

情報意味論(7) EM

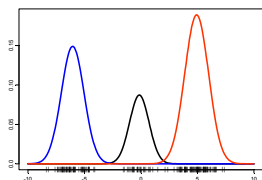
櫻井彰人
慶應義塾大学理工学部

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。
- 様々な分布を考える
 - 指数・ポアソン・対数正規...
- 分布を組み合わせる
 - 線形和 - 非観測変数の出現
 - しかし、工夫がある → EMアルゴリズム
- 分布を考えない - しょせん、多項分布
 - 変数が多いと大変。
 - しかし、工夫がある → Bayesian network

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

考え方: 各データは、まず、 j のどれかをランダムに選び(確率分布は $\{\pi_j\}$)、次に p_j に従い生成される

推定の面倒さ

データが一個の正規分布から生成されることがわかっていれば、その分布を推定することができる(平均と分散を推定すればよい。最尤推定すればよい)。例えば

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

しかし、混合分布の場合、分布が2個以上あるので、ある観測点がどの分布に属するかが分からない限り、分布(平均値等)を推定することはできない。

ところで、 分布推定は「教師なし学習」

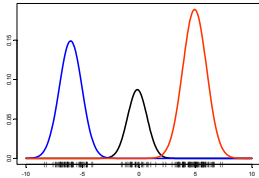
- 教師付き学習: データ $\langle x, y \rangle$
- 教師なし学習: データ x

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



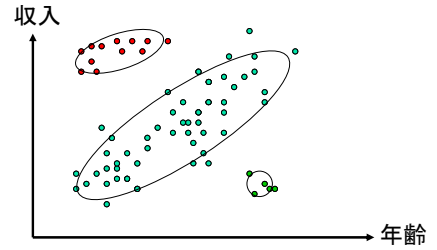
混合分布

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

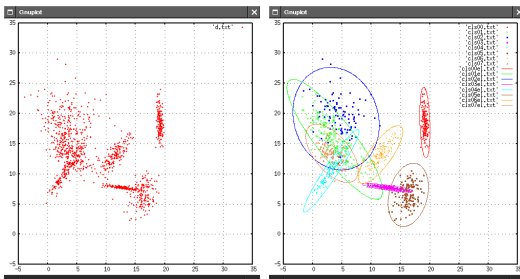
各クラスは混合分布の個々の分布に対応すると考える

- 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 $\langle x \rangle$, 全データ $\langle x, c \rangle$.
- 課題: $\langle x \rangle$ から $\langle x, c \rangle$ を推定する

クラスタリング/密度推定

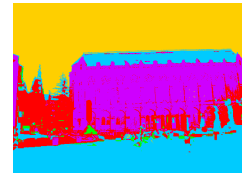
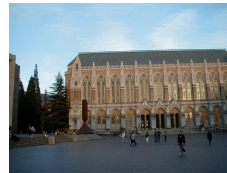


(EM) クラスタリング例



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

(K-means) クラスタリング例

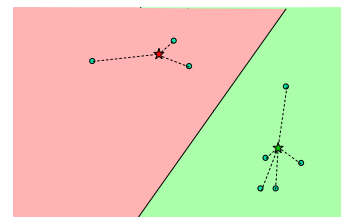


<http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster/>

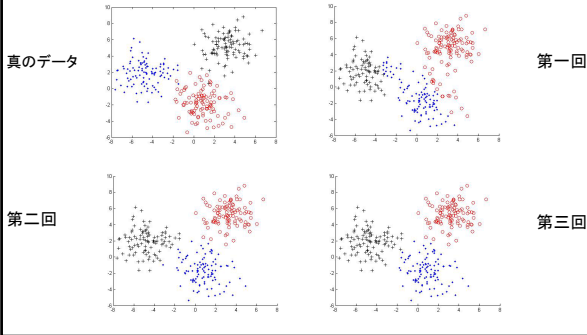
クラスタリングと分布推定

- つまり, クラスタリングができれば,
 - データ $x \rightarrow$ クラスタ(正規分布) j
 - クラスタ(正規分布) $j \rightarrow j$ の平均値・分散という具合に分布推定ができる
- しかし, 分布推定ができれば,
 - データ $x \rightarrow$ クラスタ(正規分布) j に属する確率
 - クラスタ(正規分布) j に属する確率 \rightarrow 確率最大のクラスタという具合にクラスタリングができる
- ということは, 鶏と卵。つまり, 解けない。さて, どうする?

