Day 2. GMM : Gaussian Mixture Model

A. Multivariate Gaussian Distribution

1. Vector to normal distribution.

$x \in \mathbb{R}^d$, $N(x; \mu, \Sigma)$

* Gaussian Distribution func of x

$$N(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\}}$$

inverse

$|\Sigma| \rightarrow \det(\Sigma)$

x : n-dim vector

μ : mean vector

Σ : covariance matrix

* Covariance matrix & variable relationship

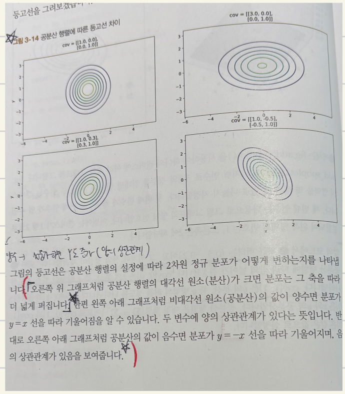
등고선을 그려보겠습니다

그림 3-14 공분산 행렬에 따른 등고선 차이

양수 → x 증가하면 y도 증가 (양의 상관관계)

그림의 등고선은 공분산 행렬의 설정에 따라 2차원 정규 분포가 어떻게 변하는지를 나타냅니다. 오른쪽 위 그래프처럼 공분산 행렬의 대각선 원소(분산)가 크면 분포는 그 축을 따라 더 넓게 퍼집니다. 한편 왼쪽 아래 그래프처럼 비대각선 원소(공분산)의 값이 양수면 분포가 $y=x$ 선을 따라 기울어짐을 알 수 있습니다. 두 변수에 양의 상관관계가 있다는 뜻입니다. 반대로 오른쪽 아래 그래프처럼 공분산의 값이 음수면 분포가 $y=-x$ 선을 따라 기울어지며, 음의 상관관계가 있음을 보여줍니다.

2. MLE (Modeling as multivariate Gaussian Distribution)

$$D = \{x^{(1)}, x^{(2)} \dots x^{(N)}\}$$

• Gaussian Modeling을 했고, 매개변수가 μ, Σ 일때 샘플 D를 얻을 수 있는 확률밀도 : 가능도

$$p(D; \mu, \Sigma) = N(x^{(1)}; \mu, \Sigma) N(x^{(2)}; \mu, \Sigma) \cdots N(x^{(N)}; \mu, \Sigma) \quad (\text{iid} \rightarrow \text{multiplication})$$

$$= \prod_{n=1}^{N} N(x^{(n)}; \mu, \Sigma)$$

$$= L(\mu, \Sigma)$$

- log 씌우기

$$\therefore L(\mu, \Sigma) = \log p(D : \mu, \Sigma)$$

- Maximize?

$$\frac{\partial L}{\partial \mu} = 0, \quad \frac{\partial L}{\partial \Sigma} = 0 \qquad \left(\frac{\partial L}{\partial \mu} \in R^{d}, \ \frac{\partial L}{\partial \Sigma} \in R^{d \times d} \right)$$

The $\hat{\mu}, \hat{\Sigma}$ that maximize the $L(\mu, \Sigma) = \log P(D; \mu, \Sigma)$ are "Sample" mean vector & "Sample" covariance matrix.

B. GMM : Gaussian Mixture Model

- GMM is a model with multiple Gaussian distributions.

- Generative model that follows GMM
  1) GMM modeling
  2) Parameter Estimation → "EM algorithm"

※ PB review : Joint probability & marginal probability & conditional probability

- J.P : $p(x, y)$, x, y가 동시에 일어날 확률

- Marginalize : 결합확률에서 특정 확률변수 제거
  - Discrete : $p(x) = \sum_{y} p(x, y)$
  - Continual : $p(x) = \int p(x, y)\, dy$

- C.P : $p(x, y) = \underbrace{p(x|y)}_{\text{C.P}} p(y)$

- Sampling in GMM
  ① 임의의 확률분포에 따라 하나의 정규분포 샘플링 → Categorical distribution
  ② 선택된 분포로부터 generate data
  $\Rightarrow$ iterate ①, ② for N times

