A tutorial on Dirichlet process mixture modeling (1)

1. Intorduction

DP : cluster의 갯수까지도 찾아주는 clustering 방법

(별 내용 없음. 이 논문은 DP를 쉽게 접하고 연습하기 위한 거다 + 심화적인 방법에 대한 reference)

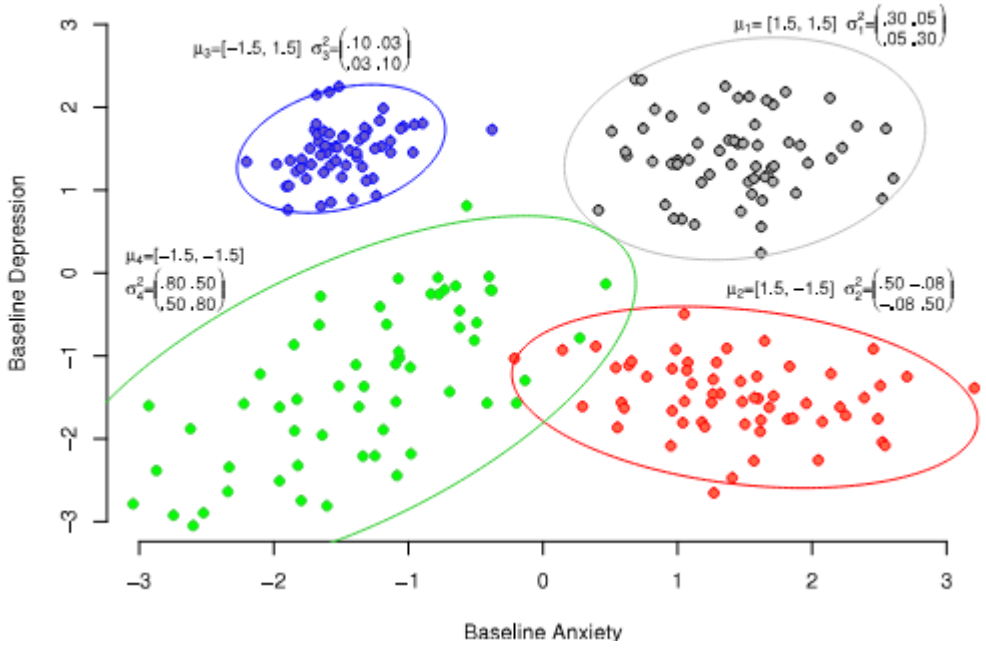
2. Finite Mixture Model

2.1 설명할 때 쓸 data

$$(y_1, y_2) \sim BVN(\mu_k, \sigma_k^2)$$

k=1,2,3,4. 즉, 4개의 평균과 분산이 다른 cluster에서 생성한 2차원 데이터. 아래 그림과 같음.

§ 2 에서는 cluster 갯수가 정해진 것이라고 가정(기존의 방식).



2.2 Model Likelihood Function In a Finite Mixture Model

GMM model을 예를 들어보자. likelihood는 다음과 같다.

$$p(y_i|\mu_1,\ldots,\mu_k,\sigma_1^2,\ldots,\sigma_k^2,\pi_1,\ldots,\pi_k)=\sum_{k=1}^{K}\pi_k\mathcal{N}(y_i;\mu_k,\sigma_k^2),$$

그 후 indicator variable c_i(c_i=1,2,...,k)를 도입할 것이다. c_i는 ith participant가 있는 cluster를 나타내는 latent variable로, 그 값은 π_k의 realization으로 볼 수 있다. 따라서 c_i의 값을 알 수 있다면 likelihood는 이렇게 쓸 수 있다.

$$p(y_i|c_i=k)=\mathcal{N}(y_i|\mu_k,\sigma_k^2).$$

2.3 Joint posterior distribution by Bayes' Theorem

$$\begin{aligned}p(\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{c}|y)&=\frac{p(y|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{c})p(\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{c})}{p(y)}\\&\propto p(y|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{c})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)p(\boldsymbol{c}).\end{aligned}$$

마지막항은 위에서 구한 data likelihood와 prior들의 곱으로 이루어져 있다. 세 parameter는 독립이므로 prior를 각각 썼다. 우선 prior부터 보자.

