06강. 가우시안 혼합 모델

◉ 세부목차

1. 가우시안 혼합 모델

⑴ 가우시안 혼합 모델의 필요성

① 확률 모델과 패턴인식

- 앞선 강의에서 살펴보아 온 것 같이, 패턴을 분류함에 있어서는 데이터들의 분포 특성을 분석하는 것이 매우 중요하다. 데이터의 분포 특성을 알기 위해서 적절한 확률밀도함수를 가정하여 데이터 분포에 대한 모델을 만드는 것을 확률 모델이라 한다.
- 가장 대표적으로 사용되는 확률 모델로 가우시안 확률 모델이 있다. 이는 하나의 클래스, 혹은 관찰된 전체 데이터 집합이 평균을 중심으로 하여 뭉쳐져 있는 분포 형태를 표현하는데 적합한 확률 모델이다. 앞서 3강과 4강에서 가우시안 확률밀도함수의 형태에 따른 데이터의 분포 형태를 살펴보고, 데이터를 이용하여 가우시안 분포의 파라미터(평균과 공분산)를 추정하는 방법에 대해서도 알아보았다.
- 그런데 가우시안 확률분포는, 가장 널리 사용되는 분포이기는 하지만, 데이터들이 평균을 중심으로 하나의 그룹으로 뭉쳐있는 unimodal한 형태만을 표현가능하다는 제약이 있다. 따라서 일반적인 확률분포를 추정하기 위해서는 보다 일반적인 형태를 표현 가능한 확률 모델이 필요한데, 이때 가장 손쉽게 생각해 볼 수 있는 것이 여러 개의 가우시안을 합하여 만들어지는 모델이다. 이를 가우시안 혼합 모델이라고 한다.

② 가우시안 확률 모델이 부적합한 데이터 분포

- {그림9-1}과 {그림9-2}에 나타난 데이터 분포를 표현하기 위한 확률 모델을 생각해 보자. 그림{9-1}의 1차원 데이터의 경우를 보면, 데이터가 밀집되어 있는 그룹이 3개 정도로 나뉘어져 있기 때문에, 평균을 중심으로 하나의 그룹으로 뭉쳐져 있는 형태를 표현

하는 가우시안 분포로는 그 특성을 제대로 표현하기 힘들다. 이를 바로 표현하기 위해서는 그림에서 보는 바와 같이, 3의 가우시안 분포를 함께 사용하는 것이 바람직하다.

- 또한 {그림9-2}를 보면, 도우넛 형태의 데이터 분포를 가지고 있어서, 이 역시 하나의 가우시안만으로 표현하는 것은 불가능하다. 그런데 그림에서와 같이 여러 개의 가우시안들이 각각 특정 영역을 맡아 표현할 수 있도록 하고 그것들을 연결하면 어느 정도 유사한 형태의 확률 모델을 찾을 수 있을 것이다.
- 이와 같이, 복수개의 가우시안 분포들의 합으로 새로운 확률분포를 나타내는 가우시안 혼합 모델을 사용하면, 가우시안 분포함수 하나가 나타낼 수 없었던 분포 특성 뿐 아니라, 아무리 복잡한 형태의 함수라도 충분한 개수의 가우시안 함수를 사용하기만 하면 원하는 만큼 정확하게 근사해 낼 수 있다.

