Spring 2023

SWCON253: Machine Learning

Lecture 16
Density Estimation

Jinwoo Choi
Assistant Professor
CSE, Kyung Hee University

Disclaimer: Many slides have been borrowed from Hui Yong Kim, who may have borrowed some of them from others. Any time a slide does not show any explicit credit, I credit the slide to Hui Yong Kim.

Contents

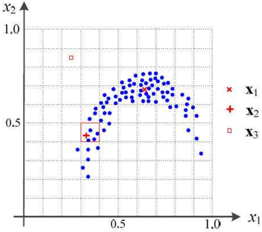
References

• "기계학습" by 오일석, "패턴인식" by 오일석

1. Density Estimation

Density Estimation

- **밀도 추정 (Density Estimation)**
  - 어떤 점 $\mathbf{x}$에서 데이터가 발생할 확률, 즉 확률분포 $P(\mathbf{x})$를 구하는 문제
    - ★ 예를 들어, 그림에서 $P(\mathbf{x}_1) > P(\mathbf{x}_2) > P(\mathbf{x}_3)$
  - 분류 모델이자 생성 모델 (Generative Model)
    - ★ 새로운 샘플을 생성할 수 있음



- **Parametric Methods**
  - ML (Maximum Likelihood) Estimation
  - MAP (Maximum A Posteriori) Estimation
- **Non-Parametric Methods**
  - 파젠 창
  - k-최근접 이웃 추정
- **Mixture Models**
  - Gaussian Mixture Model (GMM)

2. Parametric Methods

1. Parametric Methods
2. ML Estimation
3. MAP Estimation
4. ML vs. MAP

Parametric Methods

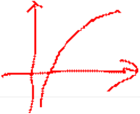
ML Estimation

$$\hat{\Theta} = \arg\max_{\Theta} P(\mathbb{X}|\Theta)$$

$$P(\mathbb{X}|\Theta) = P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n|\Theta) = \prod_{i=1}^{n} P(\mathbf{x}_i|\Theta)$$

$$\boxed{\mathcal{L}(\Theta) = \log P(\mathbb{X}|\Theta) = \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta)}$$

$$\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\Theta)$$



log x

ML Estimation (cont'd)

$$p(\mathbf{x}_i \mid \Theta) = p(\mathbf{x}_i \mid \boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right)$$

$$\ln p(\mathbf{x}_i \mid \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma|$$

$$L(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - N\left(\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma|\right)$$

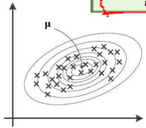
$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N} \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = 0$$

$$\Rightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

ML estimator for a Gaussian density

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

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



[Caution] Here, we assumed:

- Each samples (x_i's) are *independent* to each other.
- Each features (x_k's) are *not independent* to each other.

MAP Estimation

$$\hat{\Theta} = \arg\max_{\Theta} P(\Theta|\mathbb{X}) = \arg\max_{\Theta} P(\mathbb{X}|\Theta)P(\Theta)$$

$P(\Theta | X)$

ML: $\hat{\Theta} = \arg\max_{\Theta} P(X|\Theta)$

$$P(\mathbb{X}|\Theta) = P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n|\Theta) = \prod_{i=1}^{n} P(\mathbf{x}_i|\Theta)$$

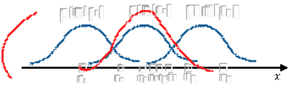
$$\log P(\mathbb{X}|\Theta)P(\Theta) = \log P(\mathbb{X}|\Theta) + \log P(\Theta) = \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta) + \log P(\Theta)$$

