경우 동일한 과정으로 클러스터링이 가능

1 가우시안 혼합 모델 개념

- K-means 클러스터링
  - 클러스터 n의 중심에 있는 데이터 점 A와 외곽 끝에 있는 데이터 점 B에 동일한 클러스터링 값이 부여 (예: Z = [1, 0, 0])



1 가우시안 혼합 모델 개념

- 3개의 클래스는 정해진 것이 아니라 추정되는 것으로 이것을 잠재 변수라 함

입력데이터	클래스	클래스 확률
$x_A = [0.02, 0.08]$	$r =$	$\gamma =$
$x_B = [0.96, 0.98]$	$[1, 0, 0]$	$[0.9\ , 0.1, 0\ ]$
	$[1, 0, 0]$	$[0.5, 0.4, 0.1]$

3개의 클러스터가 존재할 경우

- γ
  - 클러스터 k에 속할 확률
  - 어떤 클러스터에 얼마나 기여하고 있는가
    → 부담율(Responsibility)

1 가우시안 혼합 모델 개념

$$p(x) = \sum_{k=0}^{K-1} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

가우스 함수

클래스 확률

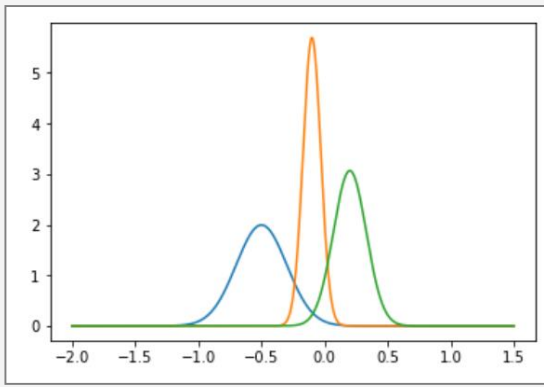
중심 벡터: $\boldsymbol{\mu}_k = (\mu_{k0}, \mu_{k1})$ ..... 각 가우스 분포 k의 중심

공분산 행렬: $\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{k00} & \sigma_{k01} \\ \sigma_{k10} & \sigma_{k11} \end{bmatrix}$ ..... 각 가우스 분포 k의 퍼짐

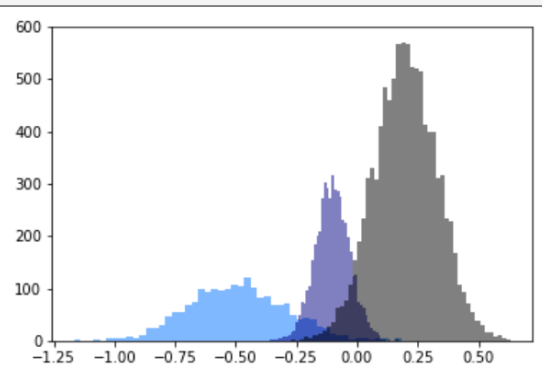
혼합 계수: $\boldsymbol{\pi}_k, 0 \leq \boldsymbol{\pi}_k \leq 1, \sum_{k=0}^{K-1} \boldsymbol{\pi}_k = 1$ ..... 각각의 가우스 분포의 크기 비율

2 GMM의 예

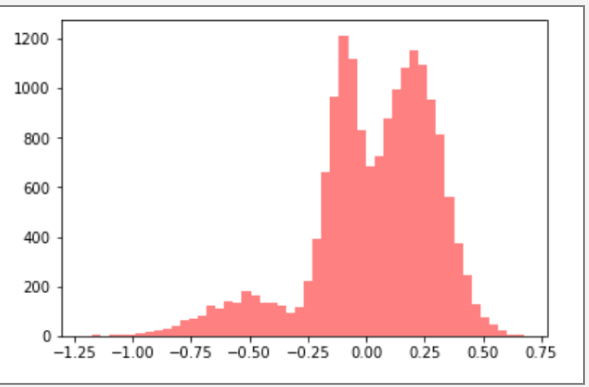
- 세 개의 gaussian 분포를 이용하여 데이터를 생성



[기저가 된 3개의 가우시안 분포]