2.3.1 Priors

$$
\begin{gathered}
p(\mu_k) \sim N(\mu_0, \sigma_0^2) \\
p(\sigma_k^2|\gamma, \beta) \sim \Gamma^{-1}(\gamma, \beta) \propto (1/\sigma_k^2)^{\gamma-1} exp(-\beta/\sigma_k^2) \\
p(\pi_1, \pi_2, ...\pi_k|\alpha) \sim Dirichlet(\alpha/K, \alpha/K, .., \alpha/K)
\end{gathered}
$$

- σ_k^2의 경우 univariate을 예로 들어 inverse-gamma를 썼지만 multivariate인 경우 inverse-wishart를 사용.
- 분포의 모수들은 일단 고려하지 않음.
- 마지막에 제시된 Dirichlet분포는 multinomal의 conjugate prior임.

2.4 Conditional posterior Distribution

μ_k, σ_k^2, π_k 를 Gibbs sampling으로 구하기 위한 conditional posterior distribution을 구할거임.

참고로 σ_k^2 대신 precision인 τ_k 를 사용할 예정.

계산 과정은 그냥 conjugacy 계산하는 거라 어렵지 않아서 생략하고, 결과만 쓰겠음. 굳이 알고싶으면 Appendix A.2,3 ㄱㄱ

$$
\begin{gathered}
p(\mu_k|\mathbf{y}, c=k) \propto p(\mathbf{y}, c=k)p(\mu_k) \sim N\left(\frac{\bar{y_k} n_k \tau_k + \mu_0 \tau_0}{n_k \tau_k + \tau_0}, \frac{1}{n_k \tau_k + \tau_0}\right) \\
p(\tau_k|\mathbf{y}) \propto \tau^{\frac{n_k}{2}} exp\left(-n_k \sum_{[i]c=k} (y_i - \bar{y_k})^2\right) \times \tau^{\alpha-1} exp(-\beta\gamma) \\
= \tau^{\alpha-1+\frac{n_k}{2}} exp\left(-\tau(\beta + \frac{1}{2} \sum_{[i]c=k} (y_i - \bar{y_k})^2)\right) \\
p(c_1, c_2, ..., c_k|\pi_1, ..., \pi_k) = \prod_{k=1}^{K} \pi_k^{n_k} (multinomial)
\end{gathered}
$$

이를 이용해 Gibbs sampling은 다음과 같이 진행된다.

$$
\begin{aligned}
\pi_k^{(t)} &\sim \text{Dirichlet}\,(n_1 + \alpha/K - 1, \ldots, \\
&\qquad n_k + \alpha/K - 1)\,, \\
\boldsymbol{\mu}_k^{(t)} | \boldsymbol{c}^{(t-1)}, \boldsymbol{y}, \tau_k^{(t-1)} &\sim \mathcal{N}\left(\frac{\bar{y}_k n_k \tau_k^{(t-1)} + \mu_0 \tau_0}{n_k \tau_k^{(t-1)} + \tau_0}, n_k \tau_k^{(t-1)} + \tau_0 \right), \\
\tau_k^{(t)} | \boldsymbol{c}^{(t-1)}, \boldsymbol{y}, \boldsymbol{\mu}_k^{(t-1)} &\sim \text{Gamma}\Big(\alpha - 1 + n_k/2, \\
&\qquad \beta + \frac{1}{2} \sum_{[i]c=k} (y_i - \bar{y}_k)^2 \Big), \\
\boldsymbol{c}_i^{(t)} &\sim \text{Multinomial}\big(1, \pi_1^{(t)} \mathcal{N}(y_i | \mu_1^{(t)}, \tau_1^{(t)}), \ldots, \\
&\qquad \pi_k^{(t)} \mathcal{N}(y_i | \mu_k^{(t)}, \tau_k^{(t)})\big).
\end{aligned}
$$

두번째줄 정규분포에는 평균과 precision을 쓴 것 같음. 분산은 역수.

