EM アルゴリズム

櫻井彰人

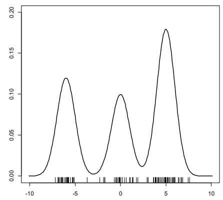
目次

動機: 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

動機: 多峰分布

- 正規分布（ガウス分布）の線形和



線形和（重みの和は1）

$\boldsymbol{p}(\boldsymbol{x}) = \Sigma\, \pi_j\, \boldsymbol{p}_j(\boldsymbol{x})$

正規分布の線形和であるなら

$\boldsymbol{p}_j(\boldsymbol{x}) = N(\boldsymbol{x};\, \mu_j, \sigma_j)$

として、

$\boldsymbol{p}(\boldsymbol{x}) = \Sigma\, \pi_j\, N(\boldsymbol{x};\, \mu_j, \sigma_j)$

課題: パラメータ推定

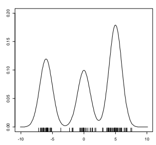
データが一個の正規分布から生成されているなら、そのパラメータ（平均と分散）の推定は容易である。例えば、平均値の最尤推定量は

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

しかし、多峰分布の場合、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）。

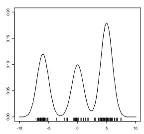
$$\mu_{\mathrm{ML}} = \operatorname*{argmax}_{(\mu_1,\mu_2,\mu_3)} \prod_{i=1}^{m} \sum_{j=1}^{3} \pi_j N(x_i;\mu_j)$$

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



解けるか？

できるかもしれない。続けてみよう。



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

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

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

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

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

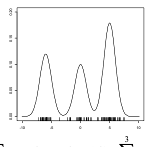
$$= \sum_{i=1}^{m} \frac{\pi_j N(x_i;\mu_j)}{\sum_{j=1}^{3} \pi_j N(x_i;\mu_j)} \Sigma_j^{-1} (x_i - \mu_j)$$

方程式は

$$0=\sum_{i=1}^{m}\frac{N(x_i;\mu_j)}{\sum_{j=1}^{3}\pi_j N(x_i;\mu_j)}-\lambda$$

$$0=\sum_{i=1}^{m}\frac{\pi_j N(x_i;\mu_j)}{\sum_{j=1}^{3}\pi_j N(x_i;\mu_j)}\Sigma_j^{-1}(x_i-\mu_j)$$

となる。とても解けるとは思えない



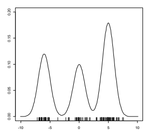
$$L=\sum_{i=1}^{m}\log\sum_{j=1}^{3}\pi_j N(x_i;\mu_j)+\lambda(1-\sum_{i=1}^{3}\pi_j)$$

ここで、通常は、EMアルゴリズムに行く(e.g. Elements of Statistical Learning, 2nd ed. p.274)

ちょっと工夫をしてみよう。次のスライドで。

$$\boxed{\tau_i^j=\frac{\pi_j N(x_i;\mu_j)}{\sum_j \pi_j N(x_i;\mu_j)}}$$

と置く



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

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

停留点を求めよう

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

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

解けているように見えるが、解けていない！
(ほぼ解いているともいえる。以下のスライドで)

補足

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

代入は、一斉に行うとする。
簡単にわかるが、どの変数値もある有界範囲内に入る。
$\pi_j\geq 0$ は和が1、μ_j は任意の値に対して、次の τ,π_j,μ_j が計算できる。
従って、上記式が定義する π_j,μ_j の変換(前頁)は、ユークリッド空間内のコンパクト凸部分集合からその上/中への連続写像となる。ブラウアーの不動点定理から、不動点が存在する。
また、ある初期点から初めて、上式により点列を生成すれば、当該コンパクト凸集合内に、それは集積点をもつ。もし一点に収束する列があるなら、当該集積点は、(方程式と考えた時の)解となる。

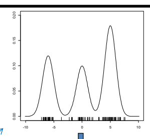
どうも、気持ちが悪い。もっとすっきりできないか。
縮小写像であればよいのだが、不明。