[생성된 데이터]

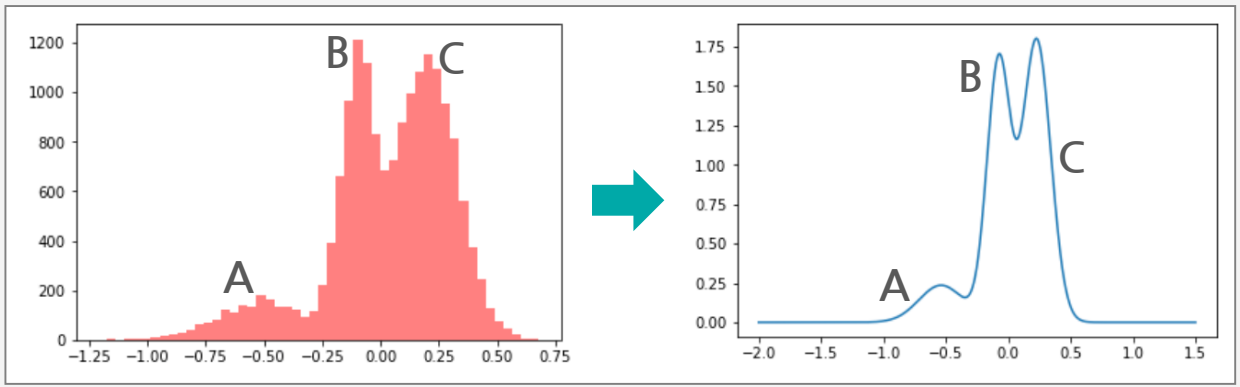


[실제 우리가 보는 데이터]

- 3가지 분포의 평균, 매개변수, 가중치를 구하여 혼합 모델을 구축
- 매개 변수 추정에 대한 다양한 알고리즘을 통해 GMM 모델 생성

2 GMM의 예

- 매개변수 추정을 통해 생성한 GMM 모델



- 왼쪽은 데이터 빈도, 오른쪽은 확률 밀도 함수
- 데이터 중 (-1, 0) 을 클러스터링 하고자 할 경우, 어떤 3개의 정규 분포 중 어느 분포에서 발생할 확률이 가장 높은지를 산출하여 클러스터링

3 GMM의 매개변수 추정

- 목적
  - 관측되지 않는 클러스터 분포의 확률 분포와 각각의 클러스터에서의 gaussian 정규분포 매개변수를 모두 추정하는 것
- 단점
  - 확률분포 함수가 선형대수 방법으로 쉽게 구할 수 없는 복잡한 형태를 가진다는 점
- 매개 변수
  - 잠재변수(latent variable) → 클러스터의 개수
  - 부담율(responsibility)

3 GMM의 매개변수 추정

잠재 변수

- 관찰은 안되지만 데이터에 영향을 준 변수를 잠재 변수 또는 숨은 변수

예

- 곤충의 질량과 크기를 나타내는 입력 데이터가 있는데 같은 곤충이라 생각하고 수집한 데이터를 클러스터링 해 보니 3개의 클러스터가 있는 것으로 추정
- 같은 곤충에 3가지 종류가 있을 수 있다고 해석이 가능
- 3개의 클러스터가 3개의 클래스라는 존재를 암시

3 GMM의 매개변수 추정

부담율(responsibility)

- 관찰되지 않은 잠재 변수의 추정치가 어떤 클러스터에 포함될 것인지를 알려 주는 조건부 확률

$$\gamma_{ik} = p(z_i = k|x_i) = \frac{p(z_i = k)p(x_i|z_i = k)}{p(x_i)} \qquad z_i\text{: class(또는 cluster)}$$

$$= \frac{p(z_i = k)p(x_i|z_i = k)}{\sum_{k=0}^{K-1} p(x_i, z_i = k)} = \frac{p(z_i = k)p(x_i|z_i = k)}{\sum_{k=0}^{K-1} p(z_i = k)p(x_i|z_i = k)}$$

