情報意味論（8） EM

櫻井彰人
慶應義塾大学理工学部

目次

EM の導入の動機（？）

- 動機（？）
  - 観測できないが、結果に関与している変数（属性）があるとき、（この変数を含む）パラメータの最尤推定をしたい。どうしたらよいか？
    - （パラメータ以外）すべて観測可能であれば、式は書ける
- 経験（？）
  - k-means クラスタリング

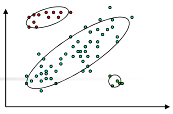
少し復習: 最尤推定

- 既知確率分布 $p(x;\theta)$ の独立サンプル $x_1,\ldots,x_N$ があるとき、パラメータ θ を推定する方法の一つ
- 尤度 $\Pi_i p(x_i;\theta)$ を最大にする θ を推定量とする

非観測変数があるときの最尤推定

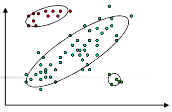
- 既知確率分布 $p(x,z;\theta)$ の独立サンプル $<x_1,z_1>, \ldots, <x_N,z_N>$ があるとき、パラメータ θ を推定したい。ただし、x は計測されてデータがあるが、z は計測されていない。
- 尤度 $\Pi_i p(x_i, z_i;\theta)$ を最大にする θ を推定量とすればよい、と思う。
- しかし、z_i が変量のまま残っているので、θ に関し尤度最大化することができない。

例: クラスタリング



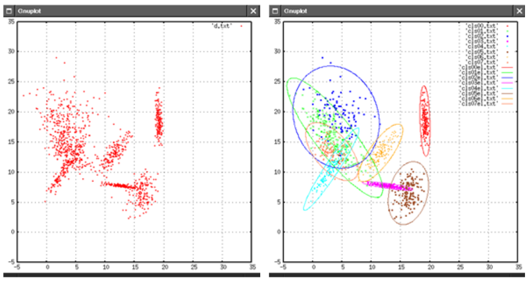
- クラスタリングは、n次元データを、クラスタに分けること。クラスタに分けるとは、
 データには、それが属するグループがあると仮定して
  - グループの発見と
  - 各データが属するグループの発見
 という二つの作業をすること
- 各データの座標（ n次元）を観測データ x_i 、属するグループを未観測データ z_i として、各グループの分布のパラメータ θ を推定することと考えることができる。

k-means 法



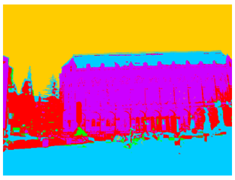
- クラスタリング方法の一つ
- 次の繰り返し
  - z_i が同じ $(=j)$ $<x_i, z_i>$ を集め、各 z_i $(=j)$ ごとその x_i を用いて θ_j を最尤推定する
  - θ_j $(j=1,\ldots,k)$ を用いて、z_i を最尤推定する。
- 結構うまくいく
- これが使えないか？

(EM)クラスタリング例



http://jormungand.net/projects/em/

(K-means) クラスタリング例



http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster/

背景

- EMアルゴリズムと名付けられて紹介されたのは、1977年の Dempster, Nan Laird, Donald Rubin による論文 Maximum Likelihood from Incomplete Data via the EM Algorithm においてである.
- 著者によれば "The EM algorithm has been proposed many times in special circumstances."
- EM は非観測量があるとき最尤推定量を求める方法である.
- EMアルゴリズムは、あるモデルのパラメータを、次の繰り返しで計算する.
  - 初期値を何等かの方法で定める。
  - 一回の計算は
    - E step - Expectation step
    - M step - Maximization step

Dempster, A.P. Laird, N.M. Rubin, D.B. (1977). "**Maximum Likelihood from Incomplete Data via the EM Algorithm**". *Journal of the Royal Statistical Society, Series B (Methodological)* **39** (1): 1–38.

応用

- 欠測値を補う
- 潜在変数の値を推定する
  - 隠れマルコフモデルのパラメータ推定
  - 有限混合モデルのパラメータ推定
  - クラスタリング
  - 半教師あり学習.