ヒント: K-Means クラスタリング



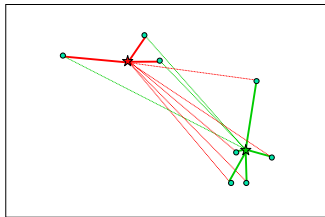
K-means クラスタリング例



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)分散は同じとする
- 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
 - 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 $\langle x \rangle \rightarrow \langle x, j \rangle$, 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
 - 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、 $o_j = \text{center of } \{x | \langle x, j \rangle\}$

K-means をソフトに



読み換えよう

- 前提
 - x の生成もとがクラスタ j ($1 \leq j \leq k$) である確率を π_j 、クラスタ j から生成された場合に生成される確率を $p_j(x)$ とする。そして $p(x) = \sum \pi_j p_j(x)$ とする。
 - まず、簡単のため、 $\pi_j = 1/k$ 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
 - 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 $\langle x \rangle \rightarrow \langle x, j \rangle$, ただし $j = \arg \min |x - o_j|$.
 - i.e. $j = \arg \max p_j(x) \Rightarrow \text{Prob}(x \in C_j) = \text{Const. } p_j(x) = (1/\sum p_j(x)) p_j(x)$
 - 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、 $o_j = \text{center of } \{x | \langle x, j \rangle\}$
 - i.e. $o_j = (1/N_j) \sum x$ (where $x \in C_j$) $\Rightarrow o_j = (1/\sum P(x \in C_j)) \sum x P(x \in C_j)$

読み換えよう (2)

- 前提
 - x の生成もとがクラスタ j ($1 \leq j \leq k$) である確率を π_j 、クラスタ j から生成された場合に生成される確率を $p_j(x)$ とする。そして $p(x) = \sum \pi_j p_j(x)$ とする。
 - 確率変数 $z_{ij} = 1$ if データ $x_i \in C_j$ or 0 otherwise
- 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
 - 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 $\langle x \rangle \rightarrow \langle x, j \rangle$, ただし $j = \arg \min |x - o_j|$.
 - i.e. $z_{ij} = "1$ if $j = \arg \min |x_i - o_j|$ or 0 otherwise" とおくと $P(x_i \in C_j) = P(z_{ij} = 1) = "1$ if $x_i \in C_j$ or 0 otherwise"
 - $\Rightarrow E[z_{ij}] = \text{Prob}(z_{ij} = 1) = \text{Const. } p_j(x_i) = (1/\sum p_j(x_i)) p_j(x_i)$
 - 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、 $o_j = \text{center of } \{x | \langle x, j \rangle\}$
 - i.e. $o_j = (1/N_j) \sum x_i$ (where $x_i \in C_j$) $= (1/\sum_i P(x_i \in C_j)) \sum_i x_i P(x_i \in C_j)$
 - $\Rightarrow o_j = (1/\sum_i P(z_{ij} = 1)) \sum_i x_i P(z_{ij} = 1) = (1/\sum_i E[z_{ij}]) \sum_i x_i E[z_{ij}]$

繰り返しの部分

- 確率変数 $z_{ij} = "1$ if データ $x_i \in C_j$ or 0 otherwise"
- 各観測点ごと、その(産みの親である)クラスタを推定する
 - $E[z_{ij}] = \text{Prob}(z_{ij} = 1) = (1/\sum_j p_j(x_i)) p_j(x_i)$
- 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
 - $o_j = (1/\sum_i E[z_{ij}]) \sum_i x_i E[z_{ij}]$

混合正規分布の学習

(分散は共通・既知)

