〈MLE, Maximum Likelihood Estimation〉

θ(모수) = (μ, σ) = (평균, 표준편차) 이를 알고 있다면 ※ 이때는 정규 분포가 1개.

해당 표본이 발생할 확률값 $P(x|\theta)$을 추정가능하고 가장 높은 것을 찾는게 MLE.

N개의 표본이 있고, 표본 데이터는 동일한 정규 분포로부터 뽑혔고, 표본이 뽑히는 과정이 서로 독립적이라고 가정.

우도함수 = $L(\theta, x_1, x_2, \cdots, x_n) = P(x_1, \cdots, x_n ; \theta) = \prod_{i=1}^{N} N(x_i|\mu, \sigma^2)$ 이 식의 의미는 PPT의 MLE의 설명을 읽으면 이해가능.

〈Gaussian Mixture Model〉

※ π_k의 조건
① $0 \le \pi_k < 1$
② $\sum_{k=1}^{K} \pi_k = 1$

※ K : 전체 데이터 세트는 K개의 Gaussian distribution을 혼합하여 표현.

주어진 데이터 x에 대하여 GMM은 x가 발생될 확률 : $p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$

※ Σ_k : GMM에서 k번째 Gaussian distribution의 표준 편차

이 식의 의미는 k번째 가우시안 분포가 선택될 확률과 k번째 가우시안 분포의 평균과 표준 편차에서 x가 등장할 곱의 합.

※ π_k : mixing coefficient (혼합 계수), k번째 Gaussian distribution이 선택될 확률.

⇒ GMM을 학습시킨다는 것은 주어진 데이터 $X = \{x_1, x_2, \cdots, x_N\}$에 대하여 적절한 π_k, μ_k, Σ_k를 추정.

⇒ 우리가 구하려고하는 것은 K개의 Gaussian distribution에 대해서 parameter, 총 $3 \times K$
(parameter를 구한다는 것은 해당 Gaussian model을 정의한 것과 마찬가지이며, 뒤의 EM 알고리즘은 반복을 통하여 이를 구한다.)

〈GMM을 이용한 classification〉

주어진 데이터 x_n에 대하여 이 데이터가 어떠한 Gaussian distribution에서 생성되었는지를 찾는 것
이를 위해 필요한 것이 responsibility : 어떤 데이터 x가 소속되어 있는 distribution을 최종 결정하는 것.

responsibility $= r(z_{nk}) = p(z_{nk}=1 | x_n)$ x_n이 선택되었는데 z_{nk}이 1이 될 확률.

$z_{nk} \in \{0, 1\}$는 x_n이 주어졌을 때 GMM의 k번째 Gaussian distribution이 선택되면 1, 아니면 0
즉, $z_{nk}=1$인 것은 x_n이 k번째 Gaussian distribution에서 생성.
GMM을 이용한 classification은 x_n이 주어졌을 때, K개의 $r(z_{nk})$을 계산하여 가장 값이 높은 Gaussian distribution을 선택하는 것. 이를 베이지안 정리를 통해 구할 수 있다.

Likelihood   A prior probability

$$P(\theta|X) = \frac{P(X|\theta)p(\theta)}{P(X)}$$

A posterior probability   Marginal probability

- **X**: 관측된 데이터, Observation을 뜻한다. 우리가 갖고 있는건 이거다!! (머신러닝에서는 트레이닝 데이터.....)
- **Θ**: Hypothesis를 말하는데 데이터를 통해 추정하고자 하는 값이 되겠다. classification문제에서는 각 discrete한 클래스가 될 수 있고 linear regression의 경우 추정하려고 하는 weight들이 될 수 있다. 그 외에 추정하고자 하는 모든 문제에서 추정하고싶은 target값이 된다.
- **Marginal probability P(X)**: 데이터 X자체의 분포를 뜻한다.
- **A prior probability(사전확률) P(Θ)**: 사전에 가지고 있는 확률을 말한다. 이때 Θ는 mutually independent해야한다.예를들어. "하늘이 파랗다"/"하늘이 파랗지 않다" 이 두가지 가정은 독립이고 두 확률의 합이 1이 되어야한다. (사전적으로 hypothesis에 대한 지식이 없을 때는 그냥 hypothesis의 element들이 가질 확률이 모두 같다고 두자)
- **Likeihood(우도) P(X|Θ)**: hypothesis를 두고, 다시말해 어떤 가정을 한 상태의 데이터의 분포를 뜻한다.
- **A posterior probability(사후확률) P(Θ|X)**: observation이 주어졌을 때의 hypothesis의 분포를 뜻한다. 얘같은 경우에는 데이터 X의 영향을 반영하는 애다.

$$r(z_{nk}) = P(z_{nk}=1 | x_n)$$

$$= \frac{\boxed{P(z_{nk}=1)}\,\boxed{P(x_n | z_{nk}=1)}}{\boxed{\sum_{j=1}^{K} P(z_{nj}=1)\, P(x_n | z_{nj}=1)}}$$

$$= \frac{\boxed{\pi_k}\,\boxed{N(x_n | \mu_k, \Sigma_k)}}{\boxed{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)}}$$

π_k와 $P(z_{nk}=1)$은 k번째 가우시안 분포가 선택될 확률.