目次

- 動機と問題設定
- 簡単な例
- ちょっと複雑な例 – ガウス混合分布
- K-meansからのアプローチ
- EMアルゴリズム: 性質とまとめ

簡単な例

あるクラスでの成績分布を考える。
事象A ＝ Aをとる　P(A)=1/2
事象B ＝ Bをとる　P(B)=μ/4
事象C ＝ Cをとる　P(C)=1/2-μ/2
事象D ＝ Dをとる　P(D)=μ/4
（ただし、0≤μ≤1）

パラメータ μ をデータから推定したい。
Aは a人、Bは b人、Cは c人、Dは d人いたとする。

a, b, c, d が与えられた時、μ を最尤推定しよう

Dempster et al 1977 の例題を簡単にしたもの

簡単な計算

P(A)=1/2
P(B)=μ/4
P(C)=1/2-μ/2
P(D)=μ/4

P(A)=1/2　P(B)=μ/4　P(C)= 1/2–μ/2　P(D)=μ/4

$P(a,b,c,d|\,\mu) = C\,(1/2)^a(\mu/4)^b(1/2-\mu/2)^c(\mu/4)^d$　ただし、$C=\frac{(a+b+c+d)!}{a!b!c!d!}$

log P(a,b,c,d| μ) =
　a log(1/2) + b log(μ/4) + c log(1/2–μ/2) + d log(μ/4) + log C
これを $L(\mu)$ としよう

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{b}{\mu} - \frac{c}{1-\mu} + \frac{d}{\mu} = 0$$

最尤推定量 $\hat{\mu}$ は $\hat{\mu} = \frac{b+d}{b+c+d}$

隠れ変数がある場合

P(A)=1/2
P(B)=μ/4
P(C)=1/2-μ/2
P(D)=μ/4

仮に、Aを取った人とBを取った人は、合計 u 人、Cをとった人は c 人、Dをとった人は d 人であるとわかったとしよう。
μの最尤推定量は何であろうか？

この場合、不完全データ (u,c,d) を観測していることになる。
完全データの対数尤度は前ページと同じであり、最尤推定量は

$$\mu = (b+d)/(b+c+d)$$

しかし、b は可観測ではないので、上記問題には適用できない
EMアルゴリズムは、これに、次のように対処する

手順は、次の通り。

1. 初期設定。
   パラメータμの値を適宜決める。
   以下を繰り返す
2. パラメータμの次の値を決める準備をする（ただし、非観測変数Bの期待値を求めることになる）
   パラメータμの現在値を用いて、Bの分布を求める。これを用いて、対数尤度の（Bの分布に基づく）期待値を求める。
   なお、この過程でBの期待値を求めることになる。
3. パラメータμの値を決める。
   上記「対数尤度の期待値」を最大化する、パラメータμの値を求める。
   なお、この時、上記「Bの期待値」を用いることになる。

ステップ2

P(A)=1/2
P(B)=μ/4
P(C)=1/2-μ/2
P(D)=μ/4

確率変数Bは、サンプルサイズ u の二項分布をしていると考えることができる。そのパラメータは、$(\mu_k/4)/(1/2+\mu_k/4)$
つまり、Bの条件付期待値は、$E\{B|h\} = u\,(\mu_k/4)/(1/2+\mu_k/4)$
従って、$b_k = u\,(\mu_k/4)/(1/2+\mu_k/4)$
また、$a_k = u - u\,(\mu_k/4)/(1/2+\mu_k/4) = u\,(1/2)/(1/2+\mu_k/4)$

より正確には

ステップ2に現れる期待尤度を $Q(\mu;\,\mu_k)$ と書くことにする。すなわち
$Q(\mu;\mu_k) = E_B\{L(\mu)|\mu_k,u,c,d\}$　　（$L(\mu)$ は対数尤度）
さて、
$L(\mu) = a\log(1/2) + b\log(\mu/4) + c\log(1/2-\mu/2) + d\log(\mu/4) + \log C$
である。なお、$a = u - b, b$ の分布がμ_kに依存している（ c, d は定数）

ステップ2（続）

従って、$Q(\mu;\mu_k) = E_B\{L(\mu)|\mu_k,u,c,d\}$ は、a, b の期待値を
$a_k = E_B\{a|\mu_k,u,c,d\}, b_k = E_B\{b|\mu_k,u,c,d\}$ とすれば

$$a_k\log(1/2) + b_k\log(\mu/4) + c\log(1/2-\mu/2) + d\log(\mu/4)$$
$$+E_B\{\log C\,|\mu_k,u,c,d\}$$