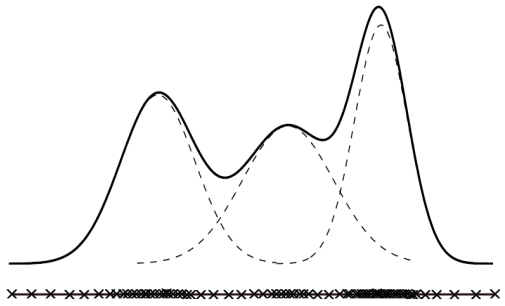


그림 9-1

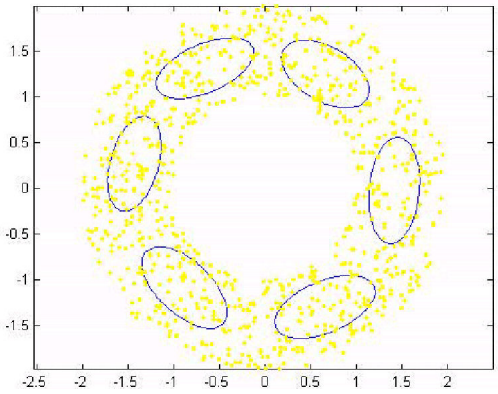


그림 9-2

2) 가우시안 혼합 모델의 정의

- M개의 간단한 확률밀도함수(혹은 성분(component))의 선형 결합으로 정의되는 전체 확률밀도함수는 다음 식과 같이 표현된다.

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \sum_{i=1}^{M} p(\mathbf{x}\,|\,\boldsymbol{\omega}_i, \boldsymbol{\theta}_i)\mathbf{P}(\boldsymbol{\omega}_i) \qquad (1)$$

- 여기서 $p(\mathbf{x}\,|\,\boldsymbol{\omega}_i, \boldsymbol{\theta}_i)$는 혼합 모델의 기본 성분을 이루는 간단한 확률밀도함수로, 가우시안 혼합 모델의 경우는 가우시안 확률밀도함수가 되지만, 다른 함수를 사용할 수도 있다.
- $\boldsymbol{\theta}_i$는 i번째 성분이 되는 확률밀도함수를 정의하는 파라미터로, 가우시안 확률밀도함수의 경우는 평균과 공분산 행렬이 된다.
- $\boldsymbol{\omega}_i$는 i번째 성분임을 나타내는 확률변수이고, $\mathbf{P}(\boldsymbol{\omega}_i)$는 i번째 성분이 전체 혼합 확률밀도함수에서 차지하는 상대적인 중요도를 의미하는 것이다.
- $\mathbf{P}(\boldsymbol{\omega}_i)$는 파라미터 α_i로 표시하기도 하며, 다음과 같은 성질을 만족해야한다.

$$0 \le \alpha_i \le 1 \text{ 그리고 } \sum_{i=1}^{M} \alpha_i = 1 \qquad (2)$$

- 이 혼합 확률 모델을 사용하여 데이터의 분포를 나타내기 위해 추정해야하는 파라미터 전체를 Θ로 나타내면, M개의 성분을 가지는 가우시안 혼합 모델의 파라미터는 다음과 같이 나타낼 수 있다.

$$\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_M, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, ..., \boldsymbol{\sigma}_M^2, \alpha_1, \alpha_2, ..., \alpha_M) \quad (3)$$

- 이 때 각 i 번째 성분의 가우시안 함수의 공분산 행렬은 $\sigma_i^2 I$인 경우를 생각하였다.

2. 가우시안 혼합 모델의 학습

⑴ 최우추정법

① 데이터 집합의 우도함수 (likelihood function of data set)

- 데이터 집합 $\mathbf{x}=\{x_1, x_2, \ldots, x_N\}$가 주어졌을 때, n번째 데이터 x_n의 확률밀도 $p(x_n)$를 가우시안 혼합 모델로 나타내고자 한다. 문제를 간단히 하기위해 여기서는 일단 데이터가 1차원인 경우를 생각한다.
- j번째 성분을 이루는 가우시안 분포가 평균 μ_j, 분산 σ_j^2를 가진다고 할 때, j번째 가우시안 확률밀도는 다음과 같이 표현된다.

$$p(\mathbf{x}_n \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_j} \exp\left(-\frac{(\mathbf{x}_n - \boldsymbol{\mu}_j)^2}{2\boldsymbol{\sigma}_j^2} \right) \quad (4)$$

- 이러한 개별 성분 M개를 결합한 혼합 확률밀도는 다음과 같이 얻을 수 있다.