$x_1, x_2, \dots, x_N \sim N(x | \mu_j, \sigma)$ with probability $\pi_j = 1/k$

$$z_{ij} = \begin{cases} 1 & \text{if } C(x_i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \text{E-Step} \quad E[z_{ij}] &\leftarrow \frac{p(C(x_i) = j)}{\sum_{j=1}^k p(C(x_i) = j)} = \frac{N(x_i | \mu_j, \sigma)}{\sum_{j=1}^k N(x_i | \mu_j, \sigma)} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\frac{1}{\sqrt{2\pi}\sigma} \sum_{j=1}^k e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}} \end{aligned}$$

$$\text{M-Step} \quad \mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] x_i$$

混合正規分布の学習

(分散は共通・既知)

$x_1, x_2, \dots, x_N \sim N(x | \mu_j, \sigma)$ with probability π_j

$$z_{ij} = \begin{cases} 1 & \text{if } C(x_i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{E-Step} \quad E[z_{ij}] \leftarrow \frac{p(C(x_i) = j)}{\sum_{j=1}^k p(C(x_i) = j)} = \frac{\pi_j N(x_i | \mu_j, \sigma)}{\sum_{j=1}^k \pi_j N(x_i | \mu_j, \sigma)}$$

$$\begin{aligned} \text{M-Step} \quad \mu_j &\leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] x_i \\ \pi_j &\leftarrow \frac{1}{N} \sum_{i=1}^N E[z_{ij}] \end{aligned}$$

混合正規分布の学習

(分散は既知)

$x_1, x_2, \dots, x_N \sim N(x | \mu_j, \sigma_j)$ with probability π_j

$$z_{ij} = \begin{cases} 1 & \text{if } C(x_i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{E-Step} \quad E[z_{ij}] \leftarrow \frac{p(C(x_i) = j)}{\sum_{j=1}^k p(C(x_i) = j)} = \frac{\pi_j N(x_i | \mu_j, \sigma_j)}{\sum_{j=1}^k \pi_j N(x_i | \mu_j, \sigma_j)}$$

$$\text{M-Step} \quad \mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] x_i$$

$$\pi_j \leftarrow \frac{1}{N} \sum_{i=1}^N E[z_{ij}]$$

混合正規分布の学習

$x_1, x_2, \dots, x_N \sim N(x | \mu_j, \sigma_j)$ with probability π_j

$$z_{ij} = \begin{cases} 1 & \text{if } C(x_i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{E-Step} \quad E[z_{ij}] \leftarrow \frac{p(C(x_i) = j)}{\sum_{j=1}^k p(C(x_i) = j)} = \frac{\pi_j N(x_i | \mu_j, \sigma_j)}{\sum_{j=1}^k \pi_j N(x_i | \mu_j, \sigma_j)}$$

$$\text{M-Step} \quad \mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] x_i \quad \pi_j \leftarrow \frac{1}{N} \sum_{i=1}^N E[z_{ij}]$$

$$\sigma_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] (x_i - \mu_j)^2$$

混合正規分布の学習

(高次元)

$x_1, x_2, \dots, x_N \sim N(x | \mu_j, \Sigma_j)$ with probability π_j

$$z_{ij} = \begin{cases} 1 & \text{if } C(x_i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{E-Step} \quad E[z_{ij}] \leftarrow \frac{p(C(x_i) = j)}{\sum_{j=1}^k p(C(x_i) = j)} = \frac{\pi_j N(x_i | \mu_j, \Sigma_j)}{\sum_{j=1}^k \pi_j N(x_i | \mu_j, \Sigma_j)}$$

$$\text{M-Step} \quad \mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] x_i \quad \pi_j \leftarrow \frac{1}{N} \sum_{i=1}^N E[z_{ij}]$$

$$\sigma_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij}]} \sum_{i=1}^N E[z_{ij}] (x_i - \mu_j)(x_i - \mu_j)^T$$

ソフト化(?) K-means

- 収束する!
- 証明 [Neal&Hinton '99, McLachlan&Krishnan '97]:
 - E/M stepは、データの尤度を減少させない
 - 鞍点に収束する
- 説明が必要ですね...