✱ P(x) for GMM

- Categorical Distribution in GMM
  $p(z=k;\phi) = \phi_k$ / 각각의 클래스 출현확률 (parameter for categorical distribution), where k is a # of G.D in GMM
  ① $\forall k,\ \phi_k \geq 0$
  ② $\sum_k \phi_k = 1$

- μ, Σ for multiple G.D
  $\mu = \{\mu_1, \mu_2 \cdots, \mu_K\}$
  $\Sigma = \{\Sigma_1, \Sigma_2 \cdots, \Sigma_K\}$

- Probability density value of x in k^{th} G.D
  $p(x \mid z=k; \mu, \Sigma) = N(x; \mu_k, \Sigma_k)$

- Induce $p(x)$ from $p(x,z)$ by marginalization.
  $p(z=k;\phi) = \phi_k \quad \cdots ①$
  $p(x \mid z=k; \mu, \Sigma) = N(x; \mu_k, \Sigma_k) \quad \cdots ②$
  $p(x, z=k) = p(z=k)\,p(x \mid z=k) \quad \cdots ③$
  $p(x) = \sum_{k=1}^{K} p(x, z=k)$

$\therefore p(x) = \sum_{k=1}^{K} \phi_k N(x; \mu_k, \Sigma_k)$ "p(x) for GMM is a weighted summation of P.D in each distribution"

✱ x는 관측된 값이지만 z는 관측되지 않는 값이므로, 잠재변수 (latent variable)라고 한다.

* Parameter estimation for GMM

$$p(x) = p(x; \phi, \mu, \Sigma)$$
$$= \sum_{k=1}^{K} \phi_k N(x; \mu_k, \Sigma_k)$$

$$D = \{x^{(1)}, x^{(2)} \dots x^{(N)}\}, \quad \theta = \{\phi, \mu, \Sigma\}$$

$$p(D;\theta) = p(x^{(1)};\theta)\, p(x^{(2)};\theta) \cdots p(x^{(N)};\theta)$$
$$= \prod_{n=1}^{N} p(x^{(n)};\theta)$$

$$\Big\downarrow \log$$

$$L(\theta) = \log p(D;\theta) = \log \prod_{n=1}^{N} p(x^{(n)};\theta)$$

likelihood of D under GMM model

$$= \sum_{n=1}^{N} \log p(x^{(n)};\theta)$$
$$= \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \phi_k N(x^{(n)}; \mu_k, \Sigma_k) \right)$$

$\therefore \underset{\theta}{\operatorname{argmax}}\, L(\theta)$, where $\theta = \{\phi, \mu, \Sigma\}$: 로그가능도를 최대로 하는 θ를 찾는 것이 목표!

$\Leftrightarrow$ Solve $\frac{\partial L}{\partial \phi} = 0$, $\frac{\partial L}{\partial \mu} = 0$, $\frac{\partial L}{\partial \Sigma} = 0$!

- Cannot solve $\frac{\partial L}{\partial \phi}, \frac{\partial L}{\partial \mu}, \frac{\partial L}{\partial \Sigma} = 0$ 해석적으로 풀 수 X $\Rightarrow$ **Use EM Algorithm.**
  (직접적으로 풀 수 x)

$\Rightarrow$ why? 샘플데이터에서는 관측되지 않은 latent variable Z로 인해 특정 데이터 x에 대한 우도 $p(x|\theta)$가 Σ 형태를 가지고 결과적으로 $L(D;\theta)$가 log-sum의 형태를 가지기 때문.

샘플 데이터에서 관측되지 않은 잠재 변수 z로 인해,
특정 데이터 x에 대한 우도 $p(x;\theta)$가 **합(summation)**의 형태를 띠고,
결과적으로 샘플 D에 대한 전체 우도 $L(D;\theta)$가 **log-sum** 형태가 되어,
해석적으로 MLE를 구할 수 없게 된다.