ML: $\sum_{i=1}^{N} \log P(x_i|\theta)$

ML vs. MAP

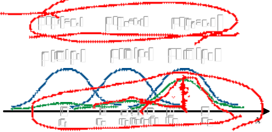
- ML Θ

p(x|M₂)



$$\log P(\mathbb{X}|\mu) = \sum_{i=1}^{n} \log P(x_i|\mu)$$

- MAP Θ



$$\log(P(\mathbb{X}|\mu)P(\mu)) = \sum_{i=1}^{n} \log P(x_i|\mu) + \log P(\mu)$$

3. Non-Parametric Methods

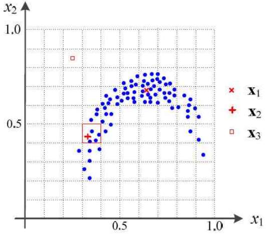
1. Parametric vs. Non-parametric
2. Histogram Method
3. Parzen Window
4. *k*-Nearest Neighbors Estimation
5. *k*-Nearest Neighbors Classification

Parametric vs. Non-Parametric

- **Parametric Density Estimation**
  - 확률 분포가 유한개의 파라메터(모수)로 표현되는 형태 (예: Gaussian, Exponential, Bernoulli, ...)
  - 매개변수 추정: ML, MAP 방법 등

- **Non-Parametric Density Estimation**
  - 확률 분포가 임의의 형태
    - ★ 유한개의 파라메터를 정할 수 없음
  - 파젠 창, k-최근접 이웃 추정 방법 등

K-NN



Histogram Method

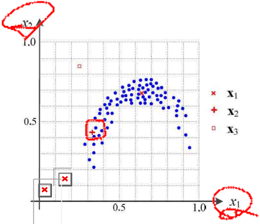
◆ 특징 공간을 칸의 집합으로 분할한 다음, 칸에 있는 샘플의 빈도를 세어 아래 식으로 추정

$$P(\mathbf{x}) = \frac{bin(\mathbf{x})}{n}$$

★ *bin*(**x**): **x**가 속한 칸(bin)의 샘플 개수

★ *n*: 전체 샘플의 개수

★ 예) 우측 그림에서, $P(\mathbf{x}=+)=4/80=0.05$

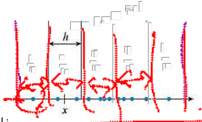


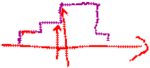
4/80 = 0.05

● 문제점

★ 매끄럽지 못하고 계단 모양을 띠는 확률밀도함수가 됨

★ 칸의 크기와 위치에 민감함





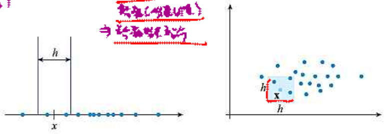
Parzen Window

Histogram 방법을 확장해 보자

- 특징공간에서 임의의 위치 $\mathbf{x}$에서의 확률 값을 다음 식과 같이 구함

$$\boxed{P_h(\mathbf{x}) = \frac{1}{h^d}\frac{k(\mathbf{x})}{n}}$$

  - ★ h: 창의 크기
  - ★ $k(\mathbf{x})$: $\mathbf{x}$를 중심으로 한 창에 포함된 샘플의 개수
  - ★ n: 전체 샘플의 개수
  - ★ d: $\mathbf{x}$의 차원 (feature의 개수)



(a) 1 차원 특징 공간 (b) 2 차원 특징 공간

- 문제점
  - ★ 여전히 매끄럽지 못함

- 해결책
  - ★ 창 안의 샘플 분포를 나타내는 가중치 함수(커널)를 도입 (중앙에 가까울 수록 높은 가중치)
  - ★ 부드러운 커널을 사용하면 매끄러운 확률 분포를 얻을 수 있음

Parzen Window (cont'd)

- **Kernel Density Estimation (KDE)**
  - 특징공간에서 샘플의 위치 $\mathbf{x}$에 커널을 씌우고, 커널 안에 있는 샘플들의 가중합을 구함
  - 부드러운 커널을 사용하면 매끄러운 확률 분포를 얻을 수 있음

$$\boxed{P_h(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{x}_i)}$$

★ $K_h(\mathbf{x}-\mathbf{x}_i)$:
$\mathbf{x}_i$를 중심으로 하는 폭 h의 커널

$K_h(x-x_i)$

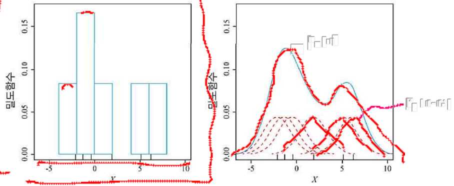


그림 6-10 히스토그램 방법(왼쪽)과 커널 밀도 추정법(오른쪽)의 비교

Parzen Window (cont'd)