*pi랑 c 분포의 모수는 뒤에 CRP와 DPMM이 적용된(적용하기 위한) 것으로 보면 됨.

3. Infinite Mixture Model

이제부터는 cluster의 갯수가 infinite하다고 가정하고 할 것임(=cluster의 갯수를 미리 지정하지 않고 문제를 품)

3.1 Tackling the problem of infinite number of clusters

1. infinite vector of mixing proportions

$$
\begin{aligned}
p(c_1, \ldots, c_k | \alpha) &= \int p(c_1, \ldots, c_k | \pi_1, \ldots, \pi_k) \\
&\quad \times p(\pi_1, \ldots, \pi_k | \alpha) d\pi_1 \cdots d\pi_k \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^k} \int \prod_{k=1}^{K} \pi_k^{n_k + \alpha/K - 1} d\pi_k \\
&= \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}. \qquad (7)
\end{aligned}
$$

π 를 통해 marginalize 하는데, 앞의 항은 multinomial(π_k), 뒤의 π_k~Dirichlet(α/K) 이므로 합칠 수 있다(두번째줄까지). π_k에 대한 적분이 문제인데, 이 적분값은 Dirichlet 분포를 적분하면 1이 됨을 보이는 식에서 적분값을 따올 수 있고 그 결과 마지막 줄과 같은 식이 나온다.

그러나 아직까지는 infinite한 건 아니고 갯수가 정해져 있는 상태

3.2 When K approaches infinity

2. indicator c_i외의 모든 indicator들이 주어졌을 때 c_i의 conditional prior를 구해야 한다. (A.3 참)

$$p(c_i = k|\boldsymbol{c}_{-i}, \alpha) = \frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha},$$

그 과정을 살펴보자.

$$\begin{aligned}
p(c_i = k|\boldsymbol{c}_{-i}, \alpha) &= p(c_i = k|c_1, \ldots, c_{i-1}) \\
&= p(c_1, \ldots, c_{i-1}, c_i = k)/p(c_1, \ldots, c_{i-1}) \\
&= \frac{\cancel{\Gamma(\alpha)}}{\Gamma(n+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\cancel{\Gamma(\alpha/K)}} \Bigg/ \\
&\quad \frac{\cancel{\Gamma(\alpha)}}{\Gamma(n+\alpha-1)} \prod_{k=1}^{K} \frac{\Gamma(n_{k,-i} + \alpha/K)}{\cancel{\Gamma(\alpha/K)}} \\
&= \frac{1}{\Gamma(n+\alpha)} \prod_{k=1}^{K} \Gamma(n_k + \alpha/K) \Bigg/ \\
&\quad \frac{1}{\Gamma(n+\alpha-1)} \prod_{k=1}^{K} \Gamma(n_{k,-i} + \alpha/K) \\
&= \frac{\Gamma(n+\alpha-1)}{\Gamma(n+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(n_{k,-i} + \alpha/K)}
\end{aligned}$$

조건부 확률을 그대로 나타냈고, 분모와 분자는 3.1에서 구한 식(7)을 활용하였다. 그 과정에서 약분되는 감마함수값을 약분해줬다.

이제 마지막 줄에서 $\prod$ 부분을 먼저 보자.

recurrence property of the Gamma function, $\Gamma(x+1) = x\,\Gamma(x)$, so that

$$\prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(n_{k,-i} + \alpha/K)} = \frac{(n_{k,-i} + \alpha/K)\,\Gamma(n_{k,-i} + \alpha/K)}{\Gamma(n_{k,-i} + \alpha/K)}$$
$$= \frac{n_{k,-i} + \alpha/K}{1}.$$

감마함수의 성질을 사용해 파이 안의 항을 약분시켜 간단하게 만들었다. 감마함수 안에 있는 값 중 분모가 ith respondent를 제외한 kth cluster의 사람수를 나타냈으므로 1차이 나서 약분이 가능한 것으로 보인다.