$$p(\mathbf{x}_n \mid \boldsymbol{\theta}) = \sum_{j=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}) \mathbf{P}(\boldsymbol{\omega}_j \mid \boldsymbol{\theta}) \quad (5)$$

$$= \sum_{j=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) \alpha_j \quad (6)$$

- 따라서, 학습 데이터 전체에 대한 로그우도(log-likelihood)는 다음과 같이 정의된다.

$$E = -\log L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) \alpha_j \quad (9)$$

② 최우추정량(MLE)

- 데이터 집합에 대한 로그우도를 최대로 하는 최우추정량은 다음과 같이 정의된다.

$$\hat{\boldsymbol{\theta}} = \arg\max \left[\sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \boldsymbol{\theta}) \right]$$

$$= \arg\max \left[\sum_{n=1}^{N} \log \sum_{j=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\theta}_j) \mathrm{P}(\boldsymbol{\omega}_j) \right]$$

$$= \arg\max \left[\sum_{n=1}^{N} \log \sum_{j=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) \alpha_j \right] \quad (10)$$

- 최우추정량은 위의 E를 파라미터에 대해 미분하여 0이 되는 값을 구함으로써 얻어진다. 평균 μ_j의 최우추정치를 계산하는 과정을 알아보면, 먼저 E를 μ_j에 대해 편미분한 식을 다음과 같이 구한다.

$$\frac{\partial}{\partial \boldsymbol{\mu}_j} E = -\frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \boldsymbol{\theta}) \tag{11}$$

$$= -\sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\mu}_j} p(\mathbf{x}_n \mid \boldsymbol{\theta}) \tag{12}$$

$$= -\sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_{i=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)\alpha_i \tag{13}$$

$$= -\sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \frac{\mathbf{1}}{\sqrt{2\pi}\boldsymbol{\sigma}_j} \exp\left(-\frac{(\mathbf{x}_n - \boldsymbol{\mu}_j)^2}{2\boldsymbol{\sigma}_j^2} \right) \frac{-2(\mathbf{x}_n - \boldsymbol{\mu}_j)}{2\boldsymbol{\sigma}_j^2} (-1)\alpha_i \tag{14}$$

$$= -\sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) \frac{(\mathbf{x}_n - \boldsymbol{\mu}_j)}{\boldsymbol{\sigma}_j^2} \alpha_j \tag{15}$$

$$= -\sum_{n=1}^{N} \mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta}) \frac{(\mathbf{x}_n - \boldsymbol{\mu}_j)}{\boldsymbol{\sigma}_j^2} \tag{16}$$

- 위 마지막 항의 $\mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta})$는 데이터 $\mathbf{x}_n$이 주어졌을 때, 그것이 j번째 성분으로부터 나왔을 확률값으로 다음과 같이 나타낼 수 있음을 이용하여 계산되었다.

$$\mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}) \mathbf{P}(\boldsymbol{\omega}_j \mid \boldsymbol{\theta})}{p(\boldsymbol{\omega}_j \mid \boldsymbol{\theta})} \tag{7}$$

$$= \frac{p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)\alpha_j}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \tag{8}$$

- 이를 이용하여 편미분 값이 0이 되는 μ_j의 값을 찾으면 다음과 같이 계산할 수 있다.

$$\mathbf{0} = -\sum_{n=1}^{N} \mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta}) \mathbf{x}_n + \boldsymbol{\mu}_j \sum_{n=1}^{N} \mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta}) \tag{19}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{n=1}^{N} \mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta}) \mathbf{x}_n}{\sum_{n=1}^{N} \mathbf{P}(\boldsymbol{\omega}_j \mid \mathbf{x}_n, \boldsymbol{\theta})} \tag{20}$$

- 마찬가지로 나머지 파라미터 σ_j^2과 α_j에 대한 추정치도 계산하면 다음과 같다.