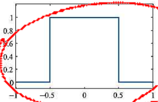
Kernel Density Estimation (KDE) (cont'd)

- K_h는 표준 커널 K로부터 만들어 낼 수 있음

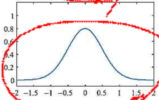
$$\boxed{K_h(\mathbf{x}) = \frac{1}{h^d} K\left(\frac{\mathbf{x}}{h}\right)}$$

$K_h(x - x_i)$

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



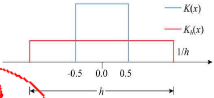
(a) 계단 함수



(b) 가우시안 함수

$K(x) \longrightarrow K\left(\frac{x}{h}\right)$

$K\left(\frac{x - x_i}{h}\right)$



Parzen Window (cont'd)

$\frac{1}{N}\sum_{i=1}^{N} N(x - x_i)$

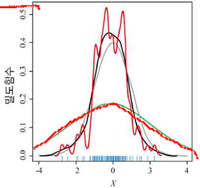


그림 6-11 대역폭이 확률밀도함수 추정에 미치는 영향

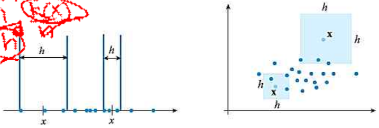
$n \ll d$

k-Nearest Neighbors *Estimation*

$$P_k(\mathbf{x}) = \frac{1}{h(\mathbf{x})^d}\frac{k}{n}$$

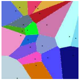
cf) Parzen window

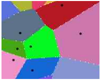
$P(x) = \frac{1}{h} \frac{K(x)}{n}$



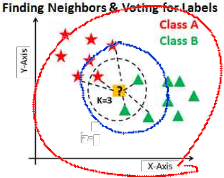
(a) 1 차원 특징 공간 (b) 2 차원 특징 공간

그림 3.8 k-최근접 이웃 방법으로 확률 밀도 함수의 추정 (k=3일 때)





k-Nearest Neighbors *Classification*



$$\hat{j} = \arg\max_{1 \le j \le c} P(\omega_j|\mathbf{x}) = \arg\max_{1 \le j \le c} k_j$$

[Algorithm] *k*-NN Classification

1) 훈련 샘플들 중에 **x**와 가장 가까운 *k*개를 찾는다.
2) 찾은 *k*개 중 가장 빈도가 높은 class를 찾는다.
3) 여러 *k*에 대해 1)과 2)를 반복한 후, 가장 많은 vote를 얻은 class를 정답으로 한다.

Contents

References

- “기계학습” by 오일석, "패턴인식" by 오일석

4. Gaussian Mixture Model

1. Mixture Models
2. Gaussian Mixture Model (GMM)

Mixture Models

Density Estimation with Mixture Models

- 두 개 이상의 서로 다른 확률 분포의 혼합으로 데이터의 확률 분포를 모델링
- 요소 확률 분포로는 통상 Gaussian 분포를 사용 → Gaussian Mixture Model (GMM)

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k P(\mathbf{x} \mid \boldsymbol{\theta}_k)$$

$$\sum_{k=1}^{K} \pi_k = 1$$

- **Example:** ***Single vs Mixture Gaussian***

$$P(\mathbf{x}) = N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

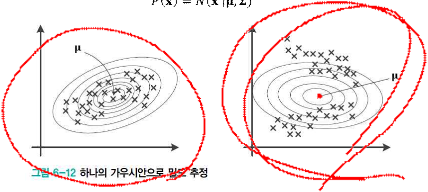


그림 6-12 하나의 가우시안으로 밀도 추정

선형결합

$$P(\mathbf{x}) = \pi_1 N(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$(\pi_1 + \pi_2 = 1)$$

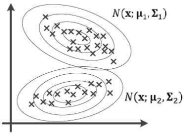


그림 6-13 가우시안 혼합으로 밀도 추정

Gaussian Mixture Model

Most Popular !