EM 一般的な定義

- $X = \{x_1, \dots, x_N\}$ 観測データ
- $Z = \{z_1, \dots, z_N\}$ 非観測データ (隠れ変数)
- $Y = X \cup Z$
- 目標: $E[\ln P(Y | h)]$ を最大化する h を見出すこと, i.e. (非観測データ含めた) データの事後確率を最大化する h を求める

EM 一般的な定義 (続)

- E-Step: Y の確率 (の対数) 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: $E[\ln P(Y | h)]$ の最大化であった)

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

ホスは決っていて

未観測確率変数の期待値を求める

- M-Step: 現在の仮説 h を, Q を最大化する h' で置き換える

$$h \leftarrow \operatorname{argmax}_{h'} Q(h' | h)$$

全員の満足の最大化

EM化 K-Means で考えてみよう

- k 個の(等荷重)正規分布から生成されたデータ:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \frac{(x_i - \mu_j)^2}{\sigma^2}}$$

- $X = N$ 個のデータ点の集合
- $Z = \{z_{ij}\}$. 但し
 $z_{ij} = 1$ iff i -番目のデータは
 j -番目の正規分布から生成

- 仮説 h は平均値の組

$$P(y_i | h) = P(x_i, z_{i1}, \dots, z_{ik} | h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^k z_{ij} \frac{(x_i - \mu_j)^2}{\sigma^2}}$$

EM化 K-Means, E-Step

$$\begin{aligned} Q(h' | h) &= E[\ln P(Y | h') | h, X] \\ &= E[\ln \prod_{i=1}^N P(y_i | h') | h, X] \\ &= E[\sum_{i=1}^N \ln P(y_i | h') | h, X] \\ &= E[\sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^k z_{ij} \frac{(x_i - \mu_j)^2}{\sigma^2}} \right) | h, X] \\ &= E[\sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^k z_{ij} (x_i - \mu_j')^2 \right) | h, X] \\ &= \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^k E[z_{ij} | h, X] (x_i - \mu_j')^2 \right) \end{aligned}$$

$$P(y_i | h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \frac{(x_i - \mu_j)^2}{\sigma^2}}$$

EM化 K-Means, M-Step

$$\begin{aligned} h &\leftarrow \operatorname{argmax}_h Q(h' | h) \\ &= \operatorname{argmax}_{h'} \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^k E[z_{ij} | h, X] (x_i - \mu_j')^2 \right) \\ &= \operatorname{argmin}_{h'} \sum_{i=1}^N \sum_{j=1}^k E[z_{ij} | h, X] (x_i - \mu_j')^2 \end{aligned}$$

仮説は μ の組であった。

すなわち, h は μ の組, h' は μ' の組である。

最小化は μ' による偏微分が0とおいて達成できる, i.e.,

$$\mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij} | h, X]} \sum_{i=1}^N E[z_{ij} | h, X] x_i \quad \text{なお} \quad E[z_{ij} | h, X] = \frac{e^{-\frac{1}{2\sigma^2} \frac{(x_i - \mu_j)^2}{\sigma^2}}}{\sum_{n=1}^k e^{-\frac{1}{2\sigma^2} \frac{(x_i - \mu_n)^2}{\sigma^2}}}$$

EM化 K-Means, まとめ

E-Step:

$$Q(h' | h) \leftarrow \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^k E[z_{ij} | h, X] (x_i - \mu_j')^2 \right)$$

$$E[z_{ij} | h, X] \leftarrow \eta e^{-\frac{1}{2\sigma^2} \frac{(x_i - \mu_j)^2}{\sigma^2}}$$

M-Step:

$$h \leftarrow \operatorname{argmin}_{h'} \sum_{i=1}^N \sum_{j=1}^k E[z_{ij} | h, X] (x_i - \mu_j')^2$$

$$\mu_j \leftarrow \frac{1}{\sum_{i=1}^N E[z_{ij} | h, X]} \sum_{i=1}^N E[z_{ij} | h, X] x_i$$

EM

- “Expectation Maximization” の略
- 探索アルゴリズム
 - 尤度 (likelihood) の最大化
 - Greedy, 局所最大点につかまるかも
 - ある問題クラスに適用
 - “隠れ変数 hidden (latent) variables” 対応
 - 最急降下法より遅いことも
 - 最急降下法より速いことも
- ポピュラー. しかし有効性は問題依存

