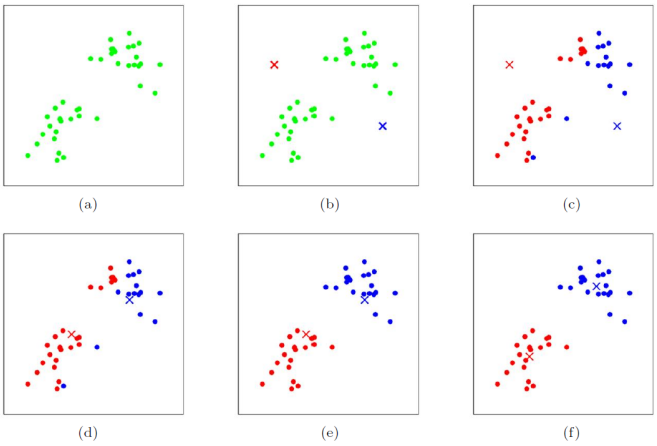
14강 : Expectation-Maximization Algorithm

Unsupervised learning

K-means clustering

비지도 학습 중 하나인 k-means clustering에 대해 알아보자.

k-means clustering은 어떻게 라벨링이 없는 데이터를 특정 그룹으로 묶을 수 있는 것일까? 다음의 예시를 통해 원리를 알아보자



(a) 라벨이 없는 training set $\{x^{(1)}, \ldots, x^{(m)}\}$이 주어졌다.

(b) 미리 지정한 클러스터 개수 (여기서는 2)만큼 랜덤하게 중심점(빨간점, 파란점)를 선정한다

(c) 데이터별로 어느 중심점에 더 가까운지를 기준으로 색을 정한다.

(d) 빨간색점들끼리의 평균을 구하고 이를 새로운 중심점을 할당한다.(파란색점도 같은 방식으로)

(e) 새로 정해진 중심점을 기준으로 데이터들이 어느 중심점에 더 가까운지를 비교하여 색을 정한다.

(f) 새롭게 할당된 클러스터의 중심점을 구한다.

이 과정을 반복하다 보면 특정 지역에 중심점이 수렴하면서 데이터를 두 클러스터로 구분할 수 있게 되었다.

- algorithm

1. Initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in R^n$ randomly

   여기에서 k는 우리가 원하는 클러스터의 개수로 알고리즘의 파라미터이다.

   데이터셋이 매우 많을 경우 보통 m개의 example 중에서 k개의 중심점을 선택한다고 한다. 물론 데이터 셋이 적을때도 가능하지만 임의로 선택하는 것과 큰 차이는 없다고 한다.

2. Repeat until convergence{

   For every i, set

$$\text{For every } i, \text{ set} \quad c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

$$\text{For each } j, \text{ set} \quad \mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

   첫번째 줄은 각 training example $x^{(i)}$를 가장 가까운 클러스터 중심점 μ_j에 할당하는 과정이다.(데이터에 색을 정하는 과정)

   두번째 줄은 클러스터에 할당된 point의 평균으로 클러스터 중심점 μ_j를 옮기는 과정이다.

- k-means algorithm guaranteed to converge?

먼저 distortion function을 다음과 같이 정의할 수 있다.

$$J(c, \mu) = \sum_{i=1}^m ||x^{(i)} - \mu_{c^{(i)}}||^2$$

즉 J는 각 training example $x^{(i)}$마다 해당 데이터가 할당된 클러스터의 중심점 $\mu_{c^{(i)}}$과의 거리를 더한 값이다. 이 값을 줄이는 방향으로 알고리즘이 진행되어야 하지만 종종 bad local minima에 빠질 수 있다. 이때는 k-means를 여러번 실행해서 가장 낮은 J값을 선택하는 걸로 문제를 해결할 수 있다.