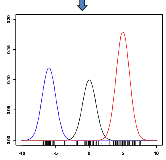
目次

- 多峰分布の最尤推定
- EMアルゴリズム

多峰分布に対し

- 混合(正規)分布と考える
  - 要素分布をサンプル毎に選択





隠れ変数(latent variable)を導入する

$$P(X;\mu,\pi)=\sum_{j=1}^{3}\pi_j N(X;\mu_j)$$

$$\Downarrow$$

$$P(X,Z;\mu,\pi)=P(X\mid Z;\mu,\pi)P(Z;\mu,\pi)$$
$$=N(X;\mu_Z)P(Z;\pi)$$

$$P(X;\mu,\pi)=\sum_{j=1}^{3}N(X;\mu_j)P(Z=j;\pi)$$

$$\sum_Z P(X,Z;\mu,\pi)=\sum_z N(X;\mu_Z)P(Z;\pi)=\sum_{j=1}^{3}N(X;\mu_j)P(Z=j;\pi)=\sum_{j=1}^{3}N(X;\mu_j)\pi_j$$

何をしたいか

- パラメータ推定
- 尤度を最大とするパラメータ値を求める

$$P(X;\mu,\pi)=\sum_{j=1}^{3}N(X;\mu_j)P(Z=j;\pi)$$

$$P(x_1,\dots x_m;\mu,\pi)=\prod_{i=1}^{m}\sum_{j=1}^{3}N(x_i;\mu_j)P(Z_i=j;\pi)$$
$$=\prod_{i=1}^{m}\sum_{j=1}^{3}P(x_i,Z_i=j;\mu,\pi)$$

式を短くするために
こちらで変形します

$$LL(\mu,\pi;x_1,\dots x_m)=\log\prod_{i=1}^{m}\sum_{j=1}^{3}P(x_i,Z_i=j;\mu,\pi)$$
$$=\sum_{i=1}^{m}\log\sum_{j=1}^{3}P(x_i,Z_i=j;\mu,\pi)$$

$$
\begin{aligned}
&LL(\mu,\pi;x_1,\ldots x_m)\\
&=\sum_{i=1}^{m}\log\sum_{j=1}^{3}P(x_i,Z_i=j;\mu,\pi)\\
&=\sum_{i=1}^{m}\log\sum_{j=1}^{3}P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})\frac{P(x_i,Z_i=j;\mu,\pi)}{P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})}
\end{aligned}
$$

$\mathrm{Note}:\sum_j P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})=1$

$$
\geq\sum_{i=1}^{m}\sum_{j=1}^{3}P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})\log\frac{P(x_i,Z_i=j;\mu,\pi)}{P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})}
$$

$\mathrm{Note: the\ equality\ holds\ when}\ \mu=\mu^{(t)},\pi=\pi^{(t)}$

$$
\begin{aligned}
&=\sum_{i=1}^{m}E_{Z_i\mid x_i;\mu^{(t)},\pi^{(t)}}\left[P(x_i,Z_i=j;\mu,\pi)\right]+\sum_{i=1}^{m}H\left(P(Z_i\mid x_i;\mu^{(t)},\pi^{(t)})\right)\\
&=Q(\mu,\pi\mid\mu^{(t)},\pi^{(t)})+\sum_{i=1}^{m}H\left(P(Z_i\mid x_i;\mu^{(t)},\pi^{(t)})\right)
\end{aligned}
$$

⇨ E step

そこで、 $\mu^{(t+1)},\pi^{(t+1)}=\arg\max_{\mu,\pi}Q(\mu,\pi\mid\mu^{(t)},\pi^{(t)})$ ⇨ M step

---

$$
\mu^{(t+1)},\pi^{(t+1)}=\arg\max_{\mu,\pi}Q(\mu,\pi\mid\mu^{(t)},\pi^{(t)})
$$