$$\hat{\sigma}_j^2 = \frac{\sum_{n=1}^{N} P(\omega_j \mid x_n)\left\| x_n - \hat{\mu}_j \right\|^2}{\sum_{n=1}^{N} P(\omega_j \mid x_n)} \qquad (22\text{-}1)$$

$$\hat{\alpha}_j = \frac{1}{N}\sum_{n=1}^{N} P(\omega_j \mid x_n) \qquad (23\text{-}1)$$

⑵ 파라미터의 반복계산 알고리즘

- 앞에서 얻어진 파라미터의 추정값에 대한 식을 보면, 각각의 파라미터의 값을 계산하기 위해서는 다른 파라미터의 추정값이 필요함을 알 수 있다.
- 따라서 다음과 같은 반복 알고리즘이 필요하다.

1. 모델 파라미터의 초기값을 임의로 설정 (Θ^{old} 값)
2. Θ^{old} 값을 이용하여 파라미터의 새로운 Θ^{new} 값을 얻기 위해 방정식의 우변을 계산
3. Θ^{old}와 Θ^{new}의 차이가 0에 가까워지면 알고리즘을 멈춤
4. 얻어진 Θ^{new}값을 Θ^{old}로 두고 2번을 반복

- 이러한 반복 알고리즘을 가우시안 혼합 모델의 EM 알고리즘이라고 한다.
- EM 알고리즘은 가우시안 혼합 모델 뿐 아니라 다양한 확률 모델에서 사용되는 학습법으로, 이후에 이를 보다 체계적으로 접근해 볼 것이다.

3. EM 알고리즘의 적용

⑴ 가우시안 혼합 모델의 예

- 남, 여 두 집단이 함께 속한 하나의 그룹으로부터 한 사람씩 뽑아서 신장을 측정하여 얻은 데이터를 $\{x_1, x_2, \ldots, x_n\}$라고 하자.
- 신장에 대한 확률변수를 x라고 했을 때, 확률밀도함수 $p(x)$를 추정하고자 한다.
- 그런데 x를 관찰한 데이터 집합에 대해 생각해 보면, 그 안에 여자들의 집단과 남자들의 집단이 존재하고, 이들 두 집단의 평균 신장은 일반적으로 다르다고 예상할 수 있다.

$$p(x) = \sum_{j=1}^{2} p(x \mid \mu_j, \sigma_j^2)\alpha_j = \alpha_1 p(x \mid \mu_1, \sigma_1^2) + \alpha_2 p(x \mid \mu_2, \sigma_2^2) \qquad (001)$$

- 따라서 $p(x)$는 두 개의 가우시안이 합쳐진 위와 같은 혼합 모델로 표현해 볼 수 있다.
- 이때, 1번째 가우시안이 여자 그룹의 데이터, 2번째 가우시안이 남자 그룹의 데이터를 나타낸다고 하자.
- 이제 데이터 집합 $\{x_1, x_2, \ldots, x_n\}$를 이용하여 파라미터 $\mu_j, \sigma_j^2, \alpha_j \ (j=1,2)$를 추정해야 한다. 그런데 이를 위해서는 각 데이터 x_i가 여자, 남자 두 그룹 중 어디에 속하는 것인지를 먼저 알고 있는 것이 도움이 될 것이다. 이러한 정보를 나타내는 새로운 변수 $z=(\omega_1, \omega_2)$ 를 생각한다. 이때, ω_1, ω_2는 각각 1 아니면 0의 값을 가진다.
- 즉, 현재 관찰된 데이터 x가 여자로부터 온 것이라면 ω_1는 1의 값, ω_2는 0의 값을 가지고, 남자로부터 얻어진 것이라면 반대로 ω_2는 1의 값, ω_1는 0의 값을 가지는 변수로

정의한다. 그런데 이 변수 **Z**는 우리가 얻은 데이터에는 포함되어 있지 않은 값으로, 외부에서 값이 측정될 수 없는 변수라 하여 숨겨진 변수(latent variable)라고 한다.