なお、$b_k = u\,(\mu_k/4)/(1/2+\mu_k/4)\,, a_k = u\,(1/2)/(1/2+\mu_k/4)$
である。それは、次から得られる

$$P(a,b,c,d;\mu) = \frac{(a+b+c+d)!}{a!b!c!d!}\left(\frac{1}{2}\right)^a\left(\frac{\mu}{4}\right)^b\left(\frac{1}{2}-\frac{\mu}{2}\right)^c\left(\frac{\mu}{4}\right)^d$$
$$= \frac{u!}{(u-b)!b!}\left(\frac{1}{2}\right)^{u-b}\left(\frac{\mu}{4}\right)^b\frac{(u+c+d)!}{u!c!d!}\left(\frac{1}{2}-\frac{\mu}{2}\right)^c\left(\frac{\mu}{4}\right)^d$$

より

$$P(b;\mu_k,u,c,d) = \frac{u!}{(u-b)!b!}\left(\frac{1}{2}\right)^{u-b}\left(\frac{\mu_k}{4}\right)^b\Big/\left(\frac{1}{2}+\frac{\mu_k}{4}\right)^u$$

ステップ3

ステップ3

$Q(\mu;\mu_k)$ を最大化するμ は

$$\mu_{k+1} = (b_k + d)/(b_k + c + d)$$

P(A)=1/2
P(B)=μ/4
P(C)=1/2-μ/2
P(D)=μ/4

結果を書き直せば

Expectation step
仮にμの値を知っているなら、a と b の期待値を計算することができる。

$$a \leftarrow \frac{1/2}{1/2+\mu/4}u,\ b \leftarrow \frac{\mu/4}{1/2+\mu/4}u$$

Maximization step
最尤推定量を計算することができる。

$$\mu \leftarrow \frac{b+d}{b+c+d}$$

計算してみると

P(A)=1/2
P(B)=μ/4
P(C)=1/2-μ/2
P(D)=μ/4

```
u <- 25
c <- 10
d <- 10
mu <- 0

for ( i in 1:8 ) {
  b <- (mu/4)*u/(1/2+mu/4)
  mu <- (b+d)/(b+c+d)
  print( c(b,mu) )
}
```

```
[1] 0.0       0.5
[1] 5.0       0.6
[1] 5.7692308 0.6119403
[1] 5.8571429 0.6132597
[1] 5.8668076 0.6134042
[1] 5.867866  0.613420
[1] 5.8679813 0.6134217
[1] 5.8679939 0.6134219
```

目次

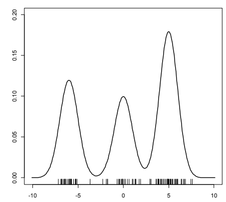
- 動機と問題設定
- 簡単な例
- ちょっと複雑な例 – ガウス混合分布
- K-meansからのアプローチ
- EMアルゴリズム: 性質とまとめ

より複雑なモデル

- 確率モデルであって、一個の著名(?)な分布で表せないもの、… で表せそうもないもの、…ではなさそうなものが、世の中にはたくさんある。
  - 例えば、多峰分布

例: 混合正規分布

- 正規分布(ガウス混合)の線形和



線形和(重みの和は1)
$$p(x) = \sum \pi_j\, p_j(x)$$

正規分布の線形和であるなら
$$p_j(x) = N(x;\ \mu_j, \sigma_j)$$
として、
$$p(x) = \sum \pi_j\, N(x;\ \mu_j, \sigma_j)$$

問題: パラメータが推定できない

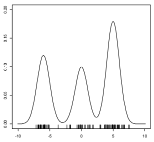
データが一個の正規分布から生成されているなら、そのパラメータ（平均と分散）の推定は容易である。例えば、平均値の最尤推定量は

$$\mu_{\mathrm{ML}} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{m} (x_i - \mu) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

しかし、混合分布の場合、最尤推定をしようと思うと、次の最大化問題を解かなければいけない（簡単にするため標準偏差は既知）。

$$\mu_{\mathrm{ML}} = \underset{(\mu_1,\mu_2,\mu_3)}{\operatorname{argmax}} \prod_{i=1}^{m} \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$

