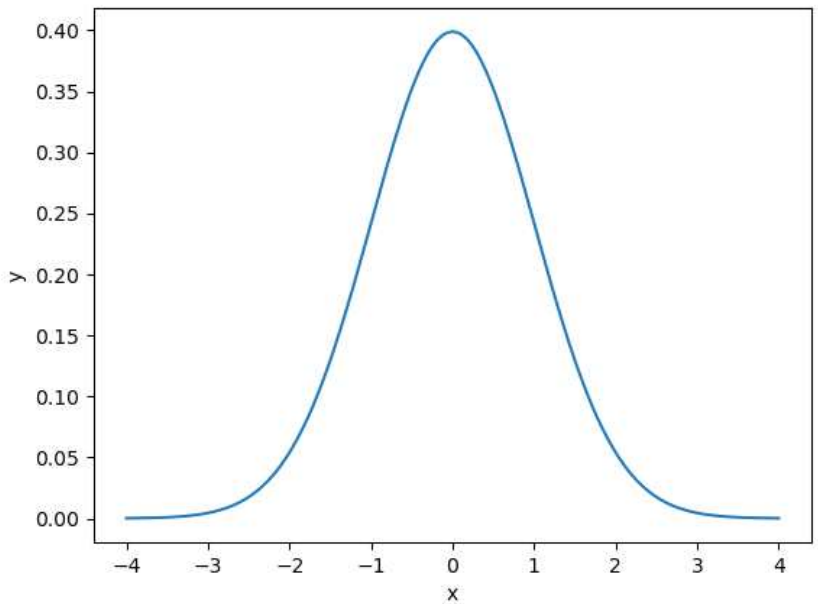
가우시안 혼합 모델 (Gaussian mixture model, GMM)

아이디어: 샘플이 파라미터(평균 μ, 분산 Σ)가 알려져 있지 않은 여러 개의 혼합된 가우시안 분포(= 정규 분포)에서 생성되었다고 가정

가우시안 분포

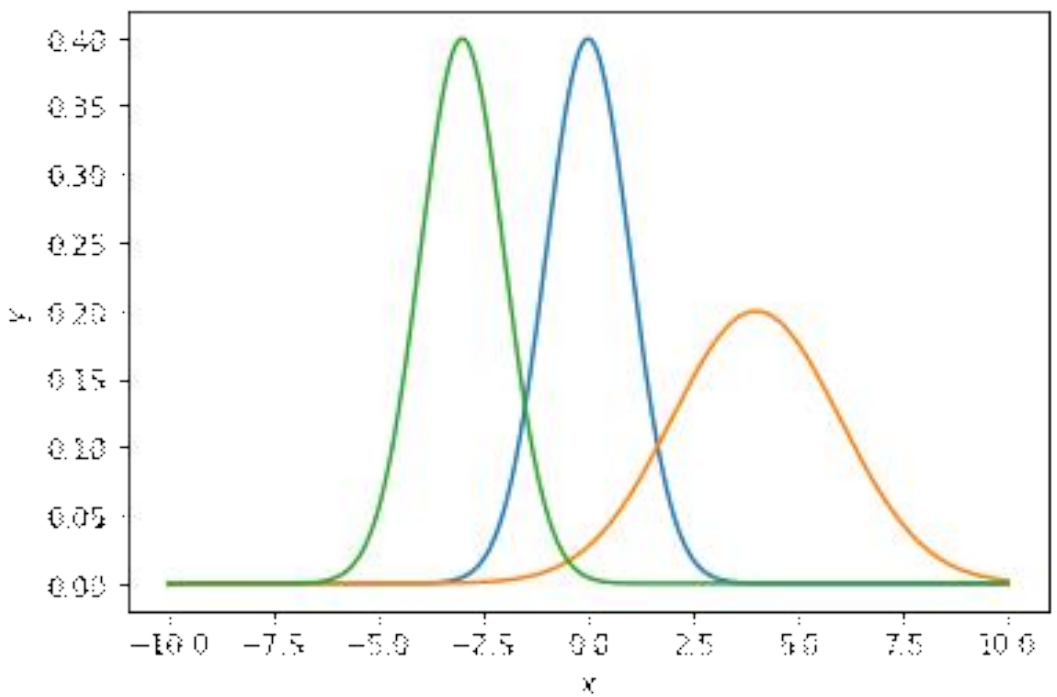
확률 밀도 함수 (Probability density function, PDF):

