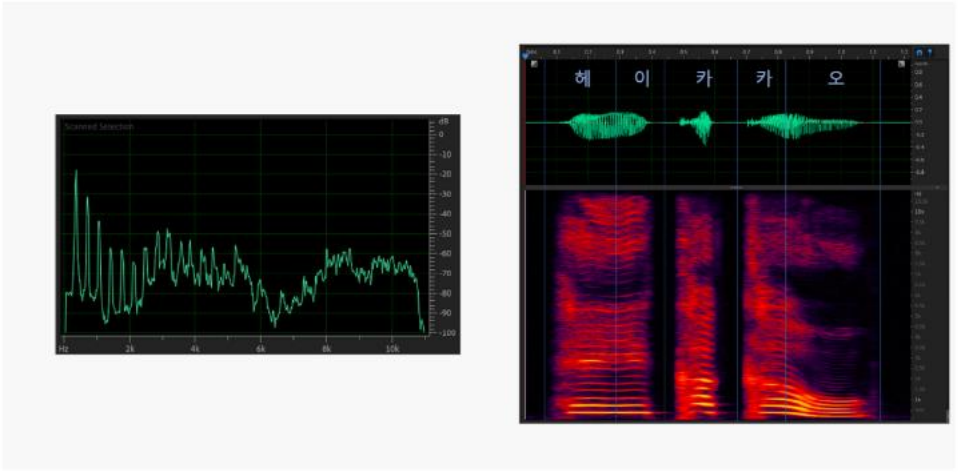
13주차 일석이조 조별보고서	
작성일 : 2023년 9월 24일	작성자 : 유정훈
조 모임 일시 : 9월 24일	모임 장소 : 구글미트
참석자 : 이준용, 유정훈, 김동규, 이학빈, 탁성재	조원 : 이준용, 유정훈, 김동규, 이학빈, 탁성재

구분	내용
학습범위와 내용	1. 13주차 온라인 강의 내용
질문 내용 (모임 전 공지된 개별 학습 범위에서 이해된 것과 못한 것)	Q(1) EM알고리즘은 Greedy 알고리즘인가? A(1) 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 최적의solution을 찾는 것이다. 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. EM은 greedy 알고리즘으로 눈 앞에 놓인 최적점을 찾으려 노력할 뿐 global 최적점은 찾지 못한다는 단점이 있다. Q(2) 블라인드 원음 분리에서의 푸리에 변환 A(2) 푸리에 변환의 기본 원리는 소리에 대한 함수를 시간 또는 공간 주파수의 합성함수로 표현이 가능하다는 것입니다. 아무런 방해가 없는 공간에서의 소리는 정확하게 함수로 표현이 가능하지만, 실제 공간에서는 감쇠, 왜곡, 잡음등에 의해 정확한 내용을 파악할 수 없습니다. 가장 대표적인 분리 방법으로 ica를 사용합니다 ica는 해당 구성 음원들이 모두 서로 다른 소리 벡터를 가지고 있는 독립적인 함수라고 가정합니다. 예를 들어 동일한 악기는 하나의 음을 연주할 때, 동일한 주파수 벡터를 가지고, 코드를 진행할때 정수배의 주파수를 추가로 가지는 특징을 가집니다. 하지만 이 악기가 아닌 다른 악기의 연주에는 또다른 주파수 벡터를 가지게 됩니다. 또다른 예시로는 음성인식을 들 수 있습니다. 우리가 어떤 단어를 이야기 할 때, 해당 단어를 표현하기 위해 발생하는 주파수의 스펙트럼들이 있습니다. 이렇게 소리가 가진 데이터를 스펙트로그램, 즉 진폭과 위상이라고 부릅니다.



[그림 4] '이' 발음에 해당하는 스펙트럼(왼쪽)과 '헤이 카카오' 음성에서 추출한 스펙트로그램(오른쪽)[03]

녹색-스펙트럼, 빨강-스펙트로그램

Q(3)

커널 밀도 추정이란?