이제 $\prod$ 앞의 감마함수항을 보자. 역시 분모와 분자가 1차이 이므로 쪼개면 약분이 가능하고, 그 결과 n+α-1 만 남게 된다. 결과적으로 제일 위와 같은 식이 나오는 것이다.

$$p(c_i = k|\boldsymbol{c}_{-i}, \alpha) = \frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha},$$

3.3 Exchangeablility

- Exchangeability

  : cluster에 할당되는 확률은 각 data point가 들어오는 순서와는 상관없이 invariant 하다.

  → 위에서 구한 conditional distribution에서 각 data point가 들어오는 순서가 가장 마지막이라고 생각할 수 있다.

- cluster에 할당될 확률(위의 conditional prob.)은 식을 보면 몇 번째 cluster에 들어가는 것은 의미없고, kth cluster에 현재 들어가있는 수에만 depend 하는 것을 알 수 있다.(only depends on $n_{-i,k}$). 이게 만족하는이유는 가정한 Dirichlet 분포의 모수가 symmetric 하기 때문이다(모두 a/K로 같기 때문).
- 이는 같은 현재원을 갖는 cluster라면 다음 data point가 할당될 확률도 같다는 것을 의미하기도 한다.

$$\text{기존}\quad \text{클러스터로} : p(c_i|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,k}}{n - 1 + \alpha}$$
$$\text{새로운}\quad \text{클러스터로} : p(c_i \neq c_k \forall k \neq i|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n - 1 + \alpha}$$

여기서 전체 cluster의 갯수 K가 무한대로 간다면 위의 conditional prob.은 다음과 같이 쓸 수 있다.

$$p(c_i|\boldsymbol{c}_{-i},\alpha)=\frac{n_{-i,k}}{n-1+\alpha}.$$

여기서, 모든 cluster에 들어갈 확률에 대해서 합을 구하면 1보다 작은 수가 나오는데, 1에서 그 값을 빼준 확률이 바로 새로운 cluster를 만들어 들어가는 확률이 된다.

$$\begin{aligned}1-\frac{\sum_k n_{-i,k}}{n-1+\alpha}&=\frac{n-1+\alpha}{n-1+\alpha}-\frac{n-1}{n-1+\alpha}\\&=\frac{(n-1+\alpha)-(n-1)}{n-1+\alpha}=\frac{\alpha}{n-1+\alpha}.\end{aligned}$$

$$\begin{aligned}&\text{clusters where } n_{-i,k}>0: && p(c_i|\boldsymbol{c}_{-i},\alpha)=\frac{n_{-i,k}}{n-1+\alpha},\\&\text{all other clusters combined}: && p(c_i\neq c_k\ \forall j\neq i|\boldsymbol{c}_{-i},\alpha)=\frac{\alpha}{n-1+\alpha}.\end{aligned}$$

이 식에서 DP의 중요한 성질 중 하나가 나오는데, 바로 α 값이 클 수록 새로 cluster를 만들 확률이 높다는 것이다. 이 α를 concentration parameter라고 부르기도 한다.