と置いたので、

$$
\begin{aligned}
LL(\mu^{(t+1)},\pi^{(t+1)};x_1,\ldots x_m)&\geq Q(\mu^{(t+1)},\pi^{(t+1)}\mid\mu^{(t)},\pi^{(t)})+\sum_i H\left(P(Z_i\mid x_i;\mu^{(t)},\pi^{(t)})\right)\\
&\geq Q\left(\mu^{(t)},\pi^{(t)}\mid\mu^{(t)},\pi^{(t)}\right)+\sum_i H\left(P(Z_i\mid x_i;\mu^{(t)},\pi^{(t)})\right)\\
&=LL(\mu^{(t)},\pi^{(t)};x_1,\ldots x_m)
\end{aligned}
$$

$$
\begin{aligned}
\mu^{(t+1)},\pi^{(t+1)}&=\arg\max_{\mu,\pi}Q(\mu,\pi\mid\mu^{(t)},\pi^{(t)})\\
&=\arg\max_{\mu,\pi}\sum_{i=1}^{m}\sum_{j=1}^{3}P(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)})\log P(x_i,Z_i=j;\mu,\pi)
\end{aligned}
$$

混合正規分布（続）

$$
\mu^{(t+1)},\pi^{(t+1)}=\arg\max_{(\mu_\bullet,\pi_\bullet)}\sum_{i=1}^{m}\sum_{j=1}^{3}p\left(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)}\right)\log\pi_j N(x_i;\mu_j)
$$

ところで

$$
p\left(Z_i=j\mid x_i;\mu^{(t)},\pi^{(t)}\right)=\frac{\pi_j N(x_i\mid\mu_j^{(t)})}{\sum_j\pi_j N(x_i\mid\mu_j^{(t)})}
$$

ゆえ、

$$
\tau_{i,j}^{(t)}=\frac{\pi_j N(x_i\mid\mu_j^{(t)})}{\sum_j\pi_j N(x_i\mid\mu_j^{(t)})}
$$

とおくと、

$$
\mu^{(t+1)},\pi^{(t+1)}=\arg\max_{(\mu_\bullet,\pi_\bullet)}\sum_{i=1}^{m}\sum_{j=1}^{3}\tau_{i,j}^{(t)}\log\pi_j N(x_i;\mu_j)
$$

混合正規分布（続々）

$$
\mu^{(t+1)},\pi^{(t+1)}=\arg\max_{(\mu_\bullet,\pi_\bullet)}\sum_{i=1}^{m}\sum_{j=1}^{3}\tau_{i,j}^{(t)}\log\pi_j N(x_i;\mu_j)
$$

が問題であった。これを解こう

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

$$
\frac{\partial L}{\partial\pi_j}=\sum_{i=1}^{m}\tau_{i,j}^{(t)}\big/\pi_j-\lambda \qquad \frac{\partial L}{\partial\mu_j}=\sum_{i=1}^{m}\tau_{i,j}^{(t)}\Sigma_j^{-1}(x_i-\mu_j)
$$

$\sum_{i=1}^{m}\tau_{i,j}^{(t)}\big/\pi_j-\lambda=0$ $\sum_{i=1}^{3}\pi_j=1,\sum_{j=1}^{3}\sum_{i=1}^{m}\tau_{i,j}^{(t)}=m$ より $\pi_j=\frac{1}{m}\sum_{i=1}^{m}\tau_{i,j}^{(t)}$

$\sum_{i=1}^{m}\tau_{i,j}^{(t)}\Sigma_j^{-1}(x_i-\mu_j)=0$ より $\mu_j=\frac{\sum_{i=1}^{m}\tau_{i,j}^{(t)}x_i}{\sum_{i=1}^{m}\tau_{i,j}^{(t)}}$

混合正規分布（続々々）

$$
\tau_{i,j}^{(t)}=\frac{\pi_j N(x_i\mid\mu_j^{(t)})}{\sum_j\pi_j N(x_i\mid\mu_j^{(t)})}
$$

$$
\pi_j^{(t+1)}=\frac{1}{m}\sum_{i=1}^{m}\tau_{i,j}^{(t)}
$$

$$
\mu_j^{(t+1)}=\frac{\sum_{i=1}^{m}\tau_{i,j}^{(t)}x_i}{\sum_{i=1}^{m}\tau_{i,j}^{(t)}}
$$