- 숨겨진 확률변수 z까지 모두 포함한 형태의 확률 모델을 다음과 같이 나타낼 수 있다.

$$p(\mathbf{X},\mathbf{Z}) = \sum_{j=1}^{2} p(\mathbf{X},\mathbf{Z}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\sigma}_j^2)\alpha_j \qquad (002)$$

⑵ EM 알고리즘의 적용 예

- 이제 관찰된 데이터 집합 $\{x_1, x_2, \ldots, x_n\}$을 이용하여 가우시안 혼합 모델의 파라미터, 즉 두 그룹의 평균과 분산, 그리고 두 그룹이 전체에 미치는 영향을 나타내는 파라미터를 추정해야 한다.
- 만약 우리가 관찰된 데이터와 함께 각 데이터들이 어떤 그룹으로부터 얻어진 것인지를 안다고 가정해보자. 즉, 각 데이터 x_i에 대해 숨겨진 확률변수의 값 $Z_i=(\omega_1,\omega_2)$ 도 관찰된 경우를 생각한다. (아래 그림 참조)

$$\begin{array}{ccccc} \overset{180}{\mathbf{x}_1}, & \overset{150}{\mathbf{x}_2}, & \ldots, & \overset{173}{\mathbf{x}_N} \\ \downarrow & \downarrow & & \downarrow \\ \underset{\text{남}}{\mathbf{z}_1}, & \underset{\text{여}}{\mathbf{z}_2}, & \ldots, & \underset{\text{남}}{\mathbf{z}_N} \end{array}$$

- 이와 같이 숨겨진 변수에 대한 값까지 모두 관찰되었다고 가정하면, 우리가 추정하고자 하는 파라미터는 다음과 같이 쉽게 얻어질 수 있다.

① 두 그룹의 평균과 표준 편차의 추정

z값에 따라 전체 데이터 집합을 여자 그룹과 남자그룹으로 각각 나눈다. 이렇게 하면 분리된 두 그룹에 대해 따로따로 가우시안 분포의 파라미터를 추정하는 문제와 같아지므로, 4강에서 배운 바와 같이 표본 평균과 표본 분산으로 추정할 수 있다.

② 각 그룹의 중요도를 타나내는 파라미터 α_j의 추정

α_j은 전체데이터에서 j번째 그룹이 차지하는 비율을 의미하므로, 다음과 같이 계산할 수 있다.

$$\alpha_j = \frac{\omega_j = 1\text{이 되는 데이터의 수}}{\text{전체 데이터의 수}} \qquad (003)$$

- 그런데, 일반적인 경우에, 숨겨진 변수에 대한 값은 알 수 없으며, 단지 **x**값만 관찰된다. 따라서 **Z**값 없이 파라미터를 추정하기 위해서는 EM 알고리즘과 같은 반복적인 학습법이 필요하게 된다. 이를 두 단계로 나누어 살펴보면 다음과 같다.
- E-step : 숨겨진 변수 **Z**의 기대치를 계산하는 단계

먼저 파라미터 값을 임의로 정한 다음, 정해진 파라미터들로 결정된 확률 모델을 이용하여 숨겨진 변수 **Z**의 기대치 E[**Z**]를 계산한다. 현재 사용하는 예에서 E[**Z**]=(E[ω_1].E[ω_2])를 계산하기 위해서는, 관찰된 각각의 데이터 x_i에 대해, 숨겨진 확률변수의 값 z_i가 (1,0)이 될 확률과 (0,1)이 될 확률을 계산하여야 한다. 현재 가지고 있는 파라미터 값을 사용하면 각각 다음과 같이 계산할 수 있다.

$$P(\omega_j = 1 \mid \mathbf{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2) = P(\omega_j \mid \mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)\alpha_j}{p(\mathbf{x}_n \mid \boldsymbol{\theta})} \qquad (004)$$

- M-step : E-step에서 계산된 숨겨진 변수 **z**의 값을 이용하여 파라미터를 새롭게 추정하는 단계

관찰된 각 데이터 $\mathbf{x}_i$가 어떤 그룹에서 나온 것인지는 정확히 알 수 없지만, 우리는 E-step에서 $P(\omega_1=1 \mid \mathbf{x}_i)$과 $P(\omega_2=1 \mid \mathbf{x}_i)$, 즉 $\mathbf{x}_i$가 여자 그룹에서 나왔을 확률과 남자 그룹에서 나왔을 확률을 각각 계산하였다. 이를 활용하면 다음과 같이 파라미터를 계산할 수 있다.