3.4 The Chinese Restaurant Process

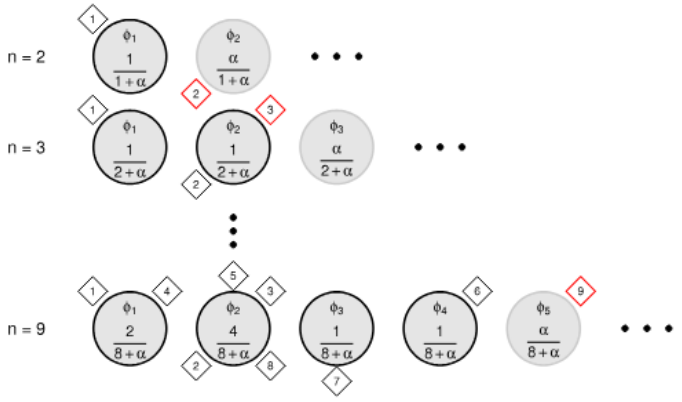


Fig. 3. An illustration of the Chinese Restaurant Process. In the CRP metaphor, imagine a Chinese restaurant with an infinite number of tables. Customers (individual data entries, shown as diamonds) enter into the restaurant one by one and are seated at tables (discrete clusters, shown as circles), in the order in which they enter into the restaurant. There are parameters associated with the clusters, represented as $\phi_k=[\mu_k,\tau_k]$ for the cluster means and precisions. The first customer who enters into the restaurant always sits at the first table. The second customer enters and sits at the first table with probability $1/(1+\alpha)$, and the second table with probability $\alpha/(1+\alpha)$, where α is a positive real number (top row, where i = 2). When the third customer enters the restaurant, he or she sits at each of the occupied tables with a probability proportional to the number of previous customers already sitting there, and at the next unoccupied table with probability proportional to α.

3.5. Prior and posterior predictive distributions

CRP에서 cluster assignment probability를 베이지안 개념을 적용하여 발전시키고자 함.

1. Posterior predictive distribution(PPD)

지금 있는 데이터에 기반하여, 새로 관측된 값의 posterior probability를 구하는 것. 앞으로 표기에서 ith 사람의 새롭게 관측된 값을 $\tilde{y}_i$라고 할 것.

위에서 제시한 식은 ith observation이 이미 있는 cluster에 얼마나 fit한지를 고려하지 않고 확률이 결정된다.

$$\begin{aligned}\tilde{y} &\sim \mathcal{N}(\mu_p, \sigma_p^2 + \sigma_y^2), \\ &= \mathcal{N}\left(\frac{\bar{y}_i n_k \tau_k + \mu_0 \tau_0}{n_k \tau_k + \tau_0}, \frac{1}{(n_k \tau_k + \tau_0)} + \sigma_y^2\right).\end{aligned}$$

수식 3.5.1(1)

어떻게 이 식이 나왔는지 살펴보기 위해선 우선 다음의 원리를 알아야 한다.

$$y \sim \mathcal{N}(\mu, \sigma_y^2).$$

Also, if the unknown mean μ is expressed with a prior belief of

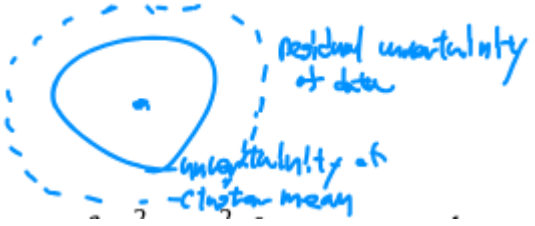
$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2),$$

then the posterior predictive distribution of a newly observed $\tilde{y}$ given the data you have already seen is

$$\tilde{y} \sim \mathcal{N}(\mu_0, \sigma_y^2 + \sigma_0^2).$$

수식 3.5.1(2). 새로 들어오는 data의 **prior** predictive distribution

이는 아래 그림처럼 cluster mean의 uncertainity에 data의 residual uncertainity가 합쳐진 것으로 볼 수 있다.



이제 y가 kth cluster라고 가정하고, 2.4에서 구한 μ_k의 condi tional dist.를 적용하면

$$\mu_k \sim N(\mu_p, \sigma_p^2)$$
$$where \quad \mu_p = \frac{\bar{y_k} n_k \tau_k + \mu_0 \tau_0}{n_k \tau_k + \tau_0}, \quad \sigma_p^2 = n_k \tau_k + \tau_0$$

이렇게 나온걸 위의 수식 3.5.1(2)에 적용한다면 수식 3.5.1(1)을 유도할 수 있다. 이것이 새로운 observation의 kth cluster의 PPD이다. 이는 prior predictive distribution에 kth cluster에 있는 정보가 추가되어 만들어진 PPD라고 볼 수 있다.