できた！

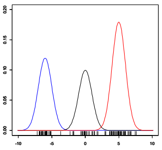
あれ？

$$
\tau_{i,j}^{(t+1)}=\frac{\pi_j N(x_i\mid\mu_j^{(t)})}{\sum_j\pi_j N(x_i\mid\mu_j^{(t)})} \qquad \tau_i^j=\frac{\pi_j N(x_i;\mu_j)}{\sum_j\pi_j N(x_i;\mu_j)}
$$

$$
\pi_j^{(t+1)}=\frac{1}{m}\sum_{i=1}^{m}\tau_{i,j}^{(t)} \qquad \text{VS.} \qquad \pi_j=\frac{1}{m}\sum_{i=1}^{m}\tau_i^j
$$

$$
\mu_j^{(t+1)}=\frac{\sum_{i=1}^{m}\tau_{i,j}^{(t)}x_i}{\sum_{i=1}^{m}\tau_{i,j}^{(t)}} \qquad \mu_j=\frac{\sum_{i=1}^{m}\tau_i^j x_i}{\sum_{i=1}^{m}\tau_i^j}
$$

EMアルゴリズム



目標

$$LL(\theta;\mathbf{x})=\sum_i \log p(x_i;\theta)=\sum_i\left(\log\sum_{z_i}p(x_i,Z_i=z_i;\theta)\right)$$

Eステップ

$$Q\left(\theta\mid\theta^{(t)}\right)=\boxed{\mathop{E}_{Z_\bullet|x_\bullet,\theta^{(t)}}\sum_i \log p(x_i,Z_i;\theta)}$$

とおき、　　　　対数尤度

Mステップ　$\theta^{(t+1)}=\arg\max_\theta Q\left(\theta\mid\theta^{(t)}\right)$

とすると、

$$LL(\theta^{(t+1)};x)=\sum_i \log p\left(x_i;\theta^{(n)}\right)\to 極大値$$

となる

$$\theta^{(t+1)}=\arg\max{}_\theta Q(\theta\mid\theta^{(t)})$$
$$=\arg\max{}_\theta \boxed{\sum_{i=1}^m\sum_j P(Z_i=j\mid x_i;\theta^{(t)})\log P(x_i,Z_i=j;\theta)}$$

別の簡単な例

まず、隠れ変数がない場合
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

全て可観測なら

事象	確率	観測数
A	1/2	a
B	μ/4	b
C	1/2−μ/2	c
D	μ/4	d

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、$C=\frac{(a+b+c+d)!}{a!b!c!d!}$

$\log P(a,b,c,d|\ \mu) =$
　$a\log(1/2) + b\log(\mu/4) + c\log(1/2-\mu/2) + d\log(\mu/4) + \log C$
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

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、次のように対処する
ただし、b を隠れ変数とする

事象	確率	観測数
A	1/2	u
B	μ/4	
C	1/2−μ/2	c
D	μ/4	d

事象	確率	観測数
A	1/2	u
B	μ/4	
C	1/2−μ/2	c
D	μ/4	d

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

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

b の分布を、現在のパラメータ $\mu^{(t)}$ を使って（可観測変数値も使って）表現する

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

$P\left(b;\mu^{(t)},u,c,d\right)$ は $\tau_{i,j}^{(t)}=p(z_i\left(\mu^{(t)}\right)=j)$ に相当

$b^{(t)}=E_{b|u,c,d,\mu^{(t)}}[b]=u\frac{\mu^{(t)}/4}{1/2+\mu^{(t)}/4}$ となり

$z_i^{(t)}=E_{z_i|\theta^{(t)}}[z_i]$ は定義しなかった。ほぼ意味がないためである

更新式

$$Q\left(\mu|\mu^{(t)}\right)$$
$$=E_{b|u,c,d,\mu^{(t)}}\log P\left(u,c,d;b;\mu^{(t)}\right)$$
$$=E_{b|u,c,d,\mu^{(t)}}\left[\log\frac{u!}{(u-b)!\,b!}+(u-b)\log\frac{1}{2}+b\log\frac{\mu}{4}\right]$$
$$+\log\frac{(u+c+d)!}{u!\,c!\,d!}+c\log\left(\frac{1}{2}-\frac{\mu}{2}\right)+d\log\frac{\mu}{4}$$