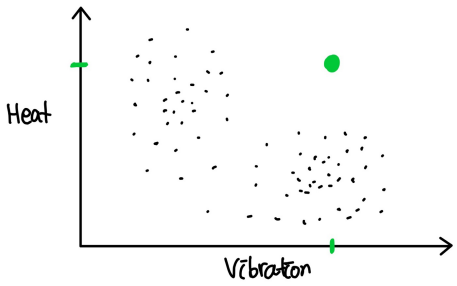
- how to decide k?

응 교수님은 “I still usually choose k by hand”라고 말씀하셨다. 즉, 우리는 목적에 맞게 k를 정하면 되는 것이다!!

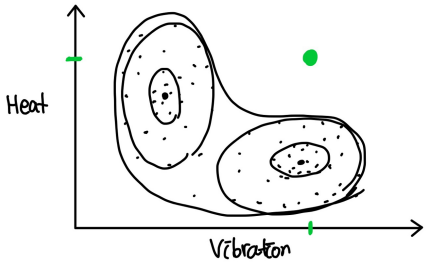
Mixtures of Gaussians and the EM algorithm

이번 파트에서는 밀도 추정(density estimation)을 위한 EM(Expectation-Maximization)에 대해 알아볼 것이다.

그 전에 비행기 엔진을 만드는 공장에서 만들어진 엔진의 불량을 탐지하는 모델을 만든다고 생각해보자. 시험 테스트 중에 나온 데이터를 진동과 발열을 가지고 그래프에 나타내면 다음과 같다. 이 때 녹색 점의 경우를 이상치로 탐지해야 한다. (anomaly detection)



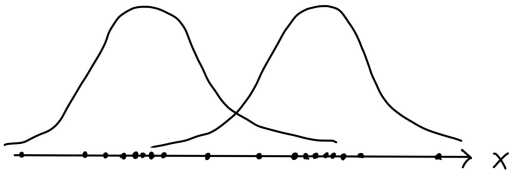
주어진 데이터셋에 대해 어떻게 모델을 세울 수 있을까? 분포가 간단하지 않은 경우 모델을 세우는 것이 어려울 수 있지만 데이터의 양상을 자세하게 살펴보면 다음과 같이 두개의 가우시안 분포가 나타나는 것을 알 수 있다.



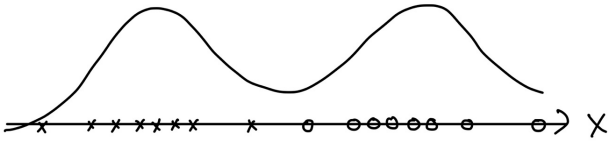
따라서 다음과 같이 분포가 간단하지 않은 경우 2개의 가우시안 분포로 전체 분포를 설명할 수 있다.

이를 GMM gaussian mixture model이라고 하는데 쉬운 설명을 위해 1차원에서 예시를 들어 설명하고자 한다.

다음과 같이 x데이터가 주어졌을 때 이것은 2개의 가우시안 분포에서 왔음을 알 수가 있다.



따라서 두개의 분포를 연결하고 각 가우시안 분포로 얻은 데이터를 x, o 으로 표시하면 다음과 같다. 확률이 낮은 중앙부분은 움푹 들어간 모양을 나타낸다.



그러나 결과적으로 우리는 어떤 데이터가 어떤 가우시안 분포에서 왔는지 알지 못한다.

이때 EM algorithm은 어떤 데이터가 어떤 가우시안 분포에서 왔는지 알지 못할 때도 모델을 세울 수 있도록 도와준다.

Mixture of Gaussians Model GMM

suppose

1. a latent(hidden/ unobserved) random variable: z
2. $x^{(i)}, z^{(i)}$ 의 joint distribution :

   $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$

   where $z^{(i)} \sim Multinomial(\phi)$ $\quad\quad$ $\phi_j = p(z^{(i)} = j)$

   $z \in \{1, ..., k\}$

   $x^{(i)}|z^{(i)} = j \sim N(\mu_j, \Sigma_j)$

그러므로 각 $x^{(i)}$는 랜덤하게 뽑힌 $z^{(i)}$에 의해 생성되고 $x^{(i)}$는 $z^{(i)}$에 의존하는 k개의 가우시안 분포로 부터 나왔다는 것을 알 수 있는데 이때의 모델을 **mixture of gaussian model**이라고 부른다.