- γ_{ik}
  - 데이터 x_i가 클러스터 k에서 생성되었을 확률
  - 확률적 추정값으로 0에서 1사이의 실수값

4 EM 알고리즘

- 1958년 Hartley에 의해서 제안된 군집 알고리즘
- K-Means 알고리즘과 마찬가지로 초기 모델을 생성한 후 반복 과정을 통하여 최적화된 모델을 생성
- 반복 과정을 통하여 각 데이터들이 혼합 모델(Mixture Model)에 속할 가능성(Probability)을 조정하여 최적의 모델을 생성
- K-Means 알고리즘에서는 유클리디언 거리 함수를 사용하는 반면에 EM 알고리즘은 log-likelihood 함수를 사용하여 모델의 적합성을 평가
- EM 알고리즘은 최적해로 수렴한다는 것이 증명
- EM은 초기 해에 따라 최종 해가 달라지는 욕심 알고리즘(Greedy algorithm)이고, 전역 최적 해가 아닌 지역 최적 해로 수렴할 수도 있음

4 EM 알고리즘

- E Step(Expectation Step)
  - π,μ,Σ 초기화로 부담율 γ를 산출
- M Step(Maximization Step)
  - 현시점의 γ를 이용하여 π,μ,Σ 를 산출
- 각 단계가 임계치보다 작을 때까지 반복

$$y = a * exp(-\frac{(x-\mu)^2}{\sigma^2})$$

Gauss Function

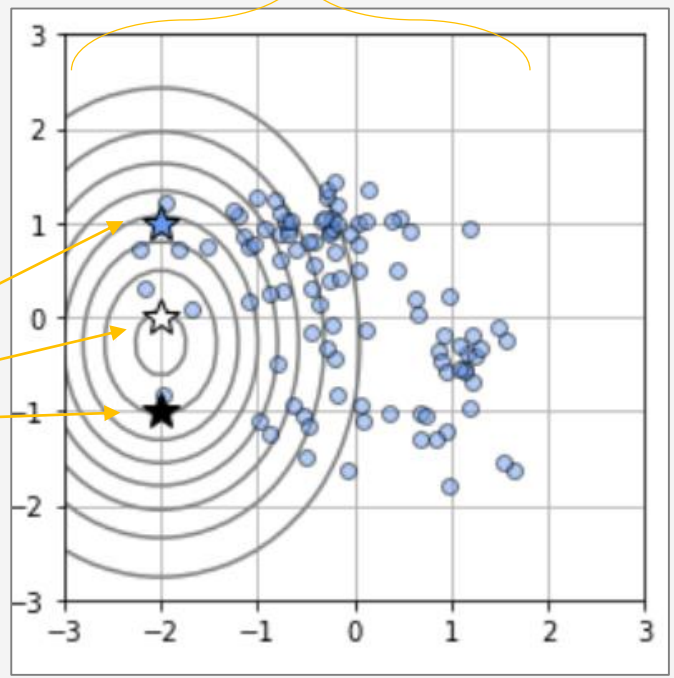
π:클러스터의 혼합 계수*
μ: 클러스터의 중심 벡터
Σ: 클러스터의 공분산 행렬

