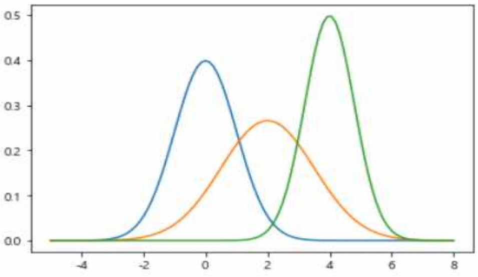
GMM(가우시안 혼합 모델)

<개요>

- 샘플이 파라미터가 알려지지 않은 여러 개의 혼합된 가우시안 분포에서 생성되었다고 가정하는 확률 모델
- 각 클러스터는 각각의 정규분포
- 각 샘플이 각 클러스터(각 파라미터를 가지는 정규분포)에서 생성될 확률을 구해서 가장 확률이 높은 클러스터에 할당



▶ GMM 변수

x_i : i번째 샘플

π_k : k번째 클러스터가 선택될 확률

z_{ik} : i번째 샘플을 k번째 클러스터가 생성하였는지에 대한 여부(0과 1의 이진변수)

u_k : k번째 클러스터의 평균

Σ_k : k번째 클러스터의 분산

▶ 각 샘플이 발생할 확률

- $P(x_i) = \sum_{j=1}^{k} \pi_j N(x_i | \mu_j, \Sigma_j)$

  $\sum_{j=1}^{k} \pi_j = 1$

  $0 \leqq \pi_j \leqq 1$
- 각 모든 클러스터에서 i번째 데이터가 생성될 확률의 합
- 각 클러스터가 전체망라, 상호배타이기 때문에 해당 식이 성립

=> GMM을 학습 : 적절한 π, μ, Σ를 찾는 것

<GMM 학습>

- EM 알고리즘을 적용하여 학습

▶ E-step

- $\gamma(z_{ik}) = P(z_{ik} = 1 | x_i)$ 를 바탕으로 함.

- 어떠한 데이터가 있는데, 이 데이터가 k번째 클러스터에서 생성되었을 확률을 구함

  -> 해당 확률을 가장 크게 만드는 클러스터가 해당 데이터의 클러스터

$\gamma(z_{ik}) = P(z_{ik}=1|x_i)$
$= \dfrac{P(x_i|z_{ik})P(z_{ik}=1)}{\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)}$
$= \dfrac{\pi_k N(x_i|u_k,\Sigma_k)}{\sum_{j=1}^{k}\pi_j N(x_i|\mu_j,\Sigma_j)}$

▶ M-step

- $\text{arg}max_\theta L(X;\theta) => \text{arg}max_\theta \ln L(X;\theta)$를 바탕으로 함

  - 각 클러스터를 구성하는 parameter가 존재할 때, 해당 데이터들이 나올 확률을 가장 크게 만드는 파라미터를 찾는 것(log가 증가함수이기 때문에 같은 의미)

- $\ln L(X;\theta) = \ln P(X|\pi,\mu,\Sigma) = \ln \Pi_{i=1}^{n} P(x_i|\pi,\mu,\Sigma) = \Sigma_{i=1}^{n} \ln P(x_i|\pi,\mu,\Sigma) = \Sigma_i^n \ln(\Sigma_{j=1}^{k}\pi_k N(x_i|\mu_k.\Sigma_k))$

-> 목적식 : $\text{arg}max_{\pi,\Sigma,\mu}\Sigma_i^n \ln(\Sigma_{j=1}^{k}\pi_k N(x_i|\mu_k.\Sigma_k))$

-> 이 때, 목적식이 μ_k,Σ_k에 대한 음의 이차식이기 때문에 미분하여 0을 만족시키는 값을 최적값으로 사용

- $\mu_k = \dfrac{1}{N_k}\Sigma_{i=1}^{n}\gamma(z_{ik})x_i,\ \ N_k = \Sigma_{i=1}^{n}\gamma(z_{ik}),\ \ \Sigma_k = \dfrac{1}{N}\Sigma_{i=1}^{n}\gamma(z_{ik})(x_i-\mu_k)(x_i-\mu_k)^T$

-> π_k의 경우 제약식($\Sigma_j^k\pi_j = 1$)이 존재하기 때문에 라그랑지안 승수 방식 사용

- $\Sigma_i^n \ln(\Sigma_{j=1}^{k}\pi_k N(x_i|\mu_k.\Sigma_k)) + \lambda(1-\Sigma_j^k\pi_k)$

-> $\lambda = N,\ \ \pi_k = \dfrac{1}{N}\Sigma_{i=1}^{n}\gamma(Z_{ik})$

<GMM 총정리>

- 초기화단계 : μ_k,Σ_k,π_k를 적당한 값으로 초기화
- E단계 : $\gamma(z_{ik})$를 계산
- M단계 : $\gamma(z_{ik})$를 이용하여 파라미터 업데이트
- 이를 반복

참고 : https://untitledtblog.tistory.com/133, PRML