A(3)

커널 함수란 다음 3가지 조건을 모두 만족하는 함수를 의미한다. (1) 적분값이 1이며, (2) 원점을 중심으로 대칭인 (3) Non-negative인 경우, 이를 커널 함수라고 한다.

$$(1)\ \int_{-\infty}^{\infty} K(u)du = 1$$

$$(2)\ K(-u) = K(u) for\ all\ values\ of\ u.$$

$$(3)\ Non\text{-}negative$$

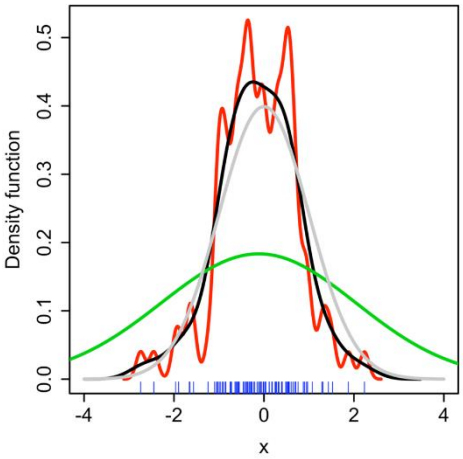
Kernel Functions, $K(u)$		
Uniform ("rectangular window")	$K(u) = \frac{1}{2}$ Support: $\lvert u\rvert \le 1$	"Boxcar function"
Triangular	$K(u) = (1 - \lvert u\rvert)$ Support: $\lvert u\rvert \le 1$	
Epanechnikov (parabolic)	$K(u) = \frac{3}{4}(1 - u^2)$ Support: $\lvert u\rvert \le 1$	
Quartic (biweight)	$K(u) = \frac{15}{16}(1 - u^2)^2$ Support: $\lvert u\rvert \le 1$	
Triweight	$K(u) = \frac{35}{32}(1 - u^2)^3$ Support: $\lvert u\rvert \le 1$	
Tricube	$K(u) = \frac{70}{81}(1 - \lvert u\rvert^3)^3$ Support: $\lvert u\rvert \le 1$	
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	
Cosine	$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right)$ Support: $\lvert u\rvert \le 1$	

KDE는 커널 함수와 데이터를 바탕으로 연속성 있는 확률 밀도 함수를 추정하는 것이다. 아래 수식을

간단히 설명해 보자면, 관측된 데이터마다 해당 데이터를 중심으로 하는 커널 함수를 생성한 후, 해당 커널 함수를 모두 더하고 데이터 개수로 나누면 KDE로 도출된 확률밀도함수를 구할 수 있다. 아래 수식에서 x는 확률 변수를 의미하고, x_i는 관측된 데이터 포인트 하나를 의미한다.

$$\widehat{f_h}(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathrm{x} - x_i) = \frac{1}{nh} K\left(\frac{\mathrm{x} - x_i}{h}\right)$$

위의 수식에서 h값이 대역폭(bandwidth)를 결정하는 파라미터이다. 대역폭은 확률밀도함수를 스무딩(smoothing)하는 역할을 하며, 대역폭의 값이 작을수록 KDE의 모양이 뾰족하고 클수록 완만한 형태를 띠게 된다. 아래의 위키피디아 예시에서 대역폭에 따른 KDE의 형태를 비교하여 볼 수 있다.



- 빨간색 선: h=0.05
- 검은색 선: h=0.337
- 초록색 선: h=2

Q(4)

가우시안 혼합 모델에 대해 알아보았습니다.

A(4)

가우시안 혼합 모델(Gaussian Mixture Model, GMM)은 비지도 학습의 일종으로, 데이터를 여러 개의 가우시안 분포를 조합하여 모델링하는 확률적 생성 모델이다. 각 데이터 포인트가 여러 개의 가우시안 분포 중 어느 하나에서 생성되었을 가능성을 확률적으로 나타낸다.

가우시안 혼합 모델은 주어진 데이터가 여러 개의 가우시안 분포를 가진 군집으로 구성되어 있다고 가정한다. 각 군집은 다른 평균과 분산을 가진 가우시안 분포를 나타낸다. GMM은 이러한 가우시안 분포들의 조합으로 데이터를 모델링한다.

가우시안 혼합 모델은 크게 세 가지 요소로 구성된다

1. 가우시안 분포들의 조합

여러 개의 가우시안 분포(또는 클러스터)를 조합하여 데이터를 모델링한다. 각각의 가우시안 분포는 데이터의 특정 부분을 대표한다.

2. 모수(Parameter)

각 가우시안 분포마다 평균과 분산을 가지고 있다. 이러한 평균과 분산은 모델의 파라미터로, EM(Expectation-Maximization) 알고리즘 등을 사용하여 추정된다.

3. EM 알고리즘

EM 알고리즘을 사용하여 모수를 추정한다. EM 알고리즘은 가우시안 혼합 모델에서 숨겨진(latent) 변수(각 데이터 포인트가 어떤 클러스터에 속하는지의 정보)를 활용하여 모델을 학습한다.

