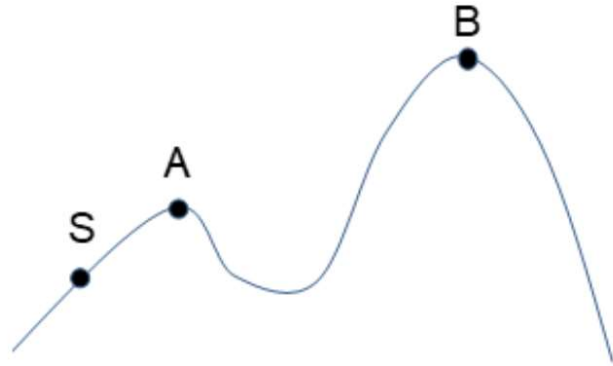


| 12 주차 조별보고서 (Default) | |
|--------------------------------|---|
| 작성일: 2019 년 11 월 29 일 | 작성자: 위성조 |
| 조 모임 일시: 2019 년 11 월 29 일 9 교시 | 모임장소: 학교 앞 카페 |
| 참석자: 위성조, 이충현, 최진성, 이재은, 김영연 | 조원: 위성조, 이충현, 최진성, 이재은, 김영연 |
| 구 분 | 내 용 |
| 학습 범위와 내용 | <p>6.1 절 비지도 학습을 지도/준지도 학습과 비교</p> <p>6.2 절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환</p> <p>6.3 절 k-평균과 친밀도 전파 알고리즘</p> <p>6.4 절 커널 밀도 추정과 가우시안 혼합</p> |
| 논의 내용 | <p>Q. EM알고리즘은 Greedy 알고리즘인가?</p> <p>A. 먼저 greedy알고리즘이란 복잡한 문제가 있을 때 그 상황에서 가장 좋다고 생각하는 solution을 선택하고 그 다음 차례에서는 그 다음 상황에서 가장 좋다고 생각하는 Solution을 선택하는 것이다. 이것을 반복하며 최적의 solution을 찾는 것이다.</p> <p>Greedy 알고리즘이 적용되는 대표적인 경우에는 EM알고리즘이 있다. EM알고리즘은 먼저 문제에 대해 정확하게 정의하고, 그 때 발생하는 parameter가 무엇인지, parameter는 어떻게 계산되는지 정의해 두어야 하며 그 다음에는 Expectation과 Maximization을 반복하여 최적의 답을 찾아낸다. Expectation은 잠재변수 z의 기대치를 계산하고, Maximization은 잠재변수 z의 기대치를 이용하여 파라미터를 추정하는 것이다.</p> |

EM은 greedy 알고리즘으로 눈 앞에 놓인 최적점을 찾으려 노력할 뿐 global 최적점은 찾지 못한다는 단점이 있다.



즉, s에서 시작해서 최대점을 찾으려 할 때 B를 찾아야 하는데 A에서 만족한다.

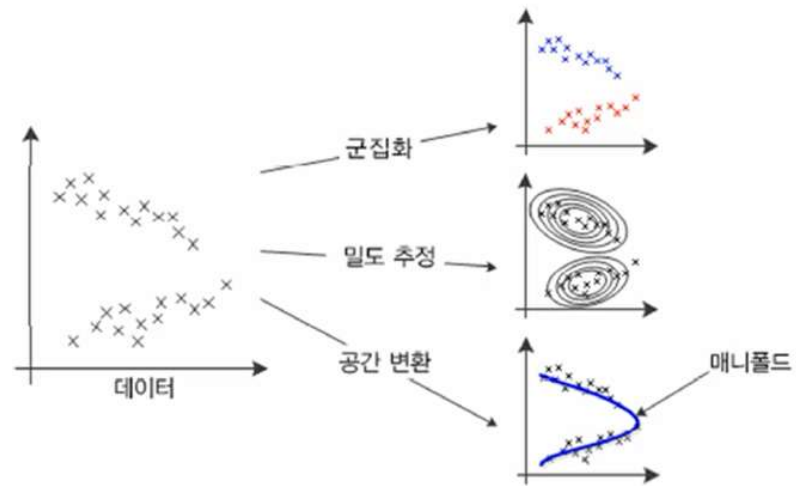
질문 내용

<첨부 개인 레포트>

-201402033 위성조

| 구분 | 내용 |
|-------|--|
| 학습 범위 | 6.1절 비지도 학습을 지도/준지도 학습과 비교 6.2절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환 6.3절 k-평균과 친밀도 전파 알고리즘 6.4절 커널 밀도 추정과 가우시안 혼합 |
| 학습 내용 | 준지도 학습 : 레이블을 가진 샘플과 가지지 않은 샘플이 섞여 있음 기계 학습이 사용하는 두 종류의 지식 훈련집합 사전 지식(prior knowledge/세상의 일반적인 규칙) 중요한 두 가지 사전 지식 매니폴드 가정 : 데이터집합은 하나의 매니폴드 또는 여러 개의 매니폴드를 구성하며, 모든 샘플은 매니폴드와 가까운 곳에 있다. 매끄러움 가정 : 샘플은 어떤 요인에 의해 변화한다. 예를 들어, 장면과 카메라 위치를 고정한 상태에서 조명을 조금씩 변화하면서 영상을 획득한 경우, 획득된 영상 샘플은 특징 공간에서 위치가 조금씩 바뀔 것이다. 이 때, 매끄러운 곡면을 따라 위치가 변한다. 비지도 학습과 준지도 학습은 사전 지식을 더 명시적으로 사용한다. 비지도 학습의 일반 과업 군집화 : 유사한 샘플을 모아 같은 그룹으로 묶는 일 밀도 추정 : 데이터로부터 확률분포를 추정하는 일 |