그리고 z를 latent random variable이라고 부르는데 이는 hidden/ unobserved라는 의미이다. 이것이 우리의 estimation problem을 어렵게 만든다!

- optimization

모델의 파라미터를 추정하기 위해 기존의 likelihood function을 적어보자…

$$\begin{aligned}\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)}; \phi, \mu, \Sigma) \\ &= \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma)p(z^{(i)}; \phi).\end{aligned}$$

그러나 이 식을 미분하고 미분한 식을 0으로 두어 푸는 방식으로는 closed form에서의 MLE를 구할 수 없다! (응 교수님께서 집에서 스스로 계산해보라고 하셔서 푸는 건 생략!)

만약 $z^{(i)}$를 알고 있다면 ML 문제는 다음과 같이 쉽게 풀렸을 것이다.

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).$$

Maximizing this with respect to ϕ, μ and Σ gives the parameters:

$$\begin{aligned}\phi_j &= \frac{1}{m}\sum_{i=1}^{m} 1\{z^{(i)} = j\}, \\ \mu_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}, \\ \Sigma_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}.\end{aligned}$$

그러나 우리의 density estimation problem에서는 $z^{(i)}$이 알려지지 않았다. 그러면 어떻게 해야할까?

EM(Expectation-Maximization)

이때 우리는 EM algorithm을 사용하게 된다.

EM 알고리즘은 2개의 main step이 반복적으로 일어나는 알고리즘이다.

1. E-step : $z^{(i)}$ 값을 추론해가는 과정이다.
2. M-step : 우리의 추론을 바탕으로 모델의 파라미터를 추정하는 과정이다.

- algorithm

$$
\begin{aligned}
&\text{Repeat until convergence: } \{ \\
&\quad \text{(E-step) For each } i, j, \text{ set} \\
&\qquad w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) \\
&\quad \text{(M-step) Update the parameters:} \\
&\qquad \phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}, \\
&\qquad \mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \\
&\qquad \Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}} \\
&\}
\end{aligned}
$$

- E-step

먼저 E-step에서는 실제 $z^{(i)}$값을 모르기 때문에 주어진 값들을 이용해서 기댓값을 찾는 단계이다. 여기서 $z^{(i)}$는 $x^{(i)}$가 뽑힐 수 있는 가우시안 분포 j 를 의미한다.

우리가 구해야 할 사후 확률은 bayes rule을 이용해서 구하면 다음의 식을 얻는다.

$$
p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}
$$

먼저 분자에서 $p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)$는 $N(\mu_j, \Sigma_j)$인 가우시안 분포로 부터 계산할 수 있다.

그리고 분자에서 $p(z^{(i)} = j)$는 $z \sim Multinomial(\phi)$이기 때문에 ϕ_j이다.

최종적으로 $w_j^{(i)}$은 $z^{(i)}$값에 대한 soft guesses를 나타낸다.

이때 soft란 0에서 1 사이의 값을 취해서 확률로 추론을 하는 것이다. 이와 반대로 hard guess란 {0 or 1}/ {1, … ,k} 중에 택1 처럼 하나의 best single값을 얻는 것을 의미하는데 k-means clustering이 여기에 속한다. 특정 클러스터 영역에 할당된 point들은 예외 없이 특정 클러스터의 point라고 규정짓기 때문이다.

마지막으로 우리는 모든 개별 training example에 대해 $w_j^{(i)}$를 계산하고 저장해야 한다.

- M-step

이제 이전 단계의 우리의 추론을 바탕으로 모델의 파라미터를 추정해보자.