가우시안 혼합 모델은 주어진 데이터를 여러 개의 가우시안 분포를 통해 모델링하므로, 이 모델을 사용하여 데이터의 군집화(clustering)를 수행할 수 있다. 또한 새로운 데이터가 주어졌을 때 해당 데이터가 각 가우시안 분포에서 발생할 확률을 계산하여 이상치 탐지나 데이터 포인트의 소속 여부를 추론하는 데 사용될 수 있다.

Q(5)

매니폴드 가정과 매끄러움 과정에 대해 더 알아보았습니다.

A(5)

"매니폴드 가정"과 "매끄러움 가정"은 주로 수학과 물리학 분야에서 사용되는 개념으로, 다양한 응용 분야에서 등장합니다. 각각의 특징과 공통점에 대해 간단히 설명하겠습니다.

1. **매니폴드 가정 (Manifold Hypothesis):**
    - **특징:**
        - 매니폴드는 국소적으로 유클리드 공간과 닮은 특징을 갖는 공간으로 생각됩니다.
        - 복잡한 데이터를 간소화하고 모델링하기 위해 사용됩니다.
        - 데이터의 차원을 줄이거나 특징을 추출하는 데에 활용됩니다.
    - **용도:**
        - 머신 러닝 및 패턴 인식에서 차원 축소나 특징 추출에 사용됩니다.
        - 데이터의 내재된 구조를 파악하고 이해하는 데 도움이 됩니다.
2. **매끄러움 가정 (Smoothness Hypothesis):**
    - **특징:**
        - 매끄러움 가정은 함수나 매핑이 근처의 입력에 대해 작은 변화에 대해 작은 출력 변화를 가진다는 가정입니다.
        - 즉, 함수의 변화가 극단적이지 않고 매끄럽게 일어난다고 가정합니다.
    - **용도:**
        - 물리학에서는 자연 현상을 모델링하고 설명하는 데 사용됩니다.
        - 수치해석, 최적화, 미분 방정식 등과 같은 수학적 문제 해결에 적용됩니다.
3. **공통점:**
    - 둘 다 데이터나 현상을 모델링하고 이해하기 위한 가정으로 사용됩니다.
    - 수학적이고 이론적인 배경을 갖고 있으며, 주로 고차원 데이터나 복잡한 구조를 처리하는 데 활용됩니다.
4. **차이점:**
    - 매니폴드 가정은 데이터의 내재된 구조를 특정한 형태의 공간으로 가정하고 모델링하는 데 중점을 둡니다.

	○ 매끄러움 가정은 함수나 매핑의 부드러운 특성을 강조하며, 입력의 작은 변화가 출력에 작은 영향을 미친다고 가정합니다. 이러한 가정들은 각각의 분야에서 모델링과 예측을 향상시키는 데 사용되며, 데이터 과학, 기계 학습, 물리학, 수학 등에서 널리 적용되고 연구되고 있습니다.
질문내용	Q1: 라벨을 부여하기 위해선 확률 분포가 필요한데, 확률 분포를 얻기 위해선 모수를 알아야 하고, 모수를 알기 위해선 다시 각 데이터에 라벨이 필요하다. 예를 들어 우리가 처한 문제는 라벨이 없는 데이터들이 주어졌다는 점이었으며, 우리가 필요한 해답은 각 라벨 별 분포가 필요로 한다. 이 문제가 어려운 이유는 라벨을 얻기 위해선 분포가 필요하고, 분포를 얻기 위해선 라벨이 필요하기 때문이었다. 문제의 해결을 위해 우리는 데이터 셋들이 정규분포를 이룰 것이라 가정했다. 그런 뒤, 랜덤하게 모수를 주어준 뒤 라벨을 얻고, 그 라벨들을 이용해 다시 분포를 얻는 방식으로 clustering을 수행한다면 해결 할 수 있다고 생각이 들었는데 맞는 생각 일까요?

학과	컴퓨터 전자시스템 공학	학번	201904458	이름	이준용
구분	내용				
학습 범위	기계학습 6장 비지도 학습 6.1 절 비지도 학습을 지도/준지도 학습과 비교 6.2 절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환 6.3 절 k-평균과 친밀도 전파 알고리즘 6.4 절 커널 밀도 추정과 가우시안 혼합				
학습 내용	**기계학습 6장 비지도 학습** **6.1 절 비지도 학습을 지도/준지도 학습과 비교** ✓ 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 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 바뀜. **6.2 절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환** 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.				