Q は、その微分さえ求めればよい

$$\frac{\partial}{\partial\mu}Q\left(\mu|\mu^{(t)}\right)=E_{b|u,c,d,\mu^{(t)}}[b]\frac{1}{\mu}-c\frac{1}{1-\mu}+d\frac{1}{\mu}$$
$$=b^{(t)}\frac{1}{\mu}-c\frac{1}{1-\mu}+d\frac{1}{\mu}$$

$b^{(t)}=E_{b|u,c,d,\mu^{(t)}}[b]=u\frac{\mu^{(t)}/4}{1/2+\mu^{(t)}/4}$ と置いた

停留点を求める

$$\left(b^{(t)}+d\right)(1-\mu)=c\mu$$
$$b^{(t)}+d=\left(c+b^{(t)}+d\right)\mu$$
$$\mu=\frac{b^{(t)}+d}{b^{(t)}+c+d}$$

すなわち、$\mu^{(t+1)}=\frac{b^{(t)}+d}{b^{(t)}+c+d}$

$$b^{(t)}=u\frac{\mu^{(t)}/4}{1/2+\mu^{(t)}/4}$$
$$\mu^{(t+1)}=\frac{b^{(t)}+d}{b^{(t)}+c+d}$$

別解

事象	確率	観測数
A	1/2	u
B	μ/4	
C	1/2−μ/2	c
D	μ/4	d

例によって、別解を考えてみよう。対数尤度を考える。

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

これをb, μについて最大化する。まず連続変数であるμについて考える

$$\frac{\partial}{\partial\mu}\log P(u, c, d; b; \mu) = b\frac{1}{\mu} - c\frac{1}{1-\mu} + d\frac{1}{\mu}$$

停留点を求める

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

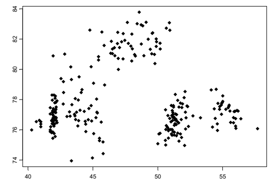
次に、bについて考える。μが与えられていれば、二項分布の性質から、最大値を与えるbは

$$b = u\frac{\mu/4}{1/2+\mu/4}$$

となる。これは2次方程式ゆえ、bを連続値とすれば、解ける。
それは、上記を反復式としても得られそうなことは、以前と同様に推定できる。

クラスタリングとの関係

クラスタリング：
観測値ベクトル x_i で記述されるサンプル i は、いずれも、いずれかのグループ（クラスタ） $j \in \{1,2,\dots,J\}$ に属するとする。サンプル i が属するクラスタを z_i で表す。この時、グループ j を記述し、また、それに属するサンプル $\{x_i \mid z_i = j\}$ を推定することを目的とする作業



クラスタリング手法

いくつもの手法が提案されているが、代表的なアプローチに、各クラスタを一つの（単峰の）分布からの i.i.d. サンプルとするものがある。
その場合の代表的な推定方法は、分布のパラメータと属するクラスタを最尤推定する方法である。すなわち、次の問題を解きたい

$$\max_\theta \max_{z_\bullet} \log\prod_i p(x_i, z_i; \theta) = \max_\theta \max_{z_\bullet} \sum_i \log p(x_i, z_i; \theta)$$

この最大化問題には、多数の離散変数 z_i が含まれるため、実際上、解けない。
そこで、例えば、次の2つの方法がとられる。
・周辺化して、属するクラスタ番号を消去し、可観測変数のみに対して、EM アルゴリズムを適用する。すなわち、次を解く。クラスタ番号はMAP推定する。

$$\max_\theta \sum_{Z_1,\dots,Z_m}\prod_{i=1}^m p(x_i, Z_i; \theta) = \max_\theta \prod_{i=1}^m \sum_{Z_i} p(x_i, Z_i; \theta)$$

・k-means 法: 元の問題を、coordinate ascent 法で（近似的に）解く

EM Alg. によるクラスタリング

今更記述する必要もないが、

$$\max_\theta \max_{z_\bullet} \log\prod_i p(x_i, z_i; \theta) = \max_\theta \max_{z_\bullet} \sum_i \log p(x_i, z_i; \theta)$$