(기존의 파라미터를 나타내는 식)과 (z를 아는 경우의 파라미터를 나타내는 식)과 비교해보면, 다음과 같은 부분에만 차이가 있다는 것을 알 수 있다.

◗ → ◗

우리는 이러한 단계를 거쳐서 여러 개의 분포가 합쳐지는 하나의 분포 곡선을 만들 수 있다.

이로써 찾은 함수를 통해 입력값 x를 실행했을 때 특정 기준인 ϵ보다 크거나 같으면 정상 데이터고 작으면 이상치(anomaly)로 판단할 수 있다.

Derivation of EM

앞서서는 EM algorithm을 GMM을 세우기 위해 이용하는 역할로 설명했는데 이번 파트에는 더 넓은 관점에서 살펴볼 것이다..!

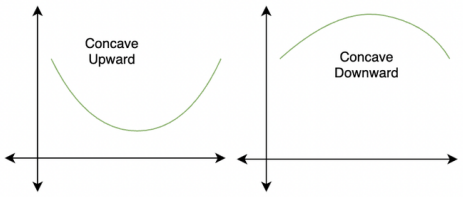
이에 앞서 Jensen's inequality에 대해 알아보자

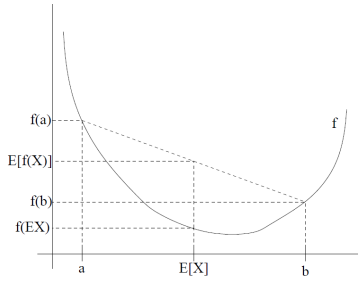
1. Jensen's inequality

let f be a convex function (e.g $f''(x) > 0$)

let x be a random variable

then, $f(EX) \leq E[f(x)]$





더 나아가

if $f''(x) > 0$ (f is strictly convec),

then $E[f(x)] = f[EX] \iff$ x is constant ($x = E[x]$ with probability 1)

우리는 jensens's inequality 공식을 로그함수에 적용하여 증명할 수 있다. 이때 로그함수는 concave 함수이기 때문에 위의 공식도 concave함수인 경우로 변형하여 사용하면된다. 이럴 경우 위의 모든 부등식을 반대로 바꾸면 된다.

2. The EM algorithm

- x에 대한 분포

우선 latent variable model $p(x, z; \theta)$를 이용하면 x에 대한 분포는 다음과 같이 나타낼 수 있다.

$$p(x; \theta) = \sum_z p(x, z; \theta) \tag{1}$$

- MLE

우리는 log likelihood를 최대화하는 파라미터값을 얻길 원하는데 log likelihood는 다음과 같이 정의 된다.

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x^{(i)}; \theta) \tag{2}$$

이를 joint density $p(x, z; \theta)$를 기준으로 다시 나타내면 다음과 같다.

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)}; \theta) && (3) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta). && (4) \end{aligned}$$

그러나 결과적으로 non-convex optimization problem을 얻기 때문에 명확하게 파라미터에 대한 MLE를 구하기는 어렵다.

- EM 알고리즘을 이용한 전략

이때 latent random variable $z^{(i)}$를 알고 있다면 MLE를 쉽게 구할 수 있다. 여기에서도 EM algorithm을 사용하면 효과적으로 답을 구할 수 있다. $l(\theta)$를 최대화하기 어렵다면 우리의 전략은 l의 lower bound를 반복적으로 만들어서(Estep), 그 lower bound에 optimize하는 것이다.(M-step)

이때 우리의 로그 가능도 함수는 합의 형태인데 계산의 편리를 위해 하나의 example x에 대해 optimize하고 나중에 sum을 더해주면서 전체 optimize 결과를 얻을 것이다.