공간 변환 : 원래 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일



비지도 학습의 응용

군집화의 응용 : 맞춤 광고, 영상 분할, 유전자 데이터 분석, SNS 실시간 검색어 분석하여 사람들의 관심 파악 등

밀도 추정의 응용 : 분류, 생성 모델 구축 등

공간 변환의 응용 : 데이터 가시화, 데이터 압축, 특징 추출(표현 학습) 등

$\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 에서 식 (6.1)을 만족하는 군집집합 $C = \{c_1, c_2, \dots, c_k\}$ 를 찾아내는 작업

$$\left. \begin{array}{l} c_i \neq \emptyset, i = 1, 2, \dots, k \\ \bigcup_{i=1}^k c_i = \mathbb{X} \\ c_i \cap c_j = \emptyset, i \neq j \end{array} \right\}$$

군집의 개수 k 는 주어지는 경우와 자동으로 찾아야 하는 경우가 있다.

군집화는 주관적이다.

| | |
|--|---|
| | <p>K-means(k-평균) 알고리즘</p> <p>원리가 단순하고, 이해하기 쉬우며, 구현하기 쉬운데, 성능이 좋아 인기가 좋음</p> <p>군집 개수 k를 알려줘야 함.</p> <p>k-means와 k-medoids</p> <p>k-means는 샘플의 평균으로 군집 중심을 갱신</p> <p>k-medoids는 샘플 중 대표를 뽑아 뽑힌 대표로 군집 중심을 갱신(k-means에 비해 잡음에 둔감)</p> <p>다중 시작 k-means</p> <p>k-means는 초기 군집 중심이 달라지면 최종 결과가 달라짐 -> 서로 다른 초기 군집 중심을 가지고 여러 번 수행한 다음 목적함숫값이 가장 작은 해를 최종해로 취한다.</p> <p>k-means 알고리즘의 한계</p> <p>군집의 크기가 불균형한 경우 / 군집의 밀도에 차이가 있는 경우 / 특이한 분포</p> <p>EM (Expectation Maximization) 기초</p> <p>E단계와 M단계를 반복하며 추정을 고도화</p> <p>친밀도 전파 알고리즘 – 군집 개수 k를 자동으로 알아냄</p> <p>책임행렬과 가용행렬을 이용함.</p> <p>자가 유사도 s_{kk} – 하이퍼 매개변수이며, 유사도의 최솟값, 중앙값(메디안), 최댓값 중에서 선택</p> <p>최솟값은 적은 수의 군집, 최댓값은 많은 수의 군집을 생성. 중앙값은 중간 정도</p> <p>자신의 자가 친밀도가 최고이면 군집 대표임.</p> <p>밀도 추정 문제 – 어떤 점 x 에서 데이터가 발생할 확률을 구하는 문제</p> |
|--|---|