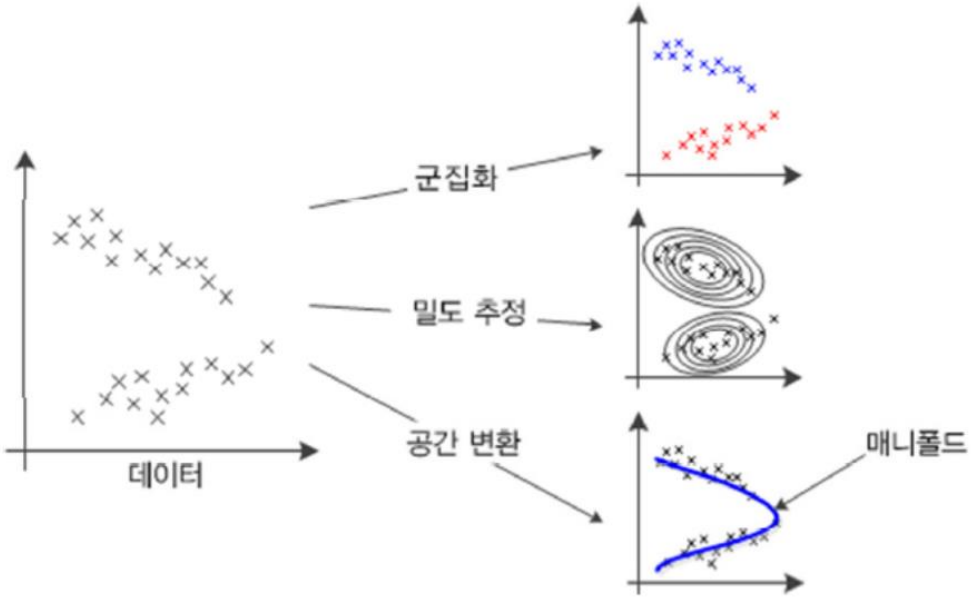


그림 6-2 비지도 학습의 군집화, 밀도 추정, 공간 변환 과업이 발견하는 정보

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.

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

- ✓ 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 적은).
- ✓ K-평균은 군집 크기 불균형, 군집 밀도 차이, 특이한 분포로 인해 한계가 발생한다.

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

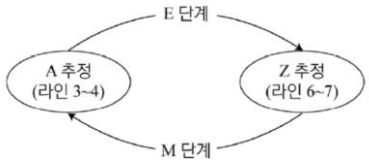


그림 6-6 *k*-평균을 EM 알고리즘으로 해석

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 ≠ k, (i, k = 1, 2, ..., n)

예) i와 k, k'라는 parameter가 있을 때 k가 k'보다 i에게 가까울 경우 k는 i에게 대표 데이터로서의 가산점을 부여하지만 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

6.4 절 커널 밀도 추정과 가우시안 혼합

밀도 추정 문제 = 어떤 점 x에서 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 게단 모양을 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.

$$P(\mathbf{x}) = N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\text{이때 } \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1,n}\mathbf{x}_i\,,\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1,n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$$

EM 알고리즘을 이용할 경우
θ를 모르므로 난수로 넣고 풀이하게 됨

	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)…

학과	컴퓨터전자시스템 공학과	학번	201800615	이름	김동규
구분	내용				
학습범위	기계학습 6장 6.1 비지도 학습을 지도, 준지도 학습과 비교 6.2 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환 6.3 k-평균과 친밀도 전파 알고리즘 6.4 커널 밀도 추정과 가우시안 혼합				
학습내용	지도 학습 주어진 데이터들의 분류 정보를 미리 줌 비지도 학습 주어진 데이터들의 분류정보를 주지 않음 준지도 학습 일부 데이터들의 분류 정보만 주어짐 비지도 학습의 분류 방법 군집화 유사한 데이터들의 묶음 생성 기준을 어떻게 잡는지에 따라 갯수, 형태가 달라짐 밀도 추정 데이터가 모여있는 곳에 대해 확률 분포 설정 공간변환 특정 차원에서 데이터를 확인해, 어느 공간에서 선형적으로 데이터가 분류 가능 한지 확인 K- 평균 샘플의 평균 위치를 계속 갱신하며 특정 위치로 수렴시킴 초기 군집 중심에 따라 결과가 달라지므로, 여러 종류의 초기 군집을 가지고 최상의 결과 만을 반영함 친밀도 전파 알고리즘 어떤 특정 값과의 거리, 벡터가 비슷할수록 큰 값을 가짐 책임행렬 R(r ik) i와 k번째의 값이 비슷하면 큰 값을 가짐				

$$r_{ik} = s_{ik} - \max_{k' \neq k}(a_{ik'} + s_{ik'}) \qquad s(i,k) = -\|x_i - x$$