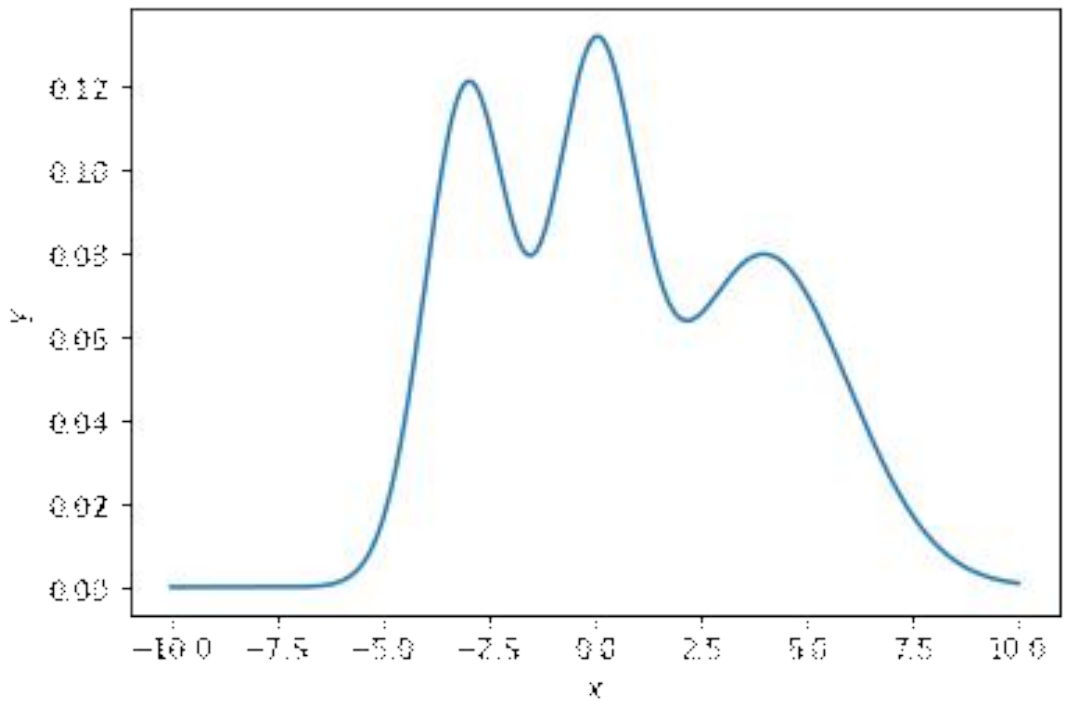
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



가우시안 혼합 모델 예시(1차원)

$N(-3, 1^2)$, $N(0, 1^2)$, $N(4, 2^2)$ 모델을 각각 0.3, 0.3, 0.4의 가중치로 혼합





각 가우시안 분포는 클러스터를 상징

각 클러스터의 가중치를 ϕ라고 한다면,
j 개의 클러스터의 가우시안 혼합 모델

$$\phi_1 \bullet N(\mu_1, \Sigma_1) + \phi_2 \bullet N(\mu_2, \Sigma_2) + \cdots + \phi_j \bullet N(\mu_j, \Sigma_j)$$

(※ 본래는 확률변수를 분포에 연결하고 확률변수끼리 가중치 평균을 연산해야 합니다. 이해의 편의를 위해)

샘플들이 이미 주어졌을 때
“이 샘플들은 가우시안 혼합 모델에서 나온 것이야” 가정
알고 싶은 것: μ, Σ, ϕ

찾는 방법: 기댓값-최대화(expectation-maximization, EM) 알고리즘

i번째 샘플 x_i가 k번째 클러스터에 속할 확률

$$p(z_{ik}=1|x_i) = \frac{p(z_{ik}=1)p(x_i|z_{ik}=1)}{\sum_{j=1}^{K} p(z_{ij}=1)p(x_i|z_{ij}=1)} = \frac{\phi_k f(x_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\phi_j f(x_i;\mu_j,\Sigma_j)}$$

z_{ik}: i번째 샘플이 k번째 클러스터에 속하면 1, 그 외 0

확률을 K개 클러스터에 대해 모두 구한 후 가장 높은 확률의 클러스터에 속하게 된다. (sklearn.mixture.GaussianMixture 클래스의 predict_proba 매서드)