EM

- 役に立つ状況は:
 - 隠れ変数の値がわかっているならば, 正しい最尤仮説 (ML hypothesis) を求めることができる
 - つまり, 混合分布のパラメータが推定できる
 - 正しい最尤仮説があれば, 上記の隠れ変数の事後分布が, 直ちに計算できる
 - つまり, 各データごと隠れ変数の値の事後分布が推定可能
- 大切なポイント: 隠れ変数
 - 「鶏と卵」状況は隠れ変数のせい。
 - しかし, 「隠れ変数」は極めて重要な概念

EM Algorithm より一般的な説明

目次

- K-means クラスタリング
 - Coordinate Descending algorithm
- 確率密度推定
 - (条件なし) 混合分布上の EM
- 回帰と分類
 - 条件付混合分布上の EM
- EM Algorithm の一般形

K-means クラスタリング

問題: 一組の観測データが所与とする

これらを K 個のクラスに分けるにはどうしたらよいか? ただし, K は所与とする.

- 第一ステップ

$$z_i^j = \begin{cases} 1 & \text{if } j = \arg \min_j \|x_i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

- 第二ステップ

$$\mu_j = \frac{1}{\sum_i z_i^j} \sum_i z_i^j x_i$$

K-means クラスタリング

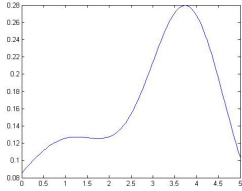
- Coordinate descent 法
- 一変数ごと下記の **歪み尺度** J の最小化を試みることに相当

$$J = \sum_{i=1}^N \sum_{j=1}^k z_i^j \|x_i - \mu_j\|^2$$

各偏微分値を0にすればよい

混合分布

問題: 与えられたサンプルデータが多峰 multimodal であつたら、その真の分布はどうやって推定したらよいのだろうか?

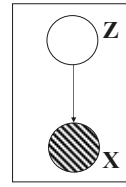


例えば、2項分布を fit しようとする

当然、アルゴリズムは収束する。しかし、結果の分布は真の分布とはかけ離れたものであろう

混合分布

- 分割統治 “divide-and-conquer” 法が使える
- 隠れ変数 Z の導入



多峰性を現すノード k 個の値のうちの一つをとる

それぞれについて、一つの分布を割り当てる、分布の全体は

$$p(x|\theta) = \sum_j p(Z^j=1|\pi_j)p(x|Z^j=1,\theta_j) = \sum_j \pi_j p(x|Z^j=1,\theta_j)$$

混合分布

混合正規分布

- このモデルでは、混合分布の一つ一つは正規分布 (未知パラメータあり) である

$$\theta_j = (\mu_j, \Sigma_j), \quad \theta = (\theta_1, \dots, \theta_k)$$

混合正規分布の確率

$$p(x|\theta) = \sum_j \pi_j N(x|\mu_j, \Sigma_j) = \sum_j \pi_j \frac{1}{(2\pi)^{m/2} |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\}$$

混合分布

- 隠れ変数 Z の事後確率:

$$\begin{aligned} \tau^j &= p(Z^j|x, \theta) && (x \text{ がクラス } j \text{ から発生した確率}) \\ &= \frac{p(x|Z^j=1, \theta)p(Z^j=1|\pi_j)}{p(x|\theta)} \\ &= \frac{\pi_j N(x|\mu_j, \Sigma_j)}{\sum_j \pi_j N(x|\mu_j, \Sigma_j)} \end{aligned}$$

- 対数尤度 Log likelihood:

$$\begin{aligned} l(\theta|x_1, \dots, x_N) &= \sum_i \log p(x_i|\theta) && p(x|\theta) = \sum_j p(Z^j=1|\pi_j)p(x|Z^j=1,\theta_j) \\ &= \sum_i \log \sum_j p(x_i, Z|\theta) && = \sum_j \pi_j p(x|Z^j=1,\theta_j) \\ &= \sum_i \log \sum_j \pi_j N(x_i|\mu_j, \Sigma_j) \end{aligned}$$