가용 행렬 A(a ik') 현재 k번째와 다른 값인 k'과의 값이 비슷하면 큰 값을 가짐

i번째 값과 k번째 값과의 거리를 유클리드 거리로 구한 후 절댓값과 음수화를 시킴

서로 차이가 작으면 결과값은 더욱 커짐

$$s_{ik} = -\|\mathbf{x}_i - \mathbf{x}_k\|_2^2, \qquad i \neq k\text{이고 } i,k = 1,2,\cdots,n$$

자가 유사도

유사도의 최솟값, 중앙값, 최댓값 중에 하나를 고름(자기 자신을 표현하는 값)

밀도 추정

어떤 특정 위치(상황)에서 해당 데이터가 발생할 확률

히스토그램 방법

특징 공간을 칸의 집합으로 분할한 후, 해당 케이스에서 발생한 횟수를 세어서 계산 커널 밀도 추정을 통해 부드럽게 표현 가능함

해당 데이터의 대역폭(h)을 지정해, 현재 데이터들, 대역폭에 대한 밀도를 추정 가능함

대역폭(h)가 너무 좁으면 뾰족해질 수 있고, 넓으면 완만해질 수 있음

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

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

가우시안 혼합

데이터가 가우시안 분포를 따른다고 가정하고, 평균 벡터, 공분산 행렬을 추정해서 사용함

여러개의 가우시안 분포를 사용해서 분리 가능함

주어진 데이터들(X)에서 가우시안 분포를 확인하기 위한 매개변수 세타를 추정하기 위해서 최대우도를 사용해서 공식화를 시킬 수 있음

인코딩

원래 공간을 다른 공간으로 변환하는 과정

데이터 압축 - 역변환으로 얻은 결과가 처음 결과와 최대한 비슷해야 함

디코딩

변환 공간을 원래 공간으로 변형

선형 인자 모델

행렬곱의 인코딩, 디코딩 과정 표현가능

각 인코딩, 디코딩은 가중치 w, 변수 x, 변수a를 가짐

인코딩에 a가 없으면 결과는 무조건적인 연관을 가지게 됨

z,a가 가우시안 분포(중앙에 많은 경우)를 따르면 pca를 따르게 됨

pca 최적화 문제

pca 기반 데이터 압축 z가 비 가우시안인 경우, ica를 따르게 됨

주성분 분석

데이터의 평균을 원점으로 이동시킴

$$\left.\begin{aligned} \mathbf{x}_i &= \mathbf{x}_i - \boldsymbol{\mu}, \qquad i = 1,2,\cdots,n \\ \text{이때 } \boldsymbol{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \end{aligned}\right\}$$

주성분의 차원 축소

각 데이터의 손실을 최소로 하는 축으로의 차원축소

분산이 클수록 정보 손실이 적음

아이겐 페이스

sparse error - 각 데이터에 대한 오류 상쇄 가능

블라인드 원음 분리

여러 소리가 적정한 비율로 섞여있는 경우 원 소리의 분리

음원 분리 혼합신호 x를 원래신호 z의 결합으로 표현

$$\left.\begin{aligned} x_1 &= a_{11}z_1 + a_{12}z_2 \\ x_2 &= a_{21}z_1 + a_{22}z_2 \end{aligned}\right\}$$

이때 a는 각 원음z가 어떤 비율로 섞인 건지에 대한 가중치 행렬을 나타냄 비율은 여러 종류가 나올 수 있으므로 추가 조건이 필요함

독립적 가정

원래 신호가 서로 독립이라고 가정함

$$P(\mathbf{z}) = P(z_1, z_2, \cdots, z_d) = \prod_{j=1}^{d} P(z_j)$$

ICA에서의 문제풀이

원래 신호의 비가우시안인 정도를 키우는 가중치로 구함

비 가우시안인 경우, 각각의 데이터는 각각의 기준을 가지고 표현되므로 구별하기가 편해짐

각 데이터의 분리가 확실하게 될 수 있는 그래프를 구할 수 있게 됨

$$kurtosis(\mathbb{z}_j) = \frac{1}{n}\sum_{i=1}^{n} z_{ji}^4 - 3\left(\frac{1}{n}\sum_{i=1}^{n} z_{ji}^2\right)^2$$