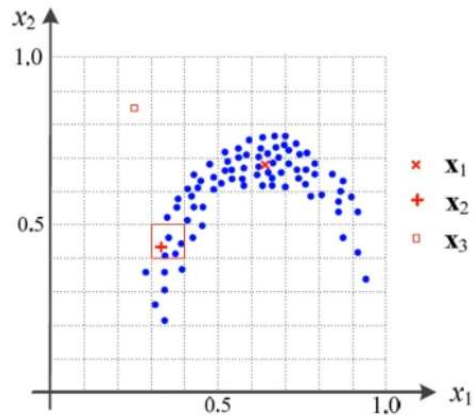


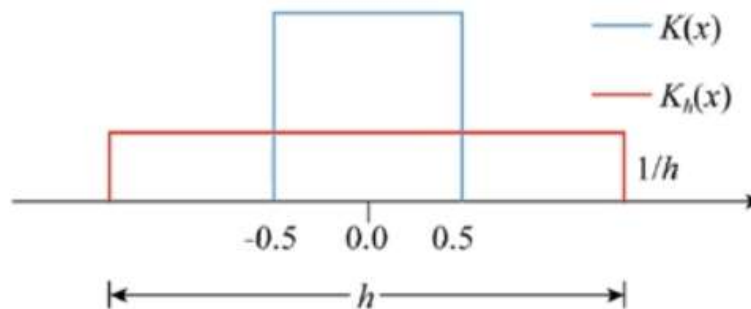
그림 6-8 밀도 추정 문제

$$P(x_1) > P(x_2) > P(x_3)$$

히스토그램 방법 - 특정 공간을 칸의 집합으로 분할한 다음, 칸에 있는 샘플의 빈도를 세어 추정

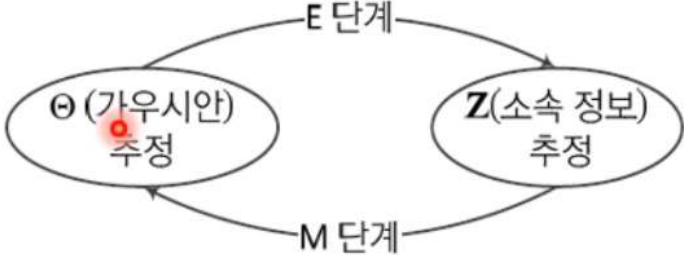
- 매끄럽지 못하고 계단 모양을 띠는 확률밀도함수가 됨/ 칸의 크기와 위치에 민감함

커널 밀도 추정법 - 점 x 에 아래 그림과 같은 커널을 씌우고 커널 안에 있는 샘플의 가중 합을 이용, 대역폭 h 의 크기가 중요



h 가 너무 작으면 뾰족한 모양, 너무 크면 뭉개지므로, 적절한 값으로 설정해야 함.

문제점 - 샘플을 모두 저장하고 있어야 하며, 새로운 샘플이 들어오면 처음부터 다시 계산해야 함.

| | |
|-------|--|
| | <p>차원이 높아지는 경우, 데이터가 적어(차원의 저주) 측정하기가 어려움 - 낮은 차원에 국한하여 사용</p> <p>가우시안 혼합 - 데이터가 가우시안 분포를 따른다고 가정하고 평균 벡터 μ와 공분산 행렬 Σ를 추정함</p> <p>대부분의 데이터가 하나의 가우시안으로 불충분</p> <p>가우시안 혼합을 위한 EM 알고리즘</p> <p>θ(가우시안)을 모르므로 난수로 설정하고 시작</p> <p>가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속 정보로 가우시안 개선(M단계) ...</p>  |
| 질문 내용 | |

-201402665 이충현

| 구분 | 내용 |
|-------|--|
| 학습 범위 | <p>6.1절 비지도 학습을 지도/준지도 학습과 비교</p> <p>6.2절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환</p> <p>6.3절 k-평균과 친밀도 전파 알고리즘</p> <p>6.4절 커널 밀도 추정과 가우시안 혼합</p> |
| 학습 내용 | <p><6.1 절 비지도 학습을 지도/준지도 학습과 비교></p> <ul style="list-style-type: none"> 지도 학습 = 모든 훈련 샘플이 레이블 정보를 가짐 비지도 학습 = 모든 훈련 샘플이 레이블 정보를 가지지 않음 준지도 학습 = 지도+비지도 중요한 두 가지 사전 지식 <ol style="list-style-type: none"> 1) 매니폴드 가정 = 모든 샘플은 매니폴드와 가까운 곳에 있다. 고차원 데이터로부터 저차원의 Locally Euclidian 를 구한다. 고차원 공간에 내재한 저차원 공간이므로 학습 데이터를 저차원 매니폴드 공간에 표현한다. 2) 매끄러움 가정 = 샘플은 어떤 요인에 의해 변화한다. 카메라의 위치, 조명 등 획득한 특징 공간에서 위치가 조금씩 바뀐. <p><6.2절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환></p> <ul style="list-style-type: none"> 군집화 = 유사한 샘플을 모아 같은 그룹으로 묶는 일. Ex) 영상 분할, 맞춤 광고 밀도 추정 = 데이터로부터 확률분포를 추정하는 일. Ex) 분류, 생성 모델 구축 공간 변환 = 원래 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일(매니폴드) Ex) 데이터 가시화, 데이터 압축, 특징 추출. |

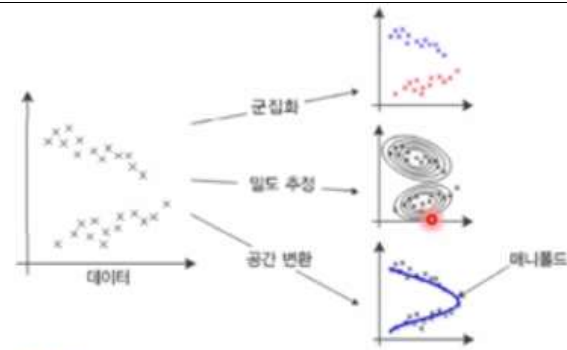


그림 6-2 비지도 학습의 군집화, 밀도 추정, 공간 변환 과정이 발견하는 정보

<6.3절 k-평균과 친밀도 전파 알고리즘>

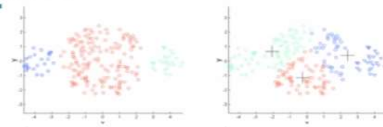
- k-평균 = x 개의 데이터 셋에 k 개의 군집단을 찾아내는 작업. 군집의 개수 k 는 주어지는 경우와 자동으로 찾아야 하는 경우가 있음. 군집화를 부류 발견 직업이라 부르기도 한다.