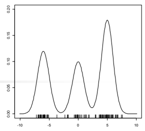
$$= \underset{(\mu_1,\mu_2,\mu_3)}{\arg\max} \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$



これは解けない

しかし近似計算なら

できるかもしれない。続けてみよう。



$$\theta_{\mathrm{ML}} = \underset{\mu_j,\pi_j}{\arg\max}\, LL(\mu_1,\ldots,\pi_1,\ldots)$$

$$LL(\mu_1,\ldots,\pi_1,\ldots) = \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$

とりあえず、停留点が求まるかどうか、Lagrange関数を微分してみよう

Lagrange関数は $L = \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j) + \lambda(1 - \sum_{i=1}^{3} \pi_j)$

$$\frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{m} \frac{N(x_i;\mu_j)}{\sum_{j=1}^{3} \pi_j N(x_i;\mu_j)} - \lambda$$

$$= \sum_{i=1}^{m} \tau_i^j \big/ \pi_j - \lambda$$

$$\frac{\partial L}{\partial \mu_j} = \sum_{i=1}^{m} \frac{\pi_j N(x_i;\mu_j)}{\sum_{j=1}^{3} \pi_j N(x_i;\mu_j)} \frac{\partial}{\partial \mu_j}\left\{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\right\}$$

$$= \sum_{i=1}^{m} \tau_i^j \Sigma_j^{-1}(x_i - \mu_j)$$

$$\tau_i^j = p(z_i = j \mid x_i, \theta) = \frac{p(x_i, z_i = j \mid \theta)}{p(x_i \mid \theta)} = \frac{p(x_i \mid z_i = j, \theta)\, p(z_i = j \mid \theta)}{p(x_i \mid \theta)} = \frac{\pi_j N(x_i;\mu_j)}{\sum_j \pi_j N(x_i;\mu_j)}$$

方程式

$$L = \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j) + \lambda(1 - \sum_{i=1}^{3} \pi_j) \qquad \tau_i^j = \frac{\pi_j N(x_i;\mu_j)}{\sum_j \pi_j N(x_i;\mu_j)}$$

$$\frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{m} \tau_i^j \big/ \pi_j - \lambda = 0,\ \sum_{i=1}^{3} \pi_j = 1,\ \sum_{j=1}^{3}\sum_{i=1}^{m} \tau_i^j = m \quad \text{より} \quad \pi_j = \frac{1}{m}\sum_{i=1}^{m} \tau_i^j$$

$$\frac{\partial L}{\partial \mu_j} = \sum_{i=1}^{m} \tau_i^j \Sigma_j^{-1}(x_i - \mu_j) = 0 \quad \text{より} \quad \mu_j = \frac{\sum_{i=1}^{m} \tau_i^j x_i}{\sum_{i=1}^{m} \tau_i^j}$$

非線形連立方程式だが、これは解けない。
ヒューリスティックスにより、下記のようにすればよさそうだが、果して収束するのだろうか

$$\tau_i^j \leftarrow \frac{\pi_j N(x_i;\mu_j)}{\sum_j \pi_j N(x_i;\mu_j)} \qquad \pi_j \leftarrow \frac{1}{m}\sum_{i=1}^{m} \tau_i^j \qquad \mu_j \leftarrow \frac{\sum_{i=1}^{m} \tau_i^j x_i}{\sum_{i=1}^{m} \tau_i^j}$$

参考: EMとの対応

勿論、対応するのですが、それは後講釈

$$\theta_{\mathrm{ML}} = \underset{\mu_j,\pi_j}{\arg\max}\, LL(\mu_1,\ldots,\pi_1,\ldots)$$

$$LL(\mu_1,\ldots,\pi_1,\ldots) = \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$

Lagrange関数は $L = \sum_{i=1}^{m} \log \sum_{j=1}^{3} \pi_j N(x_i;\mu_j) + \lambda(1 - \sum_{i=1}^{3} \pi_j)$