하나의 example에 대한 가능도 함수는 다음과 같다.

$$\log p(x; \theta) = \log \sum_z p(x, z; \theta) \tag{5}$$

그 다음으로 일단 Q를 z값이 나올 수 있는 아무 분포라고 해보자. 즉, $\sum_z Q(z) = 1, Q(z) \geq 0$이다. 그럼 우리의 가능도 함수는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \log p(x; \theta) &= \log \sum_z p(x, z; \theta) \\ &= \log \sum_z Q(z) \frac{p(x, z; \theta)}{Q(z)} && (6) \\ &\geq \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} && (7) \end{aligned}$$

이때 (7)에는 jensen's inequality를 적용하였다. 적용이 가능한 이유는 다음과 같다.

(6)에서

$$\log \sum_z Q(z) \frac{P(x,z;\theta)}{Q(z)}$$
$$= \log \mathop{\mathbb{E}}_{z\sim Q} \frac{P(x,z;\theta)}{Q(z)}$$
$$\geq \mathop{\mathbb{E}}_{z\sim Q} \log \frac{P(x,z;\theta)}{Q(z)} \quad (\log : \text{concave function})$$
$$= \sum_z Q(z) \log \frac{P(x,z;\theta)}{Q(z)}$$

따라서 Q 분포에 상관없이 (7)은 우리의 로그 가능도 함수에 lower bound를 나타내게 된다.

이때 Q를 무엇으로 선택하는 것이 좋을까? inequality를 equality로 만들게 할수록 좋은 선택이 될것이다.

위에서 jensen's inequality에 대해 배울 때 등호가 성립할 때는 다음과 같은 경우라고 알아두었다.

$E[f(x)] = f[EX] \iff$ x is constant ($x = E[x]$ with probability 1)

따라서 우리는 다음 식을 만족해야 한다.

$$\frac{p(x,z;\theta)}{Q(z)} = c$$

이때 상수가 된다는 것은 분모와 분자의 비율을 의미하는 것이기 때문에 아래와 같이도 나타낼 수 있다.

$$Q(z) \propto p(x,z;\theta).$$

이때 Q는 분포이기 때문에 합이 1이라는 점을 적용하면 다음과 같이 나오고

$$\begin{aligned} Q(z) &= \frac{p(x,z;\theta)}{\sum_z p(x,z;\theta)} \\ &= \frac{p(x,z;\theta)}{p(x;\theta)} \\ &= p(z|x;\theta) \end{aligned} \tag{8}$$

그 결과 Q는 z|x의 사후분포임을 알게 되었다.

이는 다음과 같이 직접적으로도 증명할 수 있다.

$$\begin{aligned}\sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)} &= \sum_z p(z|x;\theta) \log \frac{p(x,z;\theta)}{p(z|x;\theta)} \\ &= \sum_z p(z|x;\theta) \log \frac{p(z|x;\theta)p(x;\theta)}{p(z|x;\theta)} \\ &= \sum_z p(z|x;\theta) \log p(x;\theta) \\ &= \log p(x;\theta) \sum_z p(z|x;\theta) \\ &= \log p(x;\theta) \qquad (\text{because } \textstyle\sum_z p(z|x;\theta) = 1)\end{aligned}$$

편의를 위해 (7)을 evidence lower bound(ELBO)라고 부를 것이다.

$$\text{ELBO}(x;Q,\theta) = \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)} \tag{9}$$

모든 example에 대해 합을 취하면 log likelihood 에 대한 lower bound를 얻을 수 있다.

$$\begin{aligned}\ell(\theta) &\geq \sum_i \text{ELBO}(x^{(i)};Q_i,\theta) \\ &= \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\end{aligned} \tag{11}$$

그 결과 log likelihood의 lower bound를 maximize하는 방식으로 MLE를 구할 수 있을 것이고 이에 EM algorithm을 적용하면 다음과 같다.

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\begin{aligned}\theta &:= \arg\max_\theta \sum_{i=1}^n \text{ELBO}(x^{(i)};Q_i,\theta) \\ &= \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}.\end{aligned} \tag{12}$$

}

k시리즈는 지윤이에게,, 있어요,,, 응교수님은 강의에서 여기까지 설명함,,,,,,,!