PRML과 부수적인 자료들을 공부하여 간단히 정리하였습니다.
bg
observed variable와 latent variable에 대한 joint distribution을 정의한다고 해보자. 이 때 observed variable에 대한 확률 분포를 구하고 싶은 경우 latent variable의 marginalization을 진행하면 된다. 이 의미는 복잡한 형태의 분포를 가진 observed variable에 대한 분포를 다룰 때,
좀 더 다루기 쉬운 observed variable와 latent variable의 joint distribution을 이용할 수 있다는 것이다. 즉, latent variable을 도입함으로서 복잡한 분포 모델을 좀 더 쉬운 형태의 분포들의 조합으로 변경할 수 있다.

- 여기서는 K-mean, Gaussian mixture model, EM algorithm 에 대해 공부할 것이다.
- discrete latent variable의 경우에 해당한다. (continuous latent는 12장에서 공부한다)

9.1 K-means Clustering

- D 차원의 데이터가 N개 있다

우리는 이 데이터들을 K개의 cluster로 분류하고자 한다.

- 각 cluster의 중심점을 $\boldsymbol{\mu}_k$ 라고 하자.
- $r_{nk} \in \{0, 1\}$: data가 k cluster에 속하면 1, 나머지는 0
- object function (*distortion measure* 라고 부른다) :

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

각 cluster에 속하는 data들과 중심점의 거리를 더한식이다.

- 우리는 이 식을 최소화하는 $r_{nk}, \boldsymbol{\mu}_k$ 를 **iterative** 하게 찾으면 된다.
  - 먼저, $\boldsymbol{\mu}_k$의 초깃값을 설정한다.
  - 그리고 r_{nk}에 대해 object function을 최소화 한다. (E-step)
  - 다음 r_{nk}를 고정하고 object function을 $\boldsymbol{\mu}_k$에 대해 최소화 한다. (M-step)
  - converge할때까지 반복한다. 단, global이 아닌 local minimum으로 converge할수 있다.
- r_{nk}의 경우 쉽게 말해 데이터가 각 k개의 중심점들 중에 가장 가까운 k 에 1의 값을 가진다.
- $\boldsymbol{\mu}_k$의 경우 J에서 quadratic 형태이므로 미분하여 그 값을 구한다. J를 미분하면 $\boldsymbol{\mu}_k = \frac{\sum r_{nk}\mathbf{x}_n}{\sum r_{nk}}$ 의 값을 가지고 이는 해당 k cluster에 속하는 data들의 평균값을 의미한다.
  - 그래서 *K-means* 알고리즘이라고 불린다.

K-means 알고리즘의 몇 가지 특징을 살펴보자.

- 유클리디안 기법을 사용하기 때문에 각 변수별로 scale을 맞출 필요가 있다.
- 모든 data별로 거리를 계산하기 때문에 속도가 느릴 수 있고 이를 해결하기 위한 논문이 많이 나와 있다.(tree, sequential update...)
- cluster 갯수를 직접 정해야 한다. Bayesian 접근법으로 문제를 해결할 수 있다.
- Hard clustering이다. 확률적인 접근이 없다.
- data와 중심점을 유클리디안으로 계산하기 때문에 categorical 변수에는 적합하지 않고 outlier에 취약하다.
- 이에 대해 data point간의 dissimilarity를 다르게 계산(유클리디안 거리이외의 다른 방법)하는 *K-medoids* 알고리즘이 있다. (써봤는데 그닥 좋은지 모르겠다)

9.2 Mixtures of Gaussians

PRML 2장에서는 GMM을 Gaussian component의 linear 조합으로 생각했지만 이번에는 latent variable의 개념을 추가해서 이해해보자.

- Gaussian mixture distribution
  - π_k는 weight의 역할을 하며 multinomial distribution에서의 확률이고

- 뒤의 mixture component N은 multivariate gaussian distribution이다.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

$\mathbf{z}$ 는 discrete latent variable이다. K 차원의 binary random variable이며 하나의 원소만 1을 가지며 나머지는 0을 갖는다. $z_k \in \{1, 0\}$ 인 것이다. 이는 multinomial distribution의 확률변수라는 것을 알 수 있다.