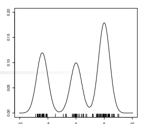
$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} \tau_i^j x_i}{\sum_{i=1}^{m} \tau_i^j} \quad \Longleftrightarrow \quad \tau_i^j \leftarrow \frac{\pi_j N(x_i;\mu_j)}{\sum_j \pi_j N(x_i;\mu_j)}$$

$$\pi_j \leftarrow \frac{1}{m}\sum_{i=1}^{m} \tau_i^j$$

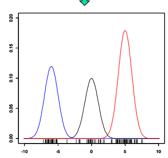
M step

E step

EM 一般的な定義



- X={x$_1$,…,x$_N$} 観測データ
- Z={z$_1$,…,z$_N$} 非観測データ（隠れ変数）
  - Y=X∪Z
- h: 分布のパラメータ（θ とも）



注: 混合分布のときの考え方

$$P(x;\mu) = \prod_{i=1}^{m} \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$

$$P(x,z;\mu) = \prod_{i=1}^{m} \prod_{j=1}^{3} \left(\pi_j N(x_i;\mu_j)\right)^{z_{i,j}} = \prod_{i=1}^{m} \pi_{z_i} N(x_i;\mu_{z_i})$$

$z_{i,j} = 1$ or 0. $z_{i,j} = 1$ iff x_i はクラスタ j に属する
$z_i = j$ iff x_i はクラスタ j に属する

EM 一般的な定義 (続)

X={x$_1$,…,x$_N$} 観測データ
Z={z$_1$,…,z$_N$} 非観測データ
Y=X∪Z

- E-Step: 次の仮説 h′ の対数尤度の期待値を求める（式で表す）. ただし, 現在の仮説 h と観測データ X は既知とする（目標: ln P(X | h) の最大化であった）

Y に関する期待値

パラメータはフリー

このパラメータ値で、Zの分布を決める

Q(h’ | h) = E[ln P(Y | h’) | h, X]

対数尤度の期待値を求める

=∫ (ln P(X, z | h’)) P(z | h, X) dz

- M-Step: Q を最大化する h’ を次の h とする
  h ← argmax$_{h'}$ Q(h’ | h)

h を決めるときに決めているのはクラスタのパラメタ
Q を決めるときに決めているのはクラスタのメンバー

EM E-step

$$
\begin{aligned}
Q(h'|h) &= E[\ln P(Y|h')|h,X] \\
&= E[\ln \prod_{i=1}^{N} P(y_i|h')|h,X] \\
&= E[\sum_{i=1}^{N} \ln P(y_i|h')|h,X] \\
&= E[\sum_{i=1}^{N}\left(\ln \prod_{j=1}^{k}\left(\pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)^{z_i^j}\right)|h,X] \\
&= E[\sum_{i=1}^{N}\left(\sum_{j=1}^{k} z_i^j \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)|h,X] \\
&= \sum_{i=1}^{N}\left(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)
\end{aligned}
$$

$$p(x|h) = \sum_j \pi_j N(x|\mu_j,\Sigma_j)$$
$$p(x, z=j|h) = \pi_j N(x|\mu_j,\Sigma_j)$$
$$p(x, z^1,..,z^k|h) = \prod_{j=1}^{k}\left(\pi_j N(x|\mu_j,\Sigma_j)\right)^{z^j}$$
$$E[z_i^j|x_i,h] = p(z_i^j=1|x_i,h) = \frac{\pi_j N(x_i|\mu_j,\Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j,\Sigma_j)}$$

EM M-Step

$$
\begin{aligned}
h &\leftarrow \arg\max_{h'} Q(h'|h) \\
&= \operatorname*{argmax}_{h'} \sum_{i=1}^{N}\left(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)
\end{aligned}
$$

h は π, μ, Σ の組、h' は π′, μ', Σ′ の組である.
最小化は π′, μ', Σ′ による偏微分が0とおいて達成できる,

π'$_j$ の推定

- 変数 π'_j に関する $Q(h'|h)$ の偏微分（言い忘れたが、Lagrange関数を用いている）

$$
\begin{aligned}
&\frac{\partial}{\partial \pi'_j}\{Q(h'|h) + \lambda(1 - \sum_j \pi'_j)\} \\
&= \frac{\partial}{\partial \pi'_j}\{\sum_{i=1}^{N}(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j))\} - \lambda \\
&= \sum_{i=1}^{N} \frac{E[z_i^j|h,X]}{\pi'_j} - \lambda
\end{aligned}
$$

$$E[z_i^j|x_i,h] = p(z_i^j=1|x_i,h) = \frac{\pi_j N(x_i|\mu_j,\Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j,\Sigma_j)}$$