Gaussian Mixture Model

- K개 Gaussian들의 선형 결합으로 확률 분포를 표현

$$\boxed{P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \qquad \boxed{\textstyle\sum_{k=1}^{K} \pi_k = 1}$$

quadratic Form

★ Gaussian의 개수 K는 Hyper parameter

다변수 가우시안

★ k^{th} Gaussian 분포: $N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \dfrac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}_k|}} \exp\left(-\dfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right)$

★ 추정할 파라메터: $\boldsymbol{\Theta} = \{(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \dots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$

- π_k: k^{th} Gaussian에 대한 혼합계수(혼합가중치)(사전확률)
- $\boldsymbol{\mu}_k$: k^{th} Gaussian의 평균벡터
- $\boldsymbol{\Sigma}_k$: k^{th} Gaussian의 공분산행렬

이 가우시안 분포에서 데이터가 발생할 확률

Gaussian Mixture Model (cont'd)

- Estimation of $\mathbf{\Theta}$
  - $\boxed{\mathbf{\Theta} = \{(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}}$
  - *Single Gaussian*의 ML 추정 결과(지난 강의 참조):

$$P(\mathbf{x}) = N(\mathbf{x}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\Theta} \log P(\mathbb{X}|\boldsymbol{\Theta}) \rightarrow$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

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

  - *Gaussian Mixture*의 경우:

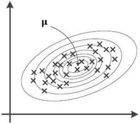
$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

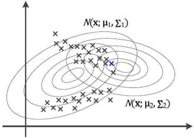
★ K쌍의 $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$들과 π_k들(파라메터)을 추정해야 함

★ 샘플이 어떤 Gaussian에 속하는 지(소속정보)를 모름

→ 소속정보 추정(Expectation)과 파라메터 추정(Maximization)을 번갈아 반복해 보자! (EM 알고리즘)

E 반복 M





5. EM Algorithm

1. EM Algorithm
2. EM Algorithm for GMM

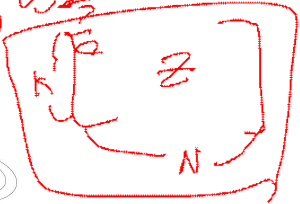
EM Algorithm

◆ 소속정보와 파라메터를 번갈아 추정해 보자

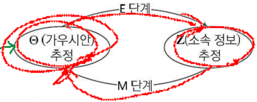
Note. **K-Means Clustering** is also a kind of EM Algorithm

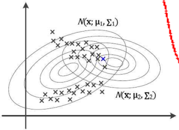
1) [파라메터 초기화] 파라메터(Θ)를 모르므로 통상 난수로 초기화
   - ★ GMM의 경우: $\Theta = \{(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$
2) [E단계] 각 샘플의 소속정보(Z)를 추정
   - ★ $Z = [z_{ki}]$, z_{ki}: 샘플 $\mathbf{x}_i$가 k번째 분포에 소속될 확률
3) [M단계] 각 분포에 소속된 샘플들을 이용하여 파라메터(Θ) 갱신
   - ★ 예: $\widehat{\boldsymbol{\theta}}_{ML} = \arg\max_{\Theta} \log P(\mathbb{X}|\boldsymbol{\Theta})$
4) Stop Condition을 만족할 때까지 2)와 3)을 반복

반복



초기값

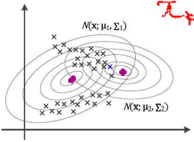


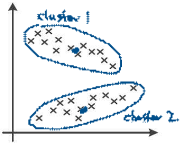


EM Algorithm for GMM

◆ 파라메터(Θ) 초기화

- 난수로 초기화
- *K*-Means Clustering 결과를 이용하여 초기화
  - ★ k^{th} Cluster에 속한 샘플들의 평균벡터(군집중심)와 공분산행렬을 구한다.
  - ★ k^{th} Gaussian의 평균벡터와 ~~공분산행렬~~을 위에서 구한 값으로 초기화한다.
  - ★ k^{th} Gaussian의 혼합계수는 (k^{th} Cluster에 속한 샘플수)/(전체 샘플수)로 초기화 한다.