$$\left. \begin{array}{l} c_i \neq \emptyset, i = 1, 2, \dots, k \\ \bigcup_{i=1}^k c_i = X \\ c_i \cap c_j = \emptyset, i \neq j \end{array} \right\}$$

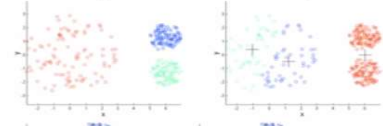
- K-평균은 샘플의 평균으로 군집 중심을 갱신
k-medoids 는 대표를 뽑아 뽑힌 대표로 군집 중심을 갱신한다(k-평균에 비해 잡음에 둔감).
- k-평균 알고리즘에서 초기 군집 중심이 달라지면 최종 결과가 달라진다. 다중 시작은 서로 다른 초기 군집 중심을 가지고 여러 번 수행한 다음, 가장 좋은 품질의 해를 취함(오차가 적음).
- K-평균은 군집 크기 불균형, 군집 밀도 차이, 특이한 분포로 인해 한계가 발생한다.

6.3.1 k-평균 알고리즘 한계

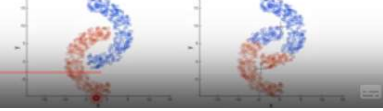
■ 군집 크기 불균형



■ 군집 밀도 차이



■ 특이한 분포



30.01 / 1:01:33

- EM 기초 = k-평균에서 훈련집합과 군집집합은 각각 입력단과 출력단에서 관찰 가능하다. K-평균은 z의 추정과 a의 추정을 번갈아 가면서 수행하는 EM 알고리즘이다.

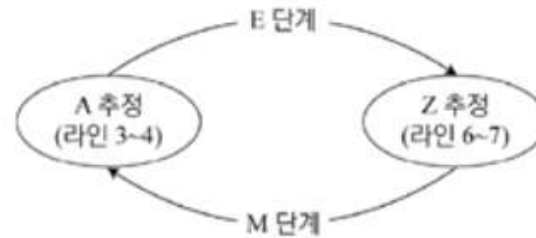


그림 6-6 k-평균을 EM 알고리즘으로 해석

- 친밀도 전파 알고리즘 = 책임 행렬 R과 가용 행렬 A의 계산.
- k번째 데이터가 i번째 데이터의 대표가 되어야 한다는 증거 =>

$$r_{ik} = s_{ik} - \max_{k' \neq k} (a_{ik'} + s_{ik'}) \quad s(i, k) = -\|x_i - x_k\|^2$$

$$a_{ik} = \min \left(0, r_{kk} + \sum_{i' \neq i, k} \max(0, r_{i'k}) \right), i \neq k$$

i번째 데이터가 k번째 데이터를 대표로 선택해야 한다는 증거 =>

<6.4 절 커널 밀도 추정과 가우시안 혼합>

- 밀도 추정 문제 = 어떤 점 x 에서 데이터가 발생할 확률, 확률분포 $P(x)$ 를 구하는 문제.
커널 밀도 추정에서 히스토그램 방법이 있는데 특정 공간을 칸의 집합으로 분할한 다음, 칸에 있는 샘플의 빈도를 세어 추정한다. 그러나 칸의 크기와 위치에 민감하고 매끄럽지 못하며 계단모양을 띠는 확률밀도함수가 되는 단점이 존재. 커널 밀도 추정의 근본적인 문제는 샘플을 모두 저장하고 있어야 하는 메모리 기반 방법이고 데이터의 희소성이 존재할 뿐만 아니라 데이터가 낮은 차원의 경우로 국한하여 활용해야 한다.
- 그걸 보완시킨 방법이 가우시안 혼합이다. 데이터가 가우시안 분포를 따른다고 가정하고 평균 벡터와 공분산 행렬을 추정한다. 대부분 데이터가 하나의 가우시안으로 불충분하다.

$$\left. \begin{aligned} P(\mathbf{x}) &= N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|} \sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ \text{이때 } \boldsymbol{\mu} &= \frac{1}{n} \sum_{i=1, n} \mathbf{x}_i, \boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1, n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \end{aligned} \right\} \quad (6.9)$$

- EM 알고리즘
가우시안으로 샘플의 소속 정보를 개선(E 단계) -> 샘플의 소속 정보로 가우시안 개선(M 단계) -> 가우시안으로 샘플의 소속 정보 개선(E 단계) -> 샘플의 소속 정보로 가우시안 개선(M 단계)... 이런 과정 반복.

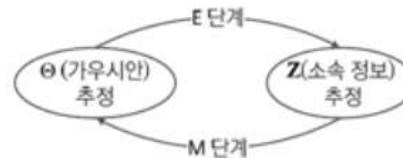


그림 6-15 가우시안 혼합을 위한 EM 알고리즘

질문 내용

EM 알고리즘은 Greedy 알고리즘인가??