- これらを 0 とおく方程式をとけば

$$\pi'_j = \frac{\sum_i E[z_i^j|h,X]}{N} = \frac{\sum_i \tau_i^j}{N}$$

$$\tau_i^j = \frac{\pi_j N(x_i|\mu_j,\Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j,\Sigma_j)}$$

μ'$_j$ の推定

- 変数 μ'_j に関する $Q(h'|h)$ の偏微分

$$
\begin{aligned}
\frac{\partial Q(h'|h)}{\partial \mu'_j} &= \frac{\partial}{\partial \mu'_j}\{\sum_{i=1}^{N}(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j))\} \\
&= \sum_{i=1}^{N} E[z_i^j|h,X] \frac{\partial}{\partial \mu'_j} \ln N(x_i|\mu'_j,\Sigma'_j) \\
&= \sum_i \tau_i^j \Sigma_j^{-1}(x_i - \mu'_j)
\end{aligned}
$$

- これを 0 とおけば、次式が得られる

$$\mu'_j = \frac{\sum_i \tau_i^j x_i}{\sum_i \tau_i^j}$$

Σ'$_j$ の推定

- 変数 Σ_j に関する l の偏微分

$$
\begin{aligned}
\frac{\partial Q(h'|h)}{\partial \Sigma'_j} &= \frac{\partial}{\partial \Sigma'_j}\{\sum_{i=1}^{N}(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j))\} \\
&= \sum_{i=1}^{N} E[z_i^j|h,X] \frac{\partial}{\partial \Sigma'_j} \ln N(x_i|\mu'_j,\Sigma'_j) \\
&= \sum_{i=1}^{N} \tau_i^j \{-\frac{1}{2}\Sigma_j'^{-1} + \frac{1}{2}\Sigma_j'^{-1}(x_i - \mu'_j)(x_i - \mu'_j)^T \Sigma_j'^{-1}\}
\end{aligned}
$$

- これを 0 とおくと次式が得られる

$$\Sigma'_j = \frac{\sum_i \tau_i^j (x_i - \mu'_j)(x_i - \mu'_j)^T}{\sum_i \tau_i^j}$$

EM 混合正規分布

E-Step:

$$Q(h'|h) \leftarrow \sum_{i=1}^{N}\left(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)$$

すなわち

$$E[z_i^j|x_i,h] = p(z_i^j=1|x_i,h) = \frac{\pi_j N(x_i|\mu_j,\Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j,\Sigma_j)} \to \tau_i^j$$

M-Step:

$$h \leftarrow \operatorname*{argmax}_{h'} \sum_{i=1}^{N}\left(\sum_{j=1}^{k} E[z_i^j|h,X] \ln \pi'_j N(x_i|\mu'_j,\Sigma'_j)\right)$$

すなわち

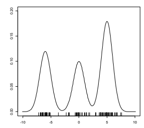
$$\pi'_j \leftarrow \frac{\sum_i \tau_i^j}{N} \qquad \mu_j \leftarrow \frac{\sum_{i=1}^{N} \tau_i^j x_i}{\sum_{i=1}^{N} \tau_i^j} \qquad \Sigma'_j \leftarrow \frac{\sum_i \tau_i^j (x_i - \mu'_j)(x_i - \mu'_j)^T}{\sum_i \tau_i^j}$$