① 두 그룹의 평균과 표준 편차의 추정

데이터 각각이 정확히 어느 한 그룹에 속한다고 말할 수는 없으므로, 각 그룹의 평균을 계산할 때에는, 각 데이터가 그 그룹에 속하는 정도 $P(\omega_1=1 \mid \mathbf{x}_i)$과 $P(\omega_2=1 \mid \mathbf{x}_i)$를 반영하여 평균을 계산할 수 있다. 즉, 다음과 같은 평균 계산식을 얻을 수 있다.

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} P(\omega_j \mid \mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} P(\omega_j \mid \mathbf{x}_n)} \qquad (005)$$

이 결과는 2절에서 얻은 평균의 추정치와 같음을 알 수 있다. 분산에 대해서도 마찬가지로 추정이 가능하다.

② 각 그룹의 중요도를 타나내는 파라미터 α_j의 추정

α_j은 전체데이터에서 j번째 그룹이 차지하는 비율을 의미하므로, 마찬가지로 각 데이터에서 j번째 그룹이 차지하는 비율 $P(z=j \mid \mathbf{x}_i)$를 고려하여 계산하면 다음과 같은 식을 얻을 수 있다.

$$\alpha_j = \frac{1}{N}\sum_{n=1}^{N} P(\omega_j \mid \mathbf{x}_n) \qquad (006)$$

-M-step을 수행하고 나면, 이전의 파라미터에서 수정된 새로운 파라미터를 얻게 되고, 이것을 이용하여 다시 숨겨진 변수 z의 기대치 혹은 확률값을 계산하는 E-step을 수행할 수 있다. 이렇게 M-step과 E-step을 반복하면서 보다 정확한 파라미터를 추정하게 된다.

⑶ EM 알고리즘의 특성과 필요성

- 앞서 살펴본 예에서와 같이, 기본 확률변수 **x**의 값이 관찰되면서, 숨겨진 확률변수 **z**의 값도 함께 관찰된다면, EM과 같은 반복적인 학습에 의한 추정은 필요하지 않다.
- 즉, EM 알고리즘은 숨겨진 확률변수 **z**를 가지고 있는 확률 모델의 파라미터를 추정하기 위해 사용되는 방법이다.
- 먼저 E-step에서 숨겨진 확률변수의 기대치를 계산하여 관찰된 값 대신 사용할 수 있도록 한다. E-step의 E는 Expectation(기대치)를 의미한다.
- M-step에서는 관찰된 데이터 x와 숨겨진 확률변수의 기대치를 이용하여, 우도함수를 최대로 하는 파라미터의 값을 찾는다. M-step의 M은 Maximization(최대화)를 의미한다.

4. EM 알고리즘의 일반화

⑴ 숨겨진 변수를 가진 확률모델

- 숨겨진 변수란, 일반변수와는 달리 외부에서는 관찰될 수 없는 값을 가지는 변수를 말한다.
- 가우시안 혼합 모델에서는 관찰된 데이터 x가 혼합 모델의 어떤 요소로부터 주어졌는지를 나타내는 확률변수를 생각하면, 이는 일반적으로 관찰되지 않는 값으로, 숨겨진 변수라고 할 수 있다.
- 이밖에도 다양한 확률 모델에서 숨겨진 변수를 가지는 경우가 많다.
- 파라미터를 θ, 관찰된 데이터를 X, 숨겨진 변수를 Z라고 하면, 확률변수 전체에 대한 완전한 데이터 우도 (Complete data likelihood)는 결합확률밀도함수로 다음과 같이 주어진다.

$$\mathbf{P(X,Z\,|\,\theta)} \qquad (24)$$

- 여기서 X는 관찰된 데이터로 상수이고, 파라미터도 정해진 상수로, 결국 $\mathbf{P(X,Z\,|\,\theta)}$는 확률변수 Z의 함수로 볼 수 있다.