-201403474 최진성

| 구분 | 내용 |
|-------|--|
| 학습 범위 | <p>기계학습 6장 비지도 학습</p> <p>6.1 지도 학습과 비지도 학습, 준지도 학습</p> <p>6.2 비지도 학습</p> <p>6.3 군집화</p> <p>6.4 밀도 추정</p> |
| 학습 내용 | <p>기계학습 6장 비지도 학습</p> <p>지도학습은 훈련집합 X와 Y가 제시되는 지도 학습</p> <p>비지도 학습은 X만 주어지는 학습</p> <ul style="list-style-type: none"> ● 현대 기계 학습에서 매우 중요한 위치를 차지. <p>6.1 지도 학습과 비지도 학습, 준지도 학습</p> <p>지도 학습 : 모든 훈련 샘플이 레이블 정보를 가짐</p> <p>비지도 학습 : 모든 훈련 샘플이 레이블 정보를 가지지 않음.</p> <p>준지도 학습 : 레이블을 가진 샘플과 가지지 않은 샘플이 섞여있음.</p> <ul style="list-style-type: none"> ● 매니폴드 가정 - 데이터 집합은 하나 이상의 매니폴드를 구성, 모든 샘플은 매니폴드와 가까운 곳에 존재 ● 매끄러움 가정 - 샘플은 어떤 요인에 의해 변화하며 획득한 샘플은 특징 공간에서 위치가 바뀌는 식으로 변화한다. 이 때 매끄러운 곡면을 따른다. <p>6.2 비지도 학습</p> <p>일반 과업</p> <ul style="list-style-type: none"> ● 군집화 : 유사한 샘플을 모아 같은 그룹으로 묶음 |

- 밀도 추정 : 데이터로부터 확률 분포를 추정
- 공간 변환 : 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일.

응용 과업

- 군집화 : 맞춤 광고, 영상 분할, 유전자 데이터 분석, 실시간 검색어 분석 등
- 밀도 추정 : 분류, 생성모델 구축 등
- 공간 변환 : 데이터 가시화, 압축, 특징 추출 등.

6.3 군집화

$X = \{x_1, x_2, \dots, x_n\}$ 에서 다음식을 만족하는 군집 집합 $C = \{c_1, c_2, \dots, c_n\}$ 를 찾아내는 작업

군집화 식

$$\left. \begin{array}{l} c_i \neq \emptyset, i = 1, 2, \dots, k \\ \bigcup_{i=1}^k c_i = X \\ c_i \cap c_j = \emptyset, i \neq j \end{array} \right\}$$

- 군집화는 주관성이 뛰어나기 때문에 같은 훈련 집합에서도 여러 군집 집합이 나올 수 있다.

K-평균 알고리즘 : 샘플의 평균으로 군집 중심을 갱신. - 목적 함수를 최소화 하는 알고리즘

K-Medoids : 대표를 뽑아 뽑힌 대표로 군집 중심을 갱신(잡음에 둔감) -

다중 시작 K-평균 알고리즘 : 서로 다른 초기 군집 중심을 가지고 여러 번 수행한 다음, 가장 좋은 품질의 해를 취함.

➔ 군집 크기의 불균형, 군집 밀도의 차이, 특이한 분포 등의 한계점이 있음.

EM - 기댓값 최대화 알고리즘 : 모수에 대한 추정값으로 로그 우도의 기댓값을 계산(E) + 기댓값을 최대화하는 모수 측정값들을 구하는 최대화 단계(M)를 번갈아가면서 적용.

친밀도 전파 알고리즘

책임 행렬 R과 가용 행렬 A라는 두 종류의 친밀도 행렬을 이용하여 군집화
군집 개수 k를 자동으로 알아냄

➔ 가까울수록 유사도가 증가하고 멀수록 유사도가 감소한다.

$$s_{ik} = -||x_i - x_k||_2^2, i \neq k, (i, k = 1, 2, \dots, n)$$

예)

i와 k, k'라는 parameter가 있을 때 k가 k'보다 i에게 가까울 경우 k는 i에게 대표 데이터로서의 가산점을 부여
하지만 i가 k'로부터 대표 데이터로서의 가산점을 받고 있을 경우 i는 굳이 k에게 대표 데이터 가산점을 부여하지 않는다.
받는 대표 가산점이 많을수록 다른 데이터에게 영향력을 행사하기 수월해진다.

- 자기 유사도 s_{kk} - 유사도의 최솟값(적은 군집), 중앙값(중간 군집), 최댓값(많은 군집) 중에서 선택
- 자가 친밀도 r_{kk}, a_{kk} - r_{ss} 는 친밀도 전파 알고리즘에서 사용하는 식을 그대로 사용한다. a_{kk} 는 새로운 식 사용.

6.4 밀도 추정

어떤 점 x에서 데이터가 발생할 확률 - 확률 분포 P(x)를 구하는 문제.

- 데이터 분포의 밀도에 따라 특정 위치를 지목하였을 때 데이터 생성 확률을 구할 수 있다.