目次

- 動機と問題設定
- 簡単な例
- ちょっと複雑な例 – ガウス混合分布
- K-meansからのアプローチ
- EMアルゴリズム: 性質とまとめ

分布推定は「教師なし学習」

- 教師付き学習: データ <x, z>
- 教師なし学習: データ x

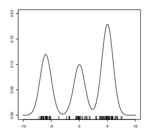


補足: 教師なし学習が必要となるところ

- 分布関数（確率密度関数）の推定
- クラスタリング
- 外れ値/新規点の検出
- データ圧縮
- 可視化

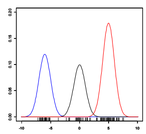
分布推定とクラスタリング

- クラスタリング: 混合分布から生成されたデータに対し、どの分布から生成されたかを推定する



混合分布

$p(x) = \sum \pi_j p_j(x)$



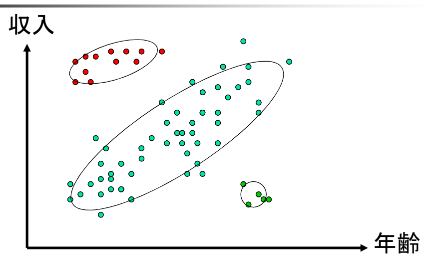
混合分布

$p(x,z) = \prod (\pi_j p_j(x))^{z_j}$

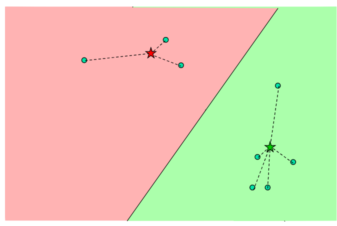
各クラスタは混合分布の個々の分布に対応すると考える

- 隠れ変数: データ点がどのガウス分布から生成されたか
- すなわち, 観測データ <x>, 全データ <x, z>.
- 課題: <x> から <x, z> を推定する

クラスタリング/密度推定付き



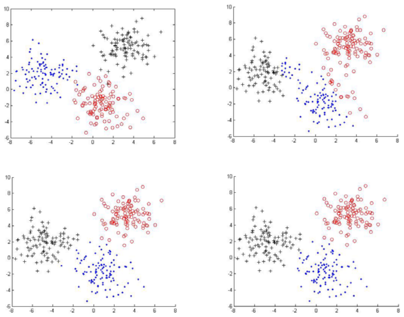
ある方法: k-means クラスタリング



K-means クラスタリング例

真のデータ 第一回

第二回 第三回



K-means の行っていること

- 前提(「動作だけ」を記述するには不要)
  - (各正規分布の)分散共分散は同じとする
  - 分散共分散行列は、対角かつ各軸で等分散とする
- 初期値
  - クラスタ中心 o_j をランダムに定め、推定を開始する
- 繰り返し
  - 分類: 各観測点ごと、その(産みの親である)クラスタを推定する 各クラスターのメンバーを推定するといってもよい
    - 各 <x> → <x, j>, ただし $j = \arg\min | x - o_j |$.
    - i.e. 最近傍のクラスタ中心を選び、そのクラスタ番号を j とする
  - 中心の再設定: クラスタごと、同一クラスタの点のみを用いて、その重心(平均値)を新たにクラスタ中心とする
    - 各 j につき、o_i = center of { x | <x, j> }

K-means原理

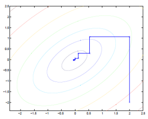
- ポテンシャル関数の最小化

$$\min_{\mu}\min_{C} F(\mu, C) = \min_{\mu}\min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- 次の2つのステップからなる
- 分類: C に関する$F(\mu, C)$の最小化
  - C のメンバーを決める
- 中心の再設定: μ に関する$F(\mu, C)$の最小化
  - $F(\mu, C)$を最小化する μ を求める

Coordinate descent

- 座標軸降下法/座標降下法(?)
- $\min_a \min_b F(a, b)$ を求めたい
- Coordinate descent
  - a を固定し、b に関して最小化
  - b を固定し、a に関して最小化
- 収束する
  - もし、F が有界であれば。
  - 実際、結構良い局所最小値に。



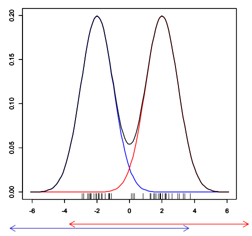
- K-means は coordinate descent アルゴリズムだ

K-means の欠点

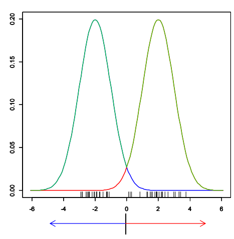
- Spherical な場合しか扱えない
  - 分散共分散行列が、σI (Iは単位行列)
- 各クラスタの重みが等しいときしか扱えない
- 混合正規分布から生成されたデータに適用すると、推定値(例えば、平均値)にbiasが発生する。