EM Algorithm for GMM (cont'd) 무조건 나옴

◆ [E단계] 소속정보(Z)의 추정

- **잠재변수** z_k가 어떤 샘플이 k^{th} Gaussian에 속하는지($z_k = 1$)아닌지($z_k = 0$)를 나타낸다고 하면,
  - ★ k^{th} Gaussian의 ***priori*** probability: $\pi_k = P(z_k = 1)$ 사전확률
  - ★ k^{th} Gaussian의 ***likelihood***: 우도 $P(\mathbf{x}|z_k = 1) = N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ → PDF
- **소속정보** 행렬 $\mathbf{Z} = [z_{ki}]$의 원소 z_{ki}를 주어진 샘플 $\mathbf{x}_i$가 k^{th} Gaussian에 속할 확률로 정의하자.
  - ★ k^{th} Gaussian의 ***posteriori*** probability: $z_{ki} = P(z_k = 1|\mathbf{x}_i)$
  - ★ 베이즈 정리를 이용하여 정리하면,

대입

$$z_{ki} = P(z_k = 1|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|z_k = 1)P(z_k = 1)}{P(\mathbf{x}_i)} = \frac{P(\mathbf{x}_i|z_k = 1)P(z_k = 1)}{\sum_{j=1}^{K} P(\mathbf{x}_i|z_j = 1)P(z_j = 1)}$$

$$\Rightarrow \boxed{z_{ki} = \frac{\pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}$$

이걸 하고 나면 소속 정보 추정이 완료됨

Bayes' Theorem

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{K} P(B|A_j)P(A_j)}$$

EM Algorithm for GMM (cont'd)

- [M단계] 확률분포(Θ)의 추정
  - $\Theta = \{(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)\}$
    - ★ π_k: k^{th} Gaussian의 발생 확률(혼합 가중치)
    - ★ $\boldsymbol{\mu}_k$: k^{th} Gaussian의 *mean vector*
    - ★ $\boldsymbol{\Sigma}_k$: k^{th} Gaussian의 *covariance matrix*
  - ***ML Estimation*** of Θ:
    - ★ 각 샘플 $\mathbf{x}_i$의 발생 확률이 서로 독립이라 가정하면, (합리적인 가정)

      The log-likelihood of Θ: $\mathcal{L}(\Theta) = \log P(\mathbb{X}|\Theta) = \log P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n|\Theta)$

      $= \log \prod_{i=1}^{n} P(\mathbf{x}_i|\Theta) = \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta)$
    - ★ The GMM: $P(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}\,|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

⇒ $$\mathcal{L}(\Theta) = \log P(\mathbb{X}|\Theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

↓ Maximize

↳ numerical issue

0보다 작은 수를 계속 곱하면
컴퓨터에선 0이 될 수 있음

EM Algorithm for GMM (cont'd)

- (...계속) ***ML Estimation*** of Θ

  Now the problem can be summarized as a ***Constrained Optimization***:

  Maximize: $\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$

  Subject to: $\sum_{k=1}^{K} \pi_k - 1 = 0$

  ★ **Lagrangian:** $L(\Theta, \lambda) = \mathcal{L}(\Theta) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$

  ★ Differentiating $L(\Theta, \lambda)$ w.r.t. $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, π_k and equating the results to $\mathbf{0}$ yields:

$$\frac{\partial L(\Theta, \lambda)}{\partial \boldsymbol{\mu}_k} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ki} \mathbf{x}_i$$

← k번째 분포에 (확률가중치로) 속한 샘플의 평균

$$\frac{\partial L(\Theta, \lambda)}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ki} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

← k번째 분포에 (확률가중치로) 속한 샘플의 공분산

$$\frac{\partial L(\Theta, \lambda)}{\partial \pi_k} = 0 \quad \Rightarrow \quad \pi_k = \frac{n_k}{n}$$