커널 밀도 추정

- 히스토그램 - 특징 공간을 칸의 집합으로 분할한 다음, 칸에 있는 샘플의 빈도를 세어 식으로 추정
- $P(x) = \frac{bin(x)}{n}$
 - ➔ 매끄럽지 못하고 계단 모양을 띠는 확률 밀도 함수가 됨.(단, 많은 수의 시행을 거듭할 경우 정밀하게 될 순 있음.)
 - ➔ 칸의 크기와 위치에 민감하여 밀도 추정이 매번 바뀌게 될 수 있다.
- 커널 밀도 추정법 - 점 x에 예시 커널(ex: 가우시안 커널)을 씌우고 커널 안에 있는 샘플의 가중 합을 이용함.

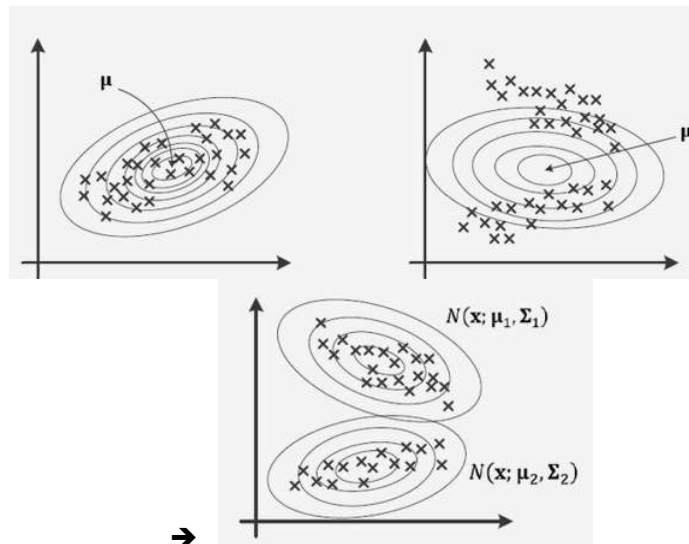
$$P_h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \left. \vphantom{\sum_{i=1}^n} \right\}$$

여기서 $K_h(\mathbf{x}) = \frac{1}{h^d} K\left(\frac{\mathbf{x}}{h}\right)$

- ➔ 대역폭 h 의 크기가 중요. - h 가 너무 작으면 파형이 너무 튀고 그렇다고 너무 작으면 파형이 뭉개진다.
- ➔ 매끄러운 확률밀도함수를 추정함
- ➔ 샘플을 모두 저장하고 있어야 하는 메모리 기반 방법이기 때문에 새로운 샘플이 주어질 때 마다 계산을 다시 한다.
- ➔ 데이터 희소성에 따른 차원의 저주 문제가 발생한다. -> (데이터가 낮은 차원인 경우로 국한하여 활용함)

가우시안 혼합

가우시안을 이용한 방법(모수적 방법) - 데이터가 가우시안 분포를 따른다고 가정, 평균 벡터 μ 와 공분산 행렬 Σ 를 추정함.
대부분 데이터가 하나의 가우시안으로 불충분하다.(오른쪽)



$$P(\mathbf{x}) = \sum_{j=1}^k \pi_j N(\mathbf{x}; \mu_j, \Sigma_j)$$

| | |
|-------|--|
| | <p>→ 주어진 데이터 – 훈련 집합 $X = \{x_1, x_2, \dots, x_n\}$, 가우시안의 개수 k</p> <p>→ 매개변수 집합 : $\theta = \{\pi = (\pi_1, \pi_2, \dots, \pi_k), (\mu_1, \Sigma_1), (\mu_2, \Sigma_2), \dots, (\mu_k, \Sigma_k)\}$</p> <p>→ 최대 우도를 이용한 최적화 – $\log P(X \theta) = \sum_{i=1}^n \log \left(\sum_{j=1}^k \pi_j N(x; \mu_j, \Sigma_j) \right)$ -> $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log P(X \theta)$</p> <p>EM 알고리즘을 이용할 경우</p> <p>θ를 모르므로 난수로 넣고 풀이하게 됨</p> <ul style="list-style-type: none"> - 가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계) -> 가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계) -> 가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계)... |
| 질문 내용 | |