bias の発生

$p(x) = \sum \pi_j p_j(x)$



$p(x) = \max \pi_j p_j(x)$



K-means の欠点の解消に向けて

- 前2者への対応
  - 混合（多項）正規分布でモデル化する
  - 分類実行時に、生起確率が最大となるクラスタを選ぶ
    - K-means: 中心（=分布の中心=平均値）からの距離が最小のクラスタを選ぶ
- 後1者への対応
  - bias の原因は、「生起確率最大のクラスタを選ぶ」故、属しうるクラスタの確率に従い、複数のクラスタに属するとする

$$\min_{\mu}\min_{C} F(\mu, C) = \min_{\mu}\min_{C} \sum_{j=1}^{k} \sum_{i:C(i)=j} \|\mu_j - x_i\|^2$$

$$= \min_{\mu}\min_{C} \sum_{j=1}^{k} \sum_{i:C(i)=j} -\log\exp\left(-\frac{\|x_i - \mu_j\|^2}{2}\right)$$

$$= \max_{\mu}\max_{C} \sum_{j=1}^{k} \sum_{i:C(i)=j} \log N(x_i; \mu_j, 1)$$

$$\Rightarrow \max_{\mu}\max_{z} \sum_{j=1}^{k} \sum_{i:z_i=j} \log \pi_j N(x_i; \mu_j, \Sigma_j) \Rightarrow \max_{\mu}\max_{z} \sum_{j=1}^{k} \sum_{i:z_i=j} \log \pi_j P(x_i; \theta_j)$$

課題: z_i をどう推定するか. 次はうまくない

$z_i = j$ iff $j = \arg\max \pi_j P(x_i; \theta_j)$

⇒ そこで、EMへ

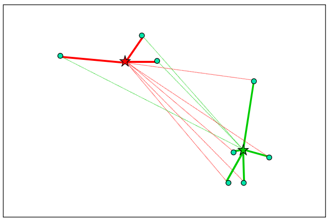
EM と k-means との対応

- E-Step: 非観測データは期待値を推定
  Q(h' | h)
  = E[ln P(Y | h') | h, X]
  =∫ (ln P(X, z | h')) P(z | h, X) dz
- M-Step: Q を最大化する h' を次の h とする. 最尤推定
  h ← argmax$_{h'}$ Q(h' | h)

- 分類: C に関する$F(\mu, C)$の最小化
  - C のメンバーを決める
- 中心の再設定: μ に関する $F(\mu, C)$の最小化
  - $F(\mu, C)$を最小化する μ を求める

（coordinate descent）

K-means をソフトにしたイメージ

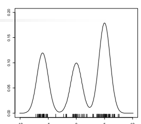


目次

- 動機と問題設定
- 簡単な例
- ちょっと複雑な例 – ガウス混合分布
- K-meansからのアプローチ
- EMアルゴリズム: 性質とまとめ

EMアルゴリズムの性質

$L(X; \theta) = \log(X = \{X_i\}_{i=1}^{n}$ の結合確率$; \theta)$



定理

θ_k をEMアルゴリズムでられる k 番目のパラメータθとする。この時、$L(X; \theta_{k+1}) \geq L(X; \theta_k)$ が成立する。また、適当な条件のもと、θ_k は、最尤推定量 $\arg\max L(X; \theta)$ に収束する。

EM まとめ1

- $X=\{x_1,\ldots,x_N\}$ 観測データ
- $Z=\{z_1,\ldots,z_N\}$ 非観測データ (隠れ変数)
  - $Y=X \cup Z$
- h: 分布のパラメータ(θ とも)

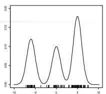
- 次を繰り返す
- E-Step: 非観測データは期待値を推定

  $Q(h' \mid h) = E[\ln P(Y \mid h') \mid h, X]$

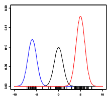
  $= \int (\ln P(X, z \mid h')) P(z \mid h, X)\, dz$
- M-Step: Q を最大化する h' を次の h とする. 最尤推定

  $h \leftarrow \mathrm{argmax}_{h'} Q(h' \mid h)$

EM まとめ2



- 混合分布の推定に用いる
  - 生成された元の分布を表す非観測変数を導入
  - EMを適用
    - 結果はソフトクラスタリングみたい



- クラスタリングに適用
  - 混合分布の推定として定式化
  - 結果中に、各サンプルのクラスタへの所属確率
  - サンプルを生成する事後確率が最大のクラスタを、それが属するクラスタとする