화이트닝

희소 코딩

	기저함수 또는 기저 벡터의 선형결합으로 신호 표현 푸리에 변환을 통해서 분리 또는 합성을 시도함 기저 함수는 데이터 별로 다름 $\hat{\mathbf{D}}, \hat{\mathbf{A}} = \underset{D,A}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\phi(\mathbf{a}_i)$

학과	철학과	학번	201802344	이름	유정훈
구분	내용				
학습범위	동영상 13주차				
학습내용	지도학습: 모든 훈련 샘플이 레이블 정보를 가짐 비지도 학습: 모든 훈련 샘플이 레이블 정보를 가지지않음 준지도 학습: 레이블을 가진 샘플과 가지지 않은 샘플이 섞여있음 <비지도 학습의 일반 과업> 1. 군집화: 유사한 샘플을 모아 같은 그룹으로 묶는 일 2. 밀도 추정: 데이터로부터 확률분포를 추정하는 일 3. 공간 변환: 원래 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일 <군집화> 1. K-평균 알고리즘: K 평균은 샘플의 평균으로 군집 중심을 갱신 -초기 군집 중심에 따라 결과가 달라지기에 여러 중심을 가지고 수행해봄으로써 좋은 결과를 낼 수 있음. 2. 친밀도 전파 알고리즘: 책임 행렬 R과 가용 행렬 A라는 두 종류의 친밀도 행렬을 이용하여 군집화 -군집 개수 k를 자동으로 알아냄 3. 밀도 추정 -커널 밀도 추정: 특정 공간을 칸의 집합으로 분할한 다음 칸에 있는 샘플의 빈도를 세어 추정 - 가우시안 혼합: 데이터가 가우시안 분포를 따른다고 가정하고 평균 벡터와 공분산 행렬로 추정 - EM 알고리즘: 가우시안으로 샘플의 소속 정보 개선				

	4. 공간변환 - 인코딩과 디코딩: 원래 공간을 다른 공간으로 변환하는 인코딩 과정, 변환 공간을 원래 공간으로 역변환하는 디코딩 과정 - 선형 인자 모델: 선형 연산을 이용한 공간 변환 기법 - 주성분 분석(선형): 데이터를 원점 중심으로 옮김 이후 더 낮은 차원으로 변환시키고 투영 -손실을 최소화, 저차원으로 변환 가능 - 독립 성분 분석(선형): 실제 세계에서는 섞이는 여러 신호를 분리하기 위해 사용 -비가우시안 가정: 원래 신호가 가우시안이라면 혼합 신호도 가우시안이 되므로 분리할 수 가 없음 -희소 코딩: 기저 함수 또는 기저 벡터의 선형 결합으로 신호를 표현 -비지도 학습이 데이터에 맞는 기저 벡터를 자동으로 알아냄 -

학과	컴퓨터 전자시스템 공학	학번	201702899	이름	이학빈
구분	내용				
학습 범위	13주차 강의				
학습 내용	-지도 학습 지도 학습은 입력 데이터와 해당 데이터에 대한 정답(레이블 또는 타깃)을 사용하여 모델을 훈련시키는 방법이다. 모델은 주어진 입력과 출력 사이의 관계를 학습하여 새로운 입력 데이터에 대한 출력 값을 예측하거나 분류를 수행한다. 주요 예시: 분류, 회귀 -비지도 학습 비지도 학습은 레이블이 없는 데이터를 기반으로 모델을 훈련시키는 방법이다. 모델은 데이터의 내재된 구조나 패턴을 발견하고 데이터를 클러스터링, 차원 축소, 이상치 탐지 등의 작업을 수행한다. 주요 예시: 클러스터링, 차원 축소, 생성 모델 -준지도 학습 준지도 학습은 지도 학습과 비지도 학습의 조합으로, 일부 데이터는 레이블이 주어지고 일부 데이터는 레이블이 없는 상황에서 학습을 수행한다. 이 방법은 레이				