| 구분 | 내용 |
|-------|---|
| 학습 범위 | <p>6.1절: 비지도 학습을 지도학습, 준지도 학습과 비교한다.</p> <p>6.2절: 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환을 소개한다.</p> <p>6.3절: 군집화 알고리즘으로 k-평균과 친밀도 전파 알고리즘을 설명한다.</p> <p>6.4절: 밀도 추정 방법으로 커널 밀도 추정과 가우시안 혼합을 설명한다.</p> |
| 학습 내용 | <ul style="list-style-type: none"> - 기계학습이 사용하는 두 종류의 지식: 훈련집합, 사전지식 - 데이터 자체는 매니폴드라 가정한다. - 비지도 학습: 군집화, 밀도 추정, 공간 변환으로 많은 곳에서 응용된다. - k-means는 평균값을 구하는 연산을 수행하기 때문에 잡음이나 이상치(아웃라이어)에 민감하다. 이러한 단점을 해결하기 위해 나온것이 k-medoids 알고리즘이다. k-medoids는 클러스터의 대표값으로 오브젝트의 중심점을 구하는 것이 아니라, 오브젝트 중에서 클러스터를 대표할 수 있는 가장 가까운 대표 오브젝트를 뽑는다. 대표로 뽑히지 않은 나머지 오브젝트는 가장 가까운 대표 오브젝트를 따라 해당 클러스터에 배정한다. - k-means : 평균을 대표 값으로 가져가서 분산을 기준으로 알고리즘 진행 - k-medoids: 중앙값을 대표 값으로 가져가므로 절대오차를 기준으로 알고리즘 진행 <div data-bbox="571 933 965 1141"> </div> <ul style="list-style-type: none"> - 밀도추정: 데이터는 어떤 변수가 가질 수 있는 다양한 가능성 중의 하나가 현실 세계에 구체화된 값이다. 그리고 우리는 이렇게 관측된 데이터들을 통해 그 변수(random variable)가 가지고 있는 본질적인 특성을 파악하고자 노력한다. 그러나 하나의 데이터는 변수의 일면에 불과하기 때문에 변수의 진면목을 파악하기 위해서는 많은 수의 데이터가 필요하다. 그리고 이렇게 얻어진(관측된) 데이터들의 분포로부터 원래 변수의 (확률) 분포 특성을 추정하고자 하는 것이 density estimation(밀도추정)이다. |

| | |
|-------|---|
| | <ul style="list-style-type: none"> - 밀도 추정에는 미리 pdf(probability density function)에 대한 모델을 정해놓고 데이터들로부터 모델의 파라미터만 추정하는 Parametric 방식과, 모델이 미리 주어지지 않고 순수하게 관측된 데이터 만으로 확률밀도 함수를 추정하는 non-parametric 방식이 있다. - 히스토그램이 가진 문제점 1) 불연속성이 나타난다. 2) 고차원 데이터에는 메모리 문제등으로 사용하기 힘들다. 3) 딱딱하다. - 커널 밀도 추정: non-parametric 밀도추정 방법 중 하나로서 커널함수(kernel function)를 이용하여 히스토그램 방법의 문제점을 개선한 방법이다 - 커널 함수: 원점을 중심으로 대칭이면서 적분값이 1인 non-negative 함수-> 가우시안, uniform 함수 등등.. 즉, 커널 밀도 추정은 아래와 같다. <ol style="list-style-type: none"> 1. 관측된 데이터 각각마다 해당 데이터 값을 중심으로 하는 커널 함수를 생성한다: $K(x-x_i)$ 2. 이렇게 만들어진 커널 함수들을 모두 더한 후 전체 데이터 개수로 나눈다. - 커널 밀도 추정의 문제점: 샘플을 모두 저장하고 있어야하는 메모리 기반 방법, 데이터 희소성 - 위의 문제점을 해결하기 위한 방법이 가우시안 혼합이다. - Log는 monotonic 하게 increasing 되는 함수이다. - K-means 군집화는 각 지에 대해서 둘 사이의 수직한 분배를 했다면, EM 알고리즘은 그것보다 부드럽게 소속된 데이터들을 가지고 아래의 과정으로 모델링한다. <div data-bbox="600 949 996 1101" data-label="Diagram"> </div> <p>그림 6-15 가우시안 혼합을 위한 EM 알고리즘</p> |
| 질문 내용 | |

-201500629 김영연

| 구분 | 내용 |
|-------|---|
| 학습 범위 | <p>6.1절: 비지도 학습을 지도학습, 준지도 학습과 비교한다.</p> <p>6.2절: 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환을 소개한다.</p> <p>6.3절: 군집화 알고리즘으로 k-평균과 친밀도 전파 알고리즘을 설명한다.</p> <p>6.4절: 밀도 추정 방법으로 커널 밀도 추정과 가우시안 혼합을 설명한다.</p> |
| 학습 내용 | <ul style="list-style-type: none"> ● 지도 학습: 속이 꼭 찬 레이블로 구성 <p>샘플은 (x_i, y_i) 형태로 주어지는데 기계학습이 할 일은 특징벡터 x_i를 입력 받아 정확하게 레이블 y_i를 출력하는 예측기를 만드는 것이다.</p> <p>지도 학습은 레이블 정보를 받으므로, 목적함수가 학습과정을 주도한다.</p> ● 비지도 학습: 레이블이 없는 샘플만 주어진다. <p>입력만 있고 출력은 없는 상황에서 수행하는 학습</p> ● 준지도 학습: 레이블이 없는 샘플도 소속 부류 정보를 모를 뿐이지, 어떤 부류에서 생성되었다는 사실은 확실하며 분명 소속 부류의 확률 분포를 따를 것이다. 따라서 이러한 사실을 잘 이용하면 레이블이 없는 샘플도 학습에 참여하여 성능 향상에 공헌할 수 있는데 이것을 준지도 학습이라고 한다. ● 사전지식: 훈련집합이라는 명시적인 정보뿐만 아니라 세상의 일반적인 규칙으로부터 얻을 수 있는 암시적인 정보 <p>매니폴드 가정: 데이터 집합은 하나의 매니폴드 또는 여러 개의 매니폴드를 구성하며, 모든 샘플은 매니폴드와 가까운 곳에 있다.</p> |