4 EM 알고리즘

- 초기화
  - π,μ,Σ의 초기화와 데이터의 분포를 시각화

각 클러스터에 대한 데이터의 부담율

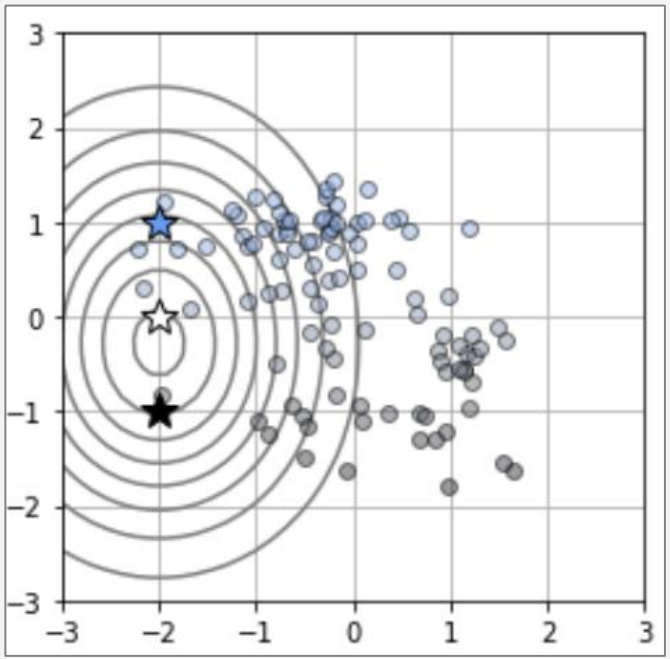
π_i, μ_i, Σ_k



[초기화 모델]

4 EM 알고리즘

- Expectation Step
  - π,μ,Σ 로 부담율 γ를 산출
  - $\gamma_{nk} = \frac{\pi_k N(\mathrm{x}_n|\mu_k, \Sigma_k)}{\sum_k^K N(\mathrm{x}_n|\mu_k, \Sigma_k)}$ 가우스 함수
  - 각 가우스 함수 값을 계산한 후 합이 1이 되도록 정규화



[데이터의 부담율을 이용하여 표시된 모델]

4 EM 알고리즘

- Maximization Step(1)
  - 각 클러스터에 대한 부담율의 합을 산출

$$N_k = \sum_{n=0}^{N-1} \gamma_{nk} \qquad \mu_k^{new} = \frac{1}{N_k} \sum_{n=0}^{N-1} \gamma_{nk} \mathrm{x}_n$$

◁ 중심 벡터 갱신

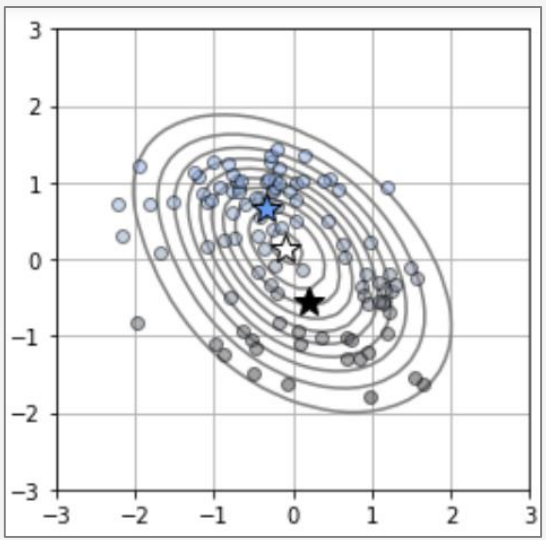
  - 산출된 값을 이용하여 혼합율 갱신

$$\pi_k^{new} = \frac{N_k}{N} \qquad \Sigma_k^{new} = \frac{1}{N_k} \sum_{n=0}^{N-1} \gamma_{nk} (\mathrm{x}_n - \mu_k^{new})^T (\mathrm{x}_n - \mu_k^{new})$$