분포 모양의 제약조건 (Σ)

2차원에서

공분산 행렬 $\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

a, d: 해당 특성 값들의 분산 / b, c: 각 특성들 간의 공분산

sklearn.mixture.GaussianMixture의 covariance_type 매개변수

spherical: 모든 클러스터가 원형, b = c = 0 / a = d

diag: 타원형, 타원의 축은 항상 좌표축과 평행, b = c = 0

tied: 모든 클러스터가 동일한 타원 모양, 크기, 방향, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$

이상치 탐지

각 샘플을 가우시안 혼합 모델의 PDF에 대입해 나온 결과에 대해
하위 n%를 걸러냄

클러스터 변수 선택하기

BIC = $\log(m)p - 2\log(\hat{L})$

AIC = $2p - 2\log(\hat{L})$

m: 샘플 개수, p: 파라미터 개수, L: 가능도

작을수록 좋음 / 파라미터 개수에 대해 규제: 과대적합 방지

샘플 개수 m은 웬만하면 e^2 = 7.38 ≒ 8 보다 큼 / BIC가 AIC보다 p에 대한 규제가 더 큼

가능도(Likelihood)

샘플이 이미 주어졌을 때 파라미터의 변화에 따른 PDF값의 곱을 비교

$$L(\mu, \Sigma; x) = \prod_{i=1}^{m} f(x_i; \mu, \Sigma)$$

클수록 좋음

$L(\mu, \Sigma; x)$를 최대화하는 μ, Σ 찾기 $\rightarrow$ maximum likelihood estimation, MLE

베이즈 통계학

기존의 확률 인식: 사건의 발생 빈도(frequency)

베이즈 확률론의 확률 인식: 믿음(Belief)의 정도

"내일 비가 올 확률이 40%래"

"90%의 확률로 이 후보가 당선 될 것이라고 믿어"

베이즈 정리

사전의 나의 믿음, 사전 확률($P(H)$)이 데이터의 획득으로 어떻게 사후 확률($P(H|D)$)로 업데이트되는지에 대한 정리

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

베이즈 정리 예시

어떤 병은 전 세계적으로 1%의 유병률을 가지고 있다.: $P(\text{병}) = 0.01$

(나도 1%의 확률로 이 병에 걸렸을 것이라는 믿음)

정확도가 90%인 검사기로 검사를 했더니 양성판정을 받았다.: $P(\text{양성}|\text{병}) = 0.9$

내가 이 병에 걸렸을 확률은?: $P(\text{병}|\text{양성}) = ?$

(음성판정 정확도 $P(\text{음성}|\text{무병})$도 90%라고 가정)

$$P(\text{병}|\text{양성}) = \frac{P(\text{양성}|\text{병})P(\text{병})}{P(\text{양성})} = \frac{P(\text{양성}|\text{병})P(\text{병})}{P(\text{양성}|\text{병})P(\text{병}) + P(\text{양성}|\text{무병})P(\text{무병})}$$

$$= \frac{0.9 \times 0.01}{0.9 \times 0.01 + (1-0.9) \times (1-0.01)} \approx 8.3\%$$

참고 자료

- 오헬리앙 제롱, *핸즈온 머신러닝 2판* (한빛미디어, 2020), 328-339

- Teayoung Park, *Lecture Note: Mathematical Statistics(1)* (2020), 84-86

- “[스탠코리아 StanKorea] 베이즈 통계학 소개 Introduction to Bayesian Statistics | 베이즈 정리 & 베이즈 추론 | 베이지안이 되어야 할 이유.” *YouTube*. 2020년 6월 3일 수정, 2021년 7월 14일 접속, https://youtu.be/ELSxxe6gMaQ.

- “베이즈 정리.” *나무위키*. 2021년 7월 17일 수정, 2021년 7월 17일 접속, https://namu.wiki/w/베이즈%20정리.