ではなく、次を行う

$$\max_\theta \sum_{Z_1,\dots,Z_m}\prod_{i=1}^m p(x_i, Z_i; \theta) = \max_\theta \prod_{i=1}^m \sum_{Z_i} p(x_i, Z_i; \theta)$$
$$\Rightarrow \max_\theta \sum_i \log\sum_{Z_i} p(x_i, Z_i; \theta)$$
$$\Rightarrow \theta^{(t+1)} = \arg\max_\theta \mathop{E}_{Z_\bullet \mid x_\bullet, \theta^{(t)}} \sum_i \log p(x_i, Z_i; \theta)$$

K-means によるクラスタリング

「実際上、解けない」クラスタリングの尤度最大化問題を、直接的に、しかし繰り返し手順で、近似的に解く方法である。
通常、等方的かつ分散の等しい正規分布の混合分布を考える。

$$\max_\theta \max_{z_\bullet} \log\prod_i p(x_i, z_i; \theta) = \max_\theta \max_{z_\bullet} \sum_i \log p(x_i, z_i; \theta)$$

を

$$z_\bullet^{(t+1)} = \arg\max_{z_\bullet} \sum_i \log p\left(x_i, z_i^{(t)}; \theta^{(t)}\right)$$ 実際上 $z_i^{(t)}$ は用いない

$$\theta^{(t+1)} = \arg\max_\theta \sum_i \log p\left(x_i, z_i^{(t+1)}; \theta^{(t)}\right)$$ 実際上 $\theta^{(t)}$ は用いない

の繰り返しで求める。ただし、

$$p(x_i, z_i; \theta) = N(x_i, \mu_{z_i})$$

とする。

K-means（続）

$$z_\bullet^{(t+1)} = \arg\max_{z_\bullet} \sum_i \log p\left(x_i, z_i^{(t)}; \theta^{(t)}\right) = \arg\max_{z_\bullet} \sum_i \log N(x_i, \mu_{z_i})$$

については、各 x_i について、$\log N(x_i, \mu_j) = -\left|x_i - \mu_j\right|^2 + const.$ が最大になるような j を選べばよい（それが可能）。
いいかえれば、「ベクトル x_i にもっとも近い中心 μ_j を求めよ。$z_i = j$ とせよ」

$$\theta^{(t+1)} = \arg\max_\theta \sum_i \log p\left(x_i, z_i^{(t+1)}; \theta^{(t)}\right) = \arg\max_\mu \sum_i \log N(x_i, \mu_{z_i})$$

については、今度は、各 j について、$\sum_{i:\, z_i=j}\log N(x_i, \mu_j)$ が最大になるように μ_j を選べばよい。これは、$\{x_i \mid z_i = j\}$ が μ_j を中心とする正規分布のサンプルだとして、中心 μ_j を求める問題である。これは最尤推定すればよい。
いいかえれば、「ベクトル $\{x_i \mid z_i = j\}$ の平均を μ_j とせよ」

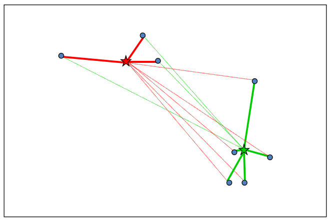
K-Means クラスタリング: Lloyd アルゴリズム

1. Lloyd アルゴリズム
2. 任意に k 個のクラスタ中心を選ぶ
3. **while** クラスタ中心が変更された
4. 各データ点をクラスタ C_i に割り当てる
   割り当てるクラスタは、最も近いクラスタ中心のクラスタ
   ($1 \leq i \leq k$)
5. すべての点への割り当てが終わったら,
   各クラスタの重心を新たなクラスタ中心とする
   すなわち、新たなクラスタ中心は、クラスタ C それぞれに

$$\sum_{\text{全}\, v \in C} v \,/\, |C|$$

*このアルゴリズムは局所解に陥ることがある.

EM-アルゴリズムの実行イメージ



K-Means クラスタリングとの比較