△ 공분산 갱신

4 EM 알고리즘

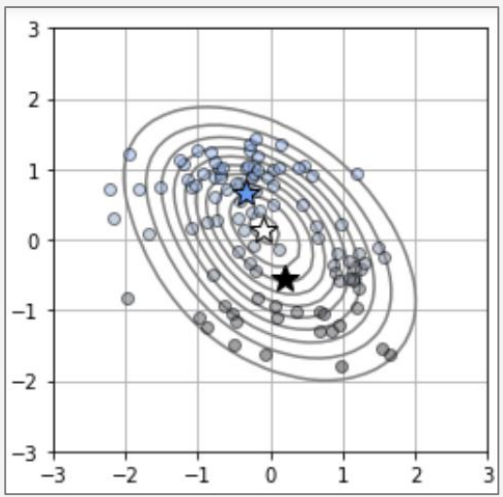
- Maximization Step(2)
  - 갱신된 혼합율과 중심벡터, 공분산을 이용하여 모델 갱신



[1회 EM 스텝을 수행 한 후 클러스터링 된 데이터들]

4 EM 알고리즘

- 반복 과정
  - 20회 반복을 통해 클러스터의 중심으로 $gauss$ 함수가 이동하여 수렴
  - K-means와는 다르게 클러스터의 경계 부분에 확률적으로 데이터가 표시되는 것을 확인
  - K-means와 마찬가지로 매개 변수의 초기값에 따라 변화
  - 클러스터링의 평가를 위해 k-means는 왜곡 척도를 사용하였고 GMM은 가능도를 사용



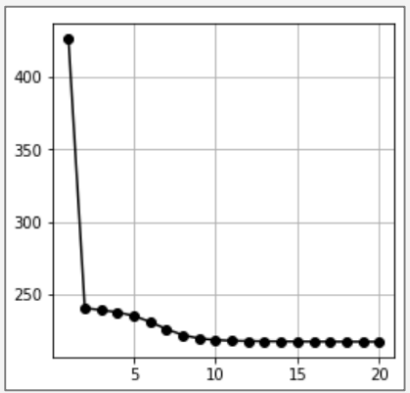
[20회 수행 한 후 클러스터링 된 데이터들]

4 EM 알고리즘

- 가능도
  - 가능도는 모든 데이터 점 X가 모델에서 생성된 확률
  - 로그를 취한 후 -1을 곱한 값
  - 가능도나 로그 가능도를 최적화할 때에는 확률을 극대화 하기 때문에 -1을 곱한 음의 로그 가능도를 오차 함수로 정의

4 EM 알고리즘

- 가능도 오차 함수
  - 스텝의 반복에 따라 가능도 값을 측정
  - 음의 로그 가능도 계산 알고리즘은 반복 계산의 종료 조건으로도 사용이 가능
  - 다양한 초기값을 이용한 클러스터링으로 음의 로그 가능도 값이 가장 작은 것이 좋은 결과임을 판단하는 기준



[가능도 오차 함수]

5 GMM 모델 응용 예(1)

퍼지 영상 분할(Fuzzy image segmentation)

- 교통 분야의 영상 처리 시 사고와 혼잡의 구분을 위해 GMM 모델을 적용
- 기존 영상 처리나 CV에서는 하나의 픽셀을 하나의 패턴에만 할당
- 만약 패턴들이 $gaussian$ 분포라면, 퍼지 영상 분할은 $gaussian$ 혼합 모델 분포를 따름
- 공간적으로 정규화 된 혼합 모델로서 더 현실적이고 계산적으로 효율적인 분할 방법

5 GMM 모델 응용 예(1)

문서의 주제(Topic Model)

- 문서가 N개의 서로 다른 단어와 전체 크기가 V인 어휘로 구성되어 있고, 각 단어는 K개의 주제 중 하나에 해당한다고 가정
- 이러한 단어들의 분포는 K 개의 서로 다른 V 차원 범주형 분포(Categorical distribution)의 혼합 모델로 표현
- 사전 분포(Prior distribution)는 아주 작은 수의 단어만 0이 아닌 확률을 가지고 있는 스파스 분포(Sparse distribution)를 만들기 위해 토픽 분포를 묘사하는 매개변수 사용
- 자연적 집단화(natural clustering)를 이용하기 위해, 몇몇 종류의 추가 조건은 단어들의 토픽 유사성(identities)을 사용