매끄러운 가정: 샘플은 어떤 요인에 의하여 변화한다. 그 때 매끄러운 곡선을 따라 위치가 변하는 것을 의미

- 비지도 학습의 일반 과업

1. 군집화: 특징 공간에서 가까이 있는 샘플을 모아 그룹으로 묶는 일이다.
2. 말도 추정: 데이터로부터 확률분포를 추정하는 일이다. 데이터가 조밀하게 분포한 곳은 높은 확률을 배정하고, 희소하게 분포한 곳은 낮은 확률로 배정해야 한다.
3. 공간 변환: 데이터가 정의된 원래 특징 공간을 저차원 공간 또는 고차원 공간으로 변환하는 일로 새로운 공간은 주어진 목적을 달성하는데 유리해야 한다.

- 군집화: 훈련집합 $X=\{x_1, x_2, \dots, x_n\}$ 이 주어지고 세가지 조건을 만족하는 군집집합 $C=\{c_1, c_2, \dots, c_n\}$ 을 찾아내는 작업이다.

1. $c_i \neq \emptyset, i = 1, 2, \dots, k$ 모든 군집이 하나 이상의 샘플을 가진다.
2. $\bigcup_{i=1}^k c_i = X$ 모든 샘플이 단 하나의 군집에 속한다.
3. $c_i \cap c_j = \emptyset, i \neq j$

군집화는 군집의 개수를 부류의 개수로 간주할 수 있으므로 부류 발견 작업이라 한다. 이는 주관성이 개입될 수 있다.

- 경성 군집화: 한 샘플이 하나의 군집에 속하도록 하는 방식
- 연성 군집화: 샘플마다 군집에 속하는 정도를 다르게 할 수 있다.

- K-평균 알고리즘: 원리가 단순하지만, 성능이 좋아 인기 좋은 군집화 알고리즘이다.
- K-medoids: 샘플 중에서 대표를 뽑고 뽑힌 대표로 군집 중심을 갱신한다. 대표를 뽑는 방법은 여러가지가 있는데, 다른 샘플까지의 거리의 합이 최소가 되는 샘플을 대표로 정하는 방법을 많이 사용한다. K-평균에 비해 잡음에 둔감되는 장점이 있다.
- K 평균이 사용되는 목적함수: $J(Z, A) = \sum_{i=1}^n \sum_{j=1}^k a_{ij} dist(x_i, z_j)$ (Z: 군집 중심, A: 샘플의 배정 정보를 나타내는 k*n 행렬)
- 다중 시작 k-평균: 서로 다른 초기 군집 중심을 가지고 k-평균을 여러 번 수행한 다음, 가장 좋은 품질의 해를 선택하는 전략을 사용한다.
- 은닉변수: 중간에 임시로 사용되다 사라지는 변수를 의미
- EM 알고리즘: 은닉변수의 추정과 매개변수 추정을 번갈아 수행하면서 최적의 해를 찾는 과정
- 친밀도 전파 알고리즘: 샘플 간의 유사도로부터 책임행렬 R과 가용행렬 A라는 두 종류 친밀도 행렬을 계산하고, 이 친밀도 정보를 이용해서 군집을 찾는 알고리즘
 $s_{ik} = -\|x_i - x_k\|_2^2$ i와 k가 유사할수록 큰 값을 부여하고, i가 k이외의 다른 샘플과 더 친밀할수록 더 큰 값을 뺀다.
 $r_{ik} = s_{ik} - \max(a_{ik'}, s_{ik})$
- 밀도 추정: 어떤 점 x에서 데이터가 발생할 확률, 즉 확률변수 밀도 P(x)를 구하는 문제
- 히스토그램 방법: 각 차원을 여러 구간으로 나누어 특정 공간을 칸의 집합으로 분할한 다음, 각각의 칸에 있는 샘플의 빈도를 세는 것

| | |
|-------|---|
| | <p>단점 1. 밀도 추정의 최종 모양은 막대의 시작점과 막대 폭에 의존한다.</p> <p>2. 다변량 데이터에 대해서도 밀도의 최종 모양은 막대의 시작점 값에 영향을 받는다</p> <p>3. 추정된 밀도가 불연속적이며 매끄럽지 못하다.</p> <ul style="list-style-type: none"> ● 커널 밀도 추정: 데이터들을 각각 커널 함수에 입력하여 합친 형태를 통해 밀도 추정을 하는 것 커널의 대역폭을 적절히 설정하는 것이 중요 ● 모수적 밀도 추정: 주어진 데이터의 분포형태를 특정 분포로 가정하고 나서 데이터 외의 숨겨진 확률함수를 모델링하는 것 ● 가우시안 혼합: 데이터가 일정한 모양의 분포를 따른다는 가정하에 확률분포를 추정한다. |
| 질문 내용 | |