⑵ EM 알고리즘

① E-step (Expectation 단계)

- EM 알고리즘에서는 관찰된 데이터와 숨겨진 확률변수에 대한 완전한 데이터 우도를 최대화하는 파라미터를 찾는 것을 목적으로 한다.
- 그런데 숨겨진 확률변수 Z에 대한 데이터는 따로 주어지지 않으므로, 완전한 데이터 우도는 얻어질 수 없다. 따라서 $\mathbf{P(X,Z\,|\,\theta)}$를 사용하는 대신, 현재 주어진 파라미터 θ^{i-1}가 주어진 경우에 확률변수 Z에 대한 로그 우도 $\mathbf{logp(X,Z\,|\,\theta)}$의 기대치를 계산한다. 이를 Q함수로 표현하면 다음과 같다.

$$\mathbf{Q(\theta\,|\,\theta^{i-1}) = E_z[logp(X,Z\,|\,\theta)\,|\,X,\theta^{i-1}]} \qquad (25)$$

- 이 Q함수에서 X와 θ^{i-1}는 주어진 값이며, 확률변수 Z는 기대치를 계산함으로써 값으로 결정되므로, 결국 Q함수는 파라미터 θ의 함수가 된다.

② M-step (Maximization 단계)

- E-step에서 얻어진 Q함수의 값이 최대가 되는 파라미터 θ의 값을 찾는다. 즉, i번째의 파라미터는 다음과 같이 얻어진다.

$$\mathbf{\theta^i = \arg\max Q(\theta\,|\,\theta^{i-1})} \qquad (27)$$

③ 수렴 성질

- E-step과 M-step을 반복하게 되면 결과적으로 로그우도값은 증가하게 된다. 결국, EM 알고리즘을 통하면 우도함수의 국부 최대값에 수렴하게 된다.

5. 가우시안 혼합 모델의 EM 알고리즘 정리

- 일반화된 공분산을 가지는 가우시안 혼합 모델의 EM 알고리즘을 정리하면, 다음 파라미터 수정 공식을 수렴할 때까지 반복하는 것이다.
- E-step에서는 데이터 $\mathbf{x}_i$가 주어졌을 때, 숨겨진 확률변수의 값 $\mathbf{z}_i$의 기대치, 즉 각 성분으로부터 데이터가 얻어졌을 확률값은 다음과 같이 계산한다. 이때, 파라미터는 현재의 값, $\boldsymbol{\theta}^{i-1} = \left\{(\boldsymbol{\mu}_k^{i-1}, \Sigma_k^{i-1}, \alpha_k)\right\}_{k=1,...,M}$ 로 고정되어 있다.

$$\mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}_n) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{i-1}, \Sigma_k^{i-1})\alpha_\kappa}{\sum_{c=1}^{M} p(\mathbf{x}_n \mid \boldsymbol{\mu}_c^{i-1}, \Sigma_c^{i-1})\alpha_c} \qquad (007)$$