$P(z_{nj}=1)$이 π_j로 치환.

〈EM 알고리즘을 이용한 학습〉

EM 알고리즘 : iteration based algorithm을 통해 GMM의 parameter를 찾는 것.

기댓값 최대화 알고리즘.

모수에 관한 추정값으로 로그가능도의 기댓값을 계산하는 기댓값(E) 단계와

이 기댓값을 최대화하는 모수 추정값들을 구하는 최대화(M) 단계를 번갈아가면서 적용.

일반적으로 GMM은 주어진 데이터 $X=\{x_1, x_2, \cdots, x_N\}$에 대해 EM 알고리즘을 적용하여 parameter인 π, μ, Σ를 추정.

parameter를 추정하기 위해 log-likelihood $\mathcal{L}(X;\theta)$를 정의.

$$\mathcal{L}(X;\theta) = \ln p(X|\pi,\mu,\Sigma) = \ln\left\{\prod_{n=1}^{N} p(x_n|\pi,\mu,\Sigma)\right\} = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\Sigma_k)\right\}$$

앞 페이지 참고 → 보조.

앞 페이지 참고.

직관적으로 likelihood는 어떠한 모델에서 데이터 $X=\{x_1, x_2, \cdots x_N\}$가 생성되었을 확률을 의미.

계산의 편의성을 위해 log-likelihood를 사용하여 log-likelihood를 최대화하는 π, μ, Σ를 추정하는 것은 주어진 데이터 X를 가장 잘 표현하는 GMM을 구성하는 것과 동일한 의미.

⇒ 이를 위해 각각의 μ_k, Σ_k, π_k에 대해 $\mathcal{L}$를 편미분

⇒ 최대화 = 미분값이 0일 때.

※. μ_k에 대하여 편미분.

$$\frac{\partial \mathcal{L}(X;\theta)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)} \Sigma_k^{-1}(x_n - \mu_k) = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k) = 0$$

미분을 했을 때 어떻게 이렇게 나왔는지 공부했는데 이해가 안돼서 식을 첨부합니다.

$$\therefore \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad (7)$$

※. Σ_k에 대하여 편미분.

$$\frac{\partial \mathcal{L}(X;\theta)}{\partial \Sigma_k} = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)} \left\{\frac{1}{2}\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1} - \frac{1}{2}\Sigma_k^{-1}\right\} = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \gamma(z_{nk})\{\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T - 1\} = 0$$

이 부분도 이해가... 잘 안돼서 이미지만 첨부하였습니다.

$$\therefore \Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad (8)$$

※· π_k에 대하여 편미분. (이해가 잘 안돼서 식에 대한 설명은 생략하겠습니다.. T)

π_k는 $\sum_{k=1}^{K} \pi_k = 1$ 의 조건을 만족하면서 log-likelihood를 최대화해야 한다. 따라서 π_k는 라그랑주 승수법을 이용하여 추정하며

라그랑지안 $J(X;\theta,\lambda)$는 다음과 같다.

$$J(X;\theta,\ \lambda) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\ \Sigma_k) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right) \tag{9}$$

$J(X;\theta,\lambda)$를 π_k에 대해 편미분하면 λ를 구할 수 있다.

$$\begin{aligned}
\frac{\partial J(X;\theta,\ \lambda)}{\partial \pi_k} &= \sum_{n=1}^{N} \frac{N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - \lambda = 0 \\
\Leftrightarrow & \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - \lambda \sum_{k=1}^{K} \pi_k = 0 \\
\Leftrightarrow & \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) - \lambda = 0 \quad \left(\because \sum_{k=1}^{K} \pi_k = 1\right)
\end{aligned}$$

$$\therefore \underline{\lambda = N} \quad \left(\because \sum_{k=1}^{K} \gamma(z_{nk}) = 1\right) \tag{10}$$

$\lambda = N$을 대입하여 π_k를 추정할 수 있다.

$$\begin{aligned}
\frac{\partial J(X;\theta,\ \lambda)}{\partial \pi_k} &= \sum_{n=1}^{N} \frac{N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - N = 0 \\
\Leftrightarrow & \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - N\pi_k = 0
\end{aligned}$$

$$\therefore \pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \tag{11}$$

GMM의 parameter 추정을 위한 EM 알고리즘의 E-step에서는 모든 데이터와 Gaussian distribution에 대해 responsibility인 $r(z_{nk})$를 계산하고 M-step에서는 보완식 변경된 식들[7,8,11]을 이용하여 모든 Gaussian distribution에 대한 π, μ, Σ를 추정한다. 이러한 E-step과 M-step을 선언한 횟수 또는 특정 값으로 수렴할 때까지 반복한다.