- marginal distribution of z

$$p(z_k = 1) = \pi_k$$

- $0 \leq \pi_k \leq 1$
- $\sum_{k=1}^{K} \pi_k = 1$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- conditional distribution of x

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k}$$

위의 식들을 이용하여 우리는 처음에 보았던 Gaussian mixture distribution을 구할 수 있다

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

- Conditional probability of z given x
  - it can be viewd as the *responsibility* the component k takes for 'explaning' the observation x
  - posterior 로 이해할 수 있다.

$$\gamma(z_{nk}) = p(z_k = 1|\mathbf{x}_n) = \frac{p(z_k = 1)p(\mathbf{x}_n|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}_n|z_j = 1)} = \frac{\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n|\boldsymbol{\mu}_j, \Sigma_j)}$$

9.2.1 Maximum likelihood

- N * D data set X
- N * K latent matrix Z
- Log likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\}$$

이를 maximize 하려고 하는데 문제가 생긴다.

- log안에 summation이 있어서 미분을 하여 closed form의 형태로 구할 수 없다.
- sigularity : 특정 점이 어떠한 평균값과 같은 값을 가지고 그 분포의 분산이 0으로 가면 그 점에서의 확률값이 무한으로 가고 logL도 무한으로 간다.
- identifiability : K! 개의 같은 solution이 생긴다.

9.2.2 EM for Gaussian mixtures

log likelihood 를 각 parameter(μ, Σ, π)에 대해 미분하면

$$0 = -\sum_{n=1}^{N} \frac{\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\Sigma_j \pi_j N(\mathbf{x}_n|\boldsymbol{\mu}_j, \Sigma_j)} \Sigma_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

- 평균이 각 data에 posterior 가중치로서 곱해져 구해진다.
- where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

mixing coefficient π 는 제약식이 있기에 Lagrange로 풀면

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$0 = \sum_{n=1}^{N} \frac{N(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_j \pi_j N(\mathbf{x}_n|\boldsymbol{\mu}_j, \Sigma_j)} + \lambda$$

$$\pi_k = \frac{N_k}{N}$$

k-means보다 많은 iteration을 해야되기 때문에 초기값을 k-means를 통해 정하면 좋다. 또한 EM을 통해 구한 값이 local maximum일 수도 있다.

- EM for GM 정리

1. 초깃값을 설정한다.
2. E-step : $\gamma(z_{nk})$ 구하기

$$\gamma(z_{nk}) = \frac{\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(\mathbf{x}_n|\mu_j, \Sigma_j)}$$

3. M-step : 주어진 responsibility로 μ, Σ, π 구하기

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k^{new} = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. log likelihood 구해서 converge할 때까지 반복

$$\ln p(\mathbf{X}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

9.3 An Alternative view of EM

$\ln p(\mathbf{X}/\pi, \mu, \Sigma)$에서 위의 식을 보면 log 안에 있는 summation(intergral) 때문에 maximum likelihood solution을 구하기 어렵다. 이를 해결하기 위해 latent variable을 추가하여 사용한다. maximization of complete-data log likelihood function은 상대적으

로 쉽다고 가정하자. 하지만 latent variable 때문에 complete log likelihood를 그대로 이용하기 어렵고 대신에 Expectation을 취해서(posterior distribution for latent variable을 통해) 이를 최대로 만드는 parameter를 구하고 반복한다.

- E-step
  - 현재 우리가 알고 있는 parameter θ^{old}를 이용하여 latent variable의 posterior $p(\mathbf{Z}/\mathbf{X}, \theta^{old})$를 구한다.
  - 이를 이용하여 expectation of the complete-data log likelihood를 구한다.

$$Q(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

- M-step
  - 이 식을 최대화하는 θ^{new}를 구한다.

$$\theta^{new} = \arg\max_{\theta} Q(\theta, \theta^{old})$$

9.3.1 Gaussian mixture revisited

위에서 설명한 대로 GM을 모델링해보자.