	블이 부족한 상황에서도 모델을 효과적으로 학습시킬 수 있다. 주로 일부 데이터에만 레이블이 있을 때 이를 사용하여 레이블이 없는 데이터를 더 잘 이해하고 예측하는 데 활용된다. -군집화 군집화(Clustering)는 비지도 학습의 한 형태로, 유사한 특성을 가진 데이터 포인트들을 그룹화하는 작업을 말한다. 군집화는 데이터 내에서 숨겨진 패턴을 찾고, 데이터를 비슷한 그룹이나 클러스터로 묶어서 이해하기 쉽게 만든다. 군집화 알고리즘은 데이터를 클러스터로 나누는 방법을 결정하며, 각 클러스터는 유사성이 높은 데이터 포인트들의 집합이다. 주어진 데이터의 특성을 고려하여 데이터를 서로 다른 그룹으로 나누는 것이 목표이다. 일반적으로 군집화의 과정은 다음과 같다: 유사성 지표 정의: 군집화를 수행하기 전에 데이터 간의 유사성을 측정할 지표(거리, 유사도 등)를 정의한다. 군집화 알고리즘 적용: K-평균(K-Means), 계층적 군집화, DBSCAN 등과 같은 군집화 알고리즘을 선택하고 데이터에 적용한다. K-평균 알고리즘은 주어진 데이터를 K개의 클러스터로 그룹화하는데 사용된다. 클러스터 중심을 반복적으로 이동시키면서 각 데이터 포인트를 가장 가까운 클러스터에 할당한다. 계층적 군집화는 계층적인 트리 구조로 데이터를 그룹화하는 방법이며, 유사성에 기반하여 클러스터를 형성한다. DBSCAN은 밀도 기반 군집화 알고리즘으로, 데이터 밀도가 높은 지역을 클러스터로 그룹화한다. 클러스터 평가: 군집화가 완료된 후, 클러스터의 품질을 평가하고 필요에 따라 세분화하거나 수정한다. 내부 평가 지표(실루엣 스코어 등)나 외부 평가(사람이 정한 정답과의 비교 등)를 사용하여 클러스터링의 효과를 평가한다. 군집화는 데이터에서 숨겨진 구조를 발견하거나 데이터를 이해하는 데 유용하다. 예를 들어, 고객 세분화, 이미지 분할, 이상치 탐지 등 다양한 분야에서 활용될 수 있다. -k-평균과 친밀도 전파 알고리즘 K-평균(K-Means)과 친밀도 전파 알고리즘은 군집화 알고리즘 중 두 가지 다른 방식으로 데이터를 클러스터링 하는 기법이다.

	K-평균(K-Means) 알고리즘: K-평균 알고리즘은 주어진 데이터를 지정된 K개의 클러스터로 그룹화하는데 사용된다. 이 알고리즘은 다음과 같은 과정으로 동작한다: K개의 클러스터 중심 초기화: 사용자가 지정한 K값에 따라 클러스터 중심을 초기화한다. 할당 단계 (Assign Step): 각 데이터 포인트를 가장 가까운 클러스터 중심에 할당한다. 업데이트 단계 (Update Step): 할당된 클러스터에 속한 데이터 포인트들의 평균을 계산하여 새로운 클러스터 중심을 업데이트한다. 2, 3단계 반복: 할당과 업데이트 단계를 번갈아가며 반복하면서 클러스터 중심과 할당된 데이터 포인트를 조정한다. 수렴: 클러스터 중심의 변화가 거의 없을 때까지 알고리즘이 수렴하고 최종적으로 각 데이터 포인트가 속한 클러스터를 결정한다. K-평균은 클러스터의 수 K를 사전에 지정해야하며, 각 클러스터의 중심이 평균으로 업데이트되기 때문에 데이터가 원형 클러스터를 형성하는 경우에 잘 작동한다. 친밀도 전파 알고리즘 (Affinity Propagation Algorithm): 친밀도 전파 알고리즘은 데이터 내에서 데이터 포인트 간의 "친밀도"라는 개념을 기반으로 클러스터링을 수행하는 알고리즘입니다. 이 알고리즘은 다음과 같은 특징을 가진다: 모든 데이터 포인트 간의 친밀도를 계산하고, 각 데이터 포인트가 다른 데이터 포인트를 대표할 수 있는 "중심"으로 메시지를 보낸다. 이 중심을 기반으로 데이터 포인트는 자신을 대표하는 중심을 선택하게 되고, 데이터 포인트는 중심들 간의 전파된 메시지를 통해 클러스터링 된다. 알고리즘은 데이터 내에서 중심 포인트를 동적으로 선택하고, 클러스터의 수를 사전에 정하지 않아도 된다. 친밀도 전파 알고리즘은 네트워크 라우팅이나 이미지 분할과 같은 영역에서도 효과적으로 사용될 수 있다. 친밀도 전파 알고리즘은 클러스터의 수를 지정할 필요가 없고, 데이터 내에서 중심을 선택하므로 다양한 형태의 클러스터를 찾을 수 있지만, 데이터가 많거나 클러스터의 크기가 큰 경우 계산 비용이 증가할 수 있다. -커널 밀도 추정