← k번째 분포가 선택되는 (확률값) 비율

$$\text{where} \quad n_k = \sum_{i=1}^{n} z_{ki}$$

← k번째 분포에 속한 샘플의 (확률합계) 개수

ML estimator for a Gaussian density

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

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

6. Summary & Remarks

1. **Summary of Density Estimation**
2. **Remarks: Density Estimation의 활용**

$$L(\theta,\lambda)=\sum_{i=1}^{n}\log\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)+\lambda\left(\sum_{j=1}^{k}\pi_j-1\right)$$

$$\frac{\partial L(\theta,\lambda)}{\partial\mu_k}=\sum_{i=1}\frac{1}{\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)}\cdot\frac{\partial}{\partial\mu_k}\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)$$

$$\Rightarrow\frac{\partial}{\partial\pi_k}\pi_k N(x_i|\mu_k,\Sigma_k)$$

$$\Sigma_k^{-1}(x_i-\mu_k)$$

$$\uparrow$$

$$=\sum_{i=1}^{n}\frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)}\frac{\partial}{\partial\mu_k}\frac{(x_i-\mu_k)^2}{2\Sigma_k}=\sum_{i=1}^{n}p(z_k=1|x_i)\frac{\partial}{\partial\mu_k}\frac{(x_i-\mu_k)^2}{2\Sigma_k}$$

Summary of Density Estimation

Non-Parametric

$$P(\mathbf{x}) = \frac{bin(\mathbf{x})}{n}$$

Histogram

$$P_k(\mathbf{x}) = \frac{1}{h^d}\frac{k(\mathbf{x})}{n}$$

→ + window

부드럽지 않음

+ Kernel ex) Gaussian

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

$$P_k(\mathbf{x}) = \frac{1}{h(\mathbf{x})^d}\frac{k}{n}$$

Parametric

$$P(\mathbf{x}) = P(\mathbf{x}|\boldsymbol{\theta})$$

Parametric Mixture GMM

$$P(\mathbf{x}) = P(\mathbf{x}|\boldsymbol{\theta})$$
$$= \sum_{k=1}^{K} \pi_k P(\mathbf{x}\,|\boldsymbol{\theta}_k)$$

$$\boldsymbol{\theta} = \{(\pi_1, \boldsymbol{\theta}_1), \ldots, (\pi_K, \boldsymbol{\theta}_K)\}$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Parameter Estimation

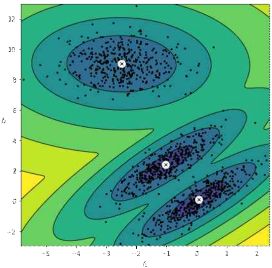
$$\widehat{\boldsymbol{\theta}}_{ML} = \arg\max_{\Theta} \log P(\mathbb{X}|\boldsymbol{\theta})$$

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\Theta} \log P(\mathbb{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

$$P(\mathbb{X}|\boldsymbol{\theta}) = P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n|\boldsymbol{\theta})$$
$$= \prod_{i=1}^{n} P(\mathbf{x}_i|\boldsymbol{\theta})$$

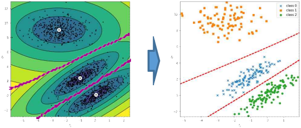
Density Estimation의 활용

◆ **After Density Estimation**

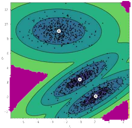


$P(x) = \sum \pi_k \cdot p(x | \theta_k)$

● Clustering (Classification) $\pi_i p(x|\theta_i) = \pi_j p(x|\theta_j)$



● Outlier (Anomaly) Detection



$p(x) < \text{Threshold}$

Q & A

- **Q:** GMM과 *(Gaussian) Kernel Density Estimation*의 차이점은?

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian = 군집의 개수 K

Gaussian = 샘플의 개수 n

$$P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{x}_i) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = \frac{1}{nh^d}\sum_{i=1}^{n} N\left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \,\middle|\, 0, 1\right)$$