- complete likelihood

$$p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} N(\mathbf{x}_n|\mu_k, \Sigma_k)^{z_{nk}}$$

- posterior of latent Z
  - $p(\mathbf{z}) = \prod_{k=1}^{K} \pi^{z_k}$
  - $p(\mathbf{x}/\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}/\mu_k, \Sigma_k)^{z_k}$

$$p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k)]^{z_{nk}}$$

이를 이용하여 expectation of complete-data likelihood function을 구해보자.

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] = E_{\mathbf{Z}}[\sum_{n} \ln p(\mathbf{x}_n, \mathbf{z}_n|\theta)] = \sum_{n} E_{\mathbf{Z}}[\ln p(\mathbf{x}_n, \mathbf{z}_n|\theta)]$$
$$= \sum_{n=1}^{N} E_{\mathbf{Z}}[\ln(p(\mathbf{z}_n)p(\mathbf{x}_n|\mathbf{z}_n, \theta))] = \sum_{n=1}^{N} E_{\mathbf{Z}}[\ln[\prod_{k=1}^{K}(\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k))^{z_{nk}}]]$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} E_z[(z_{nk}) \ln(\pi_k N(\mathbf{x}|\mu_k, \Sigma_k))] = \sum_{n=1}^{N} \sum_{k=1}^{K} E_z[z_{nk}] \ln(\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k))$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln N(\mathbf{x}_n|\mu_{\mathbf{k}}, \Sigma_k)\}$$

- $E[z_{nk}] = \frac{\sum z_{nk}[\pi_k N(\mathbf{x}_n/\mu_k, \Sigma_k)]^{z_{nk}}}{\sum [\pi_k N(\mathbf{x}_n/\mu_k, \Sigma_k)]^{z_{nj}}} = \frac{\pi_k N(\mathbf{x}_n/\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n/\mu_j, \Sigma_j)} = \gamma(z_{nk})$

이제 차례대로 E-step, M-step을 진행하면 된다.

9.3.2 Relation to K-means

k-means를 GM의 특별한 케이스라고 생각할 수 있다. covariance matrices를 추정해야되는 것이 아니라 고정된 ϵI라고 생각하고 이 값이 0으로 가는 경우 이는 k-means와 같은 결과를 낸다.

- responsibility

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n - \mu_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n - \mu_j\|^2/2\epsilon\}}$$

위의 식에서 ϵ이 0으로 간다고 가정하자.

- $\|\mathbf{x}_n - \mu_k\|^2$ 값이 가장 작은 cluster를 k라고 하자.
- k 이외의 $\exp\{-\|\mathbf{x}_n - \mu_j\|^2/2\epsilon\}$ 값들은 더 빠르게 0으로 수렴한다. (exponentially)
- 즉, $\gamma(z_{nk})$ 값이 k에서만 1이고 나머지는 0의 값을 가지는 것이다. (binary indicator)
- $\gamma(z_{nk})$이 이전의 K-means에서 봤던 γ_{nk}이 되는 것이다.

이제 ϵ이 0으로 간다고 가정하고 expected complete-data log likelihood 값을 보면

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] \rightarrow -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \mu_k\|^2 + const$$

expected complete-data log likelihood를 최대화하는 것은 결국 K-means에서 object function을 최소화하는 것과 같아졌다. wow!

9.3.3 Mixtures of Bernoulli distribution

latent class analysis의 예시이다. (구제적인 내용은 skip)

9.3.4 EM for Bayesian linear regression

$\mathbf{w}$를 marginalization했었는데 이를 latent variable로 취급하고 EM algorithm을 사용하는 예시이다. (구체적인 내용은 skip)

9.4 The EM Algorithom in General

우리의 목표는 maximize the likelihood function that is given by $p(X/\theta) = \sum_Z p(X, Z/\theta)$ 이다. (Z는 discrete, continuous 다 가능 intergral로 바꾸면 됨) 하지만 이를 바로 optimization하기는 힘들지만 complete likelihood function 를 optimization하는 것이 더 간단하자고 가정하자. 그러면 우리는 아래와 같은 decomposition을 생각할 수 있다. (임의의 $q(Z)$는 pdf라고 생각할 수 있다 defined over the latent variables)