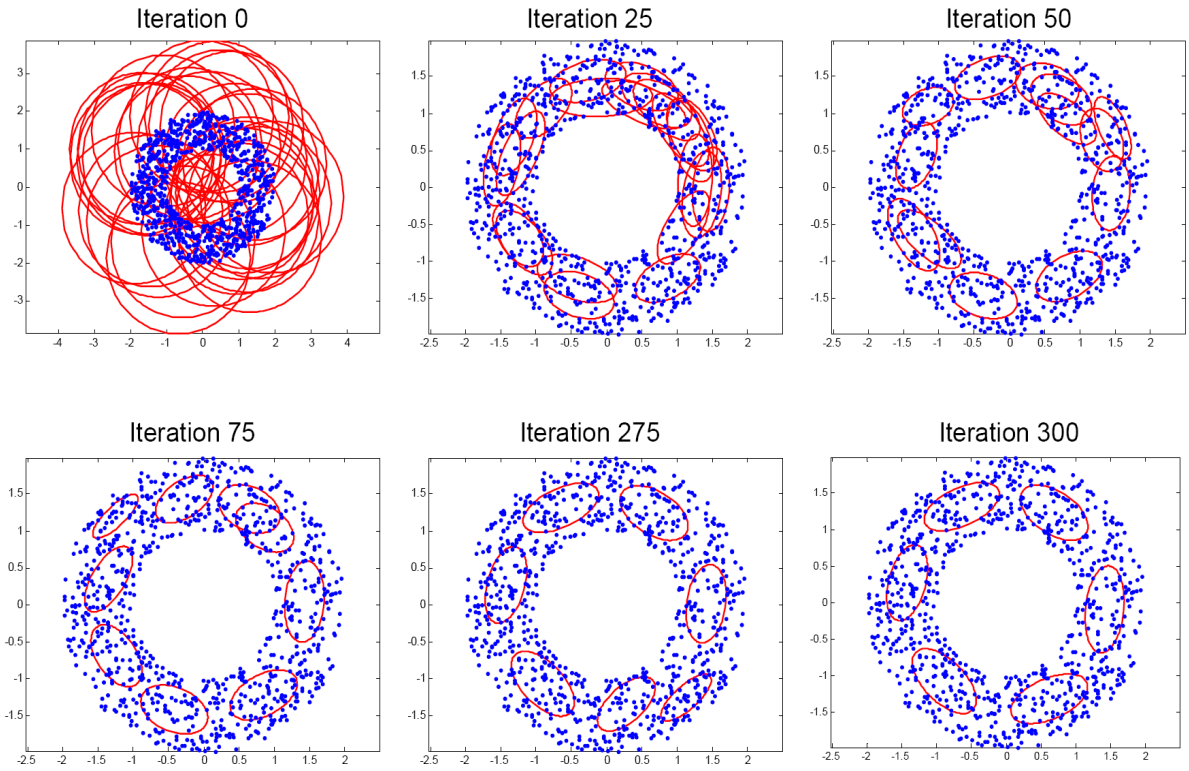
- M-step에서는 E-step에서 얻은 값을 이용하여 파라미터를 수정한다.

$$\alpha_k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}^n) \qquad (008)$$

$$\boldsymbol{\mu}_k{}^{i} = \frac{\sum_{n=1}^{N} \mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}^n)\mathbf{x}^n}{\sum_{n=1}^{N} \mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}^n)} \qquad (009)$$

$$\sum_k^i = \frac{\sum_{n=1}^{N} \mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}^n)(\mathbf{x}^n - \boldsymbol{\mu}_k{}^{i})(\mathbf{x}^n - \boldsymbol{\mu}_k{}^{i})^T}{\sum_{n=1}^{N} \mathbf{P}^{i-1}(\omega_k \mid \mathbf{x}^n)} \qquad (010)$$

- 이 알고리즘을 이용하여 1.1절의 그림 9-2에서 보인 도넛형 데이터에 대해 EM 알고리즘을 적용한 결과가 아래 그림에 나타나 있다.
- 위의 왼쪽에 있는 그림이 임의로 설정된 평균과 공분산을 가지는 각 성분들을 보여주며, E-step과 M-step을 반복해 감에 따라서 평균과 공분산이 변화하여 마지막 300번 반복한 결과 전체 데이터를 잘 표현할 수 있는 가우시안 분포 성분들을 찾아내었음을 확인할 수 있다. 이때, 각 성분들은 서로 다른 일반적인 공분산 행렬을 가진다.


Iteration 0
Iteration 25
Iteration 50
Iteration 75
Iteration 275
Iteration 300