	커널 밀도 추정(Kernel Density Estimation, KDE)은 주어진 데이터의 확률 분포를 추정하는 데 사용되는 비모수적인 방법 중 하나이다. 데이터의 분포를 그래프 상에서 부드럽게 나타내기 위해 사용된다. 커널 밀도 추정의 개념 데이터의 분포를 추정: 주어진 데이터 포인트들 로부터 어떤 확률 분포가 생성되었는지를 추정한다. 부드러운 확률 분포 추정: 데이터 포인트들의 위치에 따라 확률 밀도 함수를 표현한다. 이를 통해 데이터의 분포를 부드럽게(연속적으로) 나타낼 수 있다. 커널 함수를 이용한 추정: 각 데이터 포인트를 중심으로 하는 커널(일종의 함수)을 사용하여 데이터 주변의 확률 밀도를 추정한다. 커널은 보통 가우시안 함수와 같은 형태를 취하며, 데이터 포인트 주변에서 높은 확률을 부여한다. 모든 커널을 합산하여 전체 확률 분포 추정: 각 데이터 포인트에서 커널을 적용하고, 모든 커널들을 합산하여 전체적인 확률 분포를 구성한다. 커널 밀도 추정의 중요한 요소: 커널 함수의 선택: 커널 함수는 확률 밀도 추정에 사용되는 함수로, 가우시안 함수가 일반적으로 많이 사용된다. 하지만 커널의 종류에 따라 추정 결과가 달라질 수 있다. 밀도 추정 대역폭(Bandwidth): 커널 함수의 폭을 결정하는데, 밀도 추정 결과에 영향을 준다. 적절한 대역폭을 선택하는 것이 중요하다. 너무 작은 대역폭은 데이터를 과도하게 복잡하게 만들고, 너무 큰 대역폭은 데이터를 너무 단순하게 나타낼 수 있다. 커널 밀도 추정은 데이터의 분포를 파악하고 시각화하는 데 유용하다. 또한 확률 밀도를 추정하여 이상치 탐지, 군집화, 분류 등 다양한 데이터 분석 작업에 활용될 수 있다. -가우시안 혼합 모델 가우시안 혼합 모델(Gaussian Mixture Model, GMM)은 비지도 학습의 일종으로, 데이터를 여러 개의 가우시안 분포를 조합하여 모델링하는 확률적 생성 모델이다. 각 데이터 포인트가 여러 개의 가우시안 분포 중 어느 하나에서 생성되었을 가능성을 확률적으로 나타낸다. 가우시안 혼합 모델은 주어진 데이터가 여러 개의 가우시안 분포를 가진 군집으로 구성되어 있다고 가정한다. 각 군집은 다른 평균과 분산을 가진 가우시안 분

	포를 나타낸다. GMM은 이러한 가우시안 분포들의 조합으로 데이터를 모델링한다. 가우시안 혼합 모델은 크게 세 가지 요소로 구성된다 가우시안 분포들의 조합 여러 개의 가우시안 분포(또는 클러스터)를 조합하여 데이터를 모델링한다. 각각의 가우시안 분포는 데이터의 특정 부분을 대표한다. 모수(Parameter) 각 가우시안 분포마다 평균과 분산을 가지고 있다. 이러한 평균과 분산은 모델의 파라미터로, EM(Expectation-Maximization) 알고리즘 등을 사용하여 추정된다. EM 알고리즘 EM 알고리즘을 사용하여 모수를 추정한다. EM 알고리즘은 가우시안 혼합 모델에서 숨겨진(latent) 변수(각 데이터 포인트가 어떤 클러스터에 속하는지의 정보)를 활용하여 모델을 학습한다. 가우시안 혼합 모델은 주어진 데이터를 여러 개의 가우시안 분포를 통해 모델링하므로, 이 모델을 사용하여 데이터의 군집화(clustering)를 수행할 수 있다. 또한 새로운 데이터가 주어졌을 때 해당 데이터가 각 가우시안 분포에서 발생할 확률을 계산하여 이상치 탐지나 데이터 포인트의 소속 여부를 추론하는 데 사용될 수 있다.