이렇게 얻어진 kth cluster에 대한 PPD는 새로운 observation과 kth cluster과의 proximity로 볼 수도 있다. 따라서 각각의 cluster에 대한 PPD를 구하면 각 cluster의 proximity 가 벡터형식으로 나올 것이다.

2. Prior predictive distribution

$$\tilde{y_i} \sim N(\mu_0, \sigma_0^2 + \sigma_y^2)$$

새로운 cluster가 생기기 전엔 그 cluster의 mean과 precision을 알 수 없는 상태에서, μ의 분포를 prior로 가정하여 만든 distribution이다.

3. **DPMM** = CRP+PPD (DPMM의 정의)

Chinese Restaurant Process를 보면 크게 두 가지로 나뉘는데, (1) 기존에 있는 cluster로 배정되는 경우와 (2) 새로운 cluster를 만들어 거기에 배정되는 경우이다. (1)의 경우엔 kth cluster의 predictive dist.를 구할 때 kth cluster의 정보 (kth cluster의 size나 precision, sample mean 등)를 사용할 수 있지만 (2)의 경우는 mean과 precision을 모르므로 이를 이용할 수 없다 → (1)과 (2)에 대해서 다른 dist.를 적용해야 한다!!

∴위에서 살펴본 PPD(Posterior Predictive Dist.)와 Prior predictive dist.의 차이에 기반하여, (1)에는 PPD를, (2)에는 Prior predictive dist.를 적용할 것이다.

- (1) 기존의 cluster에 배정되는 경우

$$p(c_i | \mathbf{c}_{-\mathbf{i}}, \mu_k, \tau_k, \alpha)$$
$$\propto p(c_i | \mathbf{c}_{-\mathbf{i}}, \alpha) p(\tilde{y_i} | \mu_k, \tau_k, \mathbf{c}_{-\mathbf{i}})$$
$$\propto \frac{n_{-i,k}}{n - 1 + \alpha} N(\tilde{y_i}; \frac{\bar{y_k} n_k \tau_k + \mu_0 \tau_0}{n_k \tau_k + \tau_0}, \frac{1}{n_k \tau_k + \tau_0} + \sigma_y^2)$$

- (2) 새로운 cluster를 만들어 배정되는 경우

$$
\begin{aligned}
& p(c_i \neq c_k \quad \forall k \neq i | \mathbf{c}_{-i}, \mu_0, \tau_0, \alpha) \\
\propto\ & p(c_i \neq c_k \quad \forall k \neq i | \mathbf{c}_{-i}, \alpha) \times \int p(\tilde{y_i} | \mu_k, \tau_k) p(\mu_k, \tau_k | \mu_0, \tau_0) d\mu_k d\tau_k \\
\propto\ & \frac{\alpha}{n-1+\alpha} N(\tilde{y_i}; \mu_0, \sigma_0^2 + \sigma_y^2)
\end{aligned}
$$

적분 안에서 두 번째 항은 prior에서 나올 수 있는 μ_k, τ_k 의 확률이고, 첫 번째 항은 이렇게 나온 μ_k, τ_k 의 realization을 모수로 했을때 구해지는 $\tilde{y_i}$의 likelihood이다. 이렇게 쪼갠 뒤에 나올 수 있는 모든 realization에 대해 적분해주면 prior predictive distribution이 나온다. (Appendix A.1,2

*CRP는 DP의 probabiliistic realization를 나타내는 확률 과정이라는 사실을 기억하자.

3.6. Relationship between DP, CRP, and DPMM

DP : a distribution over distributions. ‘fixed number categories’를 가진 다른 많은 분포들에 대해 Dirichlet dist.를 확장하는 과정

CRP : Infinite clustering을 위한, 일종의 distribution. ‘predictive distribution는 고려하지 않음’

DPMM : predictive distribution을 고려하여 CRP를 적용한 mixture model.

4. DPMM Algorithm