- log likelihood는

$$\ln p(\mathbf{X}|\theta) = \ln \sum_z q(\mathbf{Z})\frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

by Jansen's inequality

$$\ln p(\mathbf{X}|\theta) \geq \sum_z q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}/\theta)}{q(\mathbf{Z})} = L(\theta, q)$$

여기서 $L(\theta, q)$는 **log likelihood function의 Lower Bound** 라고 할 수 있으며 우리는 이를 최대한 높여서 log likelihood를 최대화하고자 한다. 그런데 지금 $q(Z)$에 대해 optimization를 할 수 없는 상황(임의의 q를 구할만한 정보가 없음)이고 이를 위해 추가적인 접근법이 필요하다.

$$\begin{aligned}
L(\theta, q) &= \sum_z q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \\
&= \sum_z q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta)}{q(\mathbf{Z})} \\
&= \sum_z \{q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} + q(\mathbf{Z}) \ln p(\mathbf{X}|\theta)\} \\
&= \ln p(\mathbf{X}|\theta) + \sum_z q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \\
&= \ln p(\mathbf{X}|\theta) - \sum_z q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)}
\end{aligned}$$

여기서 first term은 log likelihood이고 second term이 KL-divergence이다.

$$KL(q||p) = -\sum_{z} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}\} = \sum_{z} q(\mathbf{Z}) \ln\{\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\theta)}\} \geq 0$$

KL-divergence는 항상 0이상의 값을 가지기 때문에 **Lower Bound를 최대화하기 위해서는 KL 을 0 의 값을 갖게 해야한다.** 여기서 우리는 $q(Z)$에 대한 정보를 찾을 수 있다. 즉, t 시점에서 $q(Z^t) = p(Z/X, \theta^t)$ 로 하면 된다. 이를 통해 우리는 다시 $\theta^{t+1} = argmax Q(\theta, q^t)$ 를 하여 계속 iterative하게 반복하여 MLE parameter(with latent variable)의 값을 구할 수 있다.

- **E-step**
  - lower bound L 은 θ^{old}는 고정한채로 $q(Z)$에 대해 최대화한다.
  - 하지만 $\ln p(X/\theta)$는 $q(Z)$와 상관이 없기 때문에 Lower Bound의 최대값은 KL값이 0을 가져야한다.
  - KL = 0 을 위해서는 $q(Z) = p(Z/X, \theta^{old})$ (posterior)의 조건을 만족해야한다.
  - 그러면 lower bound랑 log likelihood가 같아진다. (Lower Bound가 최대화되며)
- **M-step**
  - 위의 과정에 따라 $q(Z)$는 posterior로 fixed 되고 L을 최대화하는 새로운 θ^{new}를 구한다.
  - 이 새로운 값 때문에 KL은 non zero가 되고 log likelihood는 lower bound보다 더 큰 값을 가진다.
  - converge할 때까지 iterative하게 반복한다.

또 다른 표현으로는

$$\begin{aligned} L(q,\theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{Z}|\mathbf{X},\theta^{old}) \\ &= Q(\theta,\theta^{old}) + const \end{aligned}$$

- 첫번째 항이 expectation of the complete-data log likelihood
  - joint distribution이 exponential family인 경우, log를 취했을 때 쉽게 maximize할 수 있다.
- 두번째 항은 θ에 independent하기에 const

따라서 **lower bound를 높이는 것이 expectation of the complete-data log likelihood를 최대화하는 것** 이다.

한줄결론 : observed log likelihood를 최대화하는 parameter(MLE)를 구하고 싶다. latent variable에 대한 posterior distribution으로 E[complete log likelihood]를 최대로 하게 만드는 θ를 구하면 된다.

+ EM for MAP

- monk의 유투브를 보고 간단히 정리

MLE를 구하는 과정과 거의 비슷하다. 단지 prior만 추가될뿐이다!

- E-step

$$R(\theta, \theta_{t-1}) = E_{\theta_{t-1}}[\ln p(\mathbf{X},\mathbf{Z}|\theta)|\mathbf{X}=\mathbf{x}] + \ln p(\theta)$$

- M-step

$$\theta_t = argmax R(\theta, \theta_{t-1})$$