混合分布

- 変数 π_j に関する l の偏微分 (言い忘れたが、Lagrange 乗数を用いている)

$$\begin{aligned} \frac{\partial}{\partial \pi_j} \{l + \lambda(1 - \sum_j \pi_j)\} &= \frac{\partial}{\partial \pi_j} \left\{ \sum_i \log \sum_j \pi_j N(x_i|\mu_j, \Sigma_j) \right\} - \lambda \\ &= \sum_i \frac{N(x_i|\mu_j, \Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j, \Sigma_j)} - \lambda \\ &= \sum_i \frac{\tau_i^j}{\pi_j} - \lambda \end{aligned}$$

- これらを 0 とおくと方程式をよければ

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

混合分布

- 変数 μ_j に関する l の偏微分

$$\begin{aligned} \frac{\partial l}{\partial \mu_j} &= \frac{\partial}{\partial \mu_j} \left\{ \sum_i \log \sum_j \pi_j N(x_i|\mu_j, \Sigma_j) \right\} \\ &= \sum_i \frac{\pi_j N(x_i|\mu_j, \Sigma_j)}{\sum_j \pi_j N(x_i|\mu_j, \Sigma_j)} \frac{\partial}{\partial \mu_j} \left\{ -\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j) \right\} \\ &= \sum_i \tau_i^j \Sigma_j^{-1}(x_i - \mu_j) \end{aligned}$$

- これを 0 とおけば、次式が得られる

$$\mu_j = \frac{\sum_i \tau_i^j x_i}{\sum_i \tau_i^j}$$

混合分布

- 変数 Σ_j に関する l の偏微分

$$\begin{aligned} \frac{\partial l}{\partial \Sigma_j} &= \frac{\partial}{\partial \Sigma_j} \left\{ \sum_i \log \sum_j \pi_j N(x_i | \mu_j, \Sigma_j) \right\} \\ &= \sum_i \frac{\pi_j N(x_i | \mu_j, \Sigma_j)}{\sum_j \pi_j N(x_i | \mu_j, \Sigma_j)} \frac{\partial}{\partial \Sigma_j} \left\{ -\frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right\} \\ &= \sum_i \tau_i^j \left\{ -\frac{1}{2} \Sigma_j^{-1} - \frac{1}{2} \Sigma_j^{-1} (x_i - \mu_j)(x_i - \mu_j)^T \Sigma_j^{-1} \right\} \end{aligned}$$

- これを 0 とおくと次式が得られる

$$\Sigma_j = \frac{\sum_i \tau_i^j (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_i \tau_i^j}$$

混合分布

- EM Algorithm
- 第一ステップ

$$\tau_i^{j(t)} = \frac{\pi_j^{(t)} N(x_i | \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_j \pi_j^{(t)} N(x_i | \mu_j^{(t)}, \Sigma_j^{(t)})} \quad (\text{ } x_i \text{ がクラス } j \text{ から発生した確率})$$

- 第二ステップ

$$\mu_j^{(t+1)} = \frac{1}{\sum_i \tau_i^{j(t)}} \sum_i \tau_i^{j(t)} x_i, \quad \pi_j^{(t)} = \frac{1}{N} \sum_i \tau_i^{j(t-1)},$$

$$\Sigma_j^{(t+1)} = \frac{1}{\sum_i \tau_i^{j(t)}} \sum_i \tau_i^{j(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T$$

混合分布

- 完全観測の対数尤度の期待値という観点から EM アルゴリズムをみてみよう

隠れ変数 $Z_i = (Z_i^1, \dots, Z_i^k)$ を観測したと仮定する
データ集合 $\{(x_i, z_i)\}$ は完全に観測されたことになり、
尤度は **完全観測の対数尤度** となり

$$\begin{aligned} l_c(\theta | \{(x_i, z_i)\}) &= \sum_i \log p(x_i, z_i | \theta) \\ &= \sum_i \log \prod_j [\pi_j N(x_i | \mu_j, \Sigma_j)]^{z_i^j} \\ &= \sum_i \sum_j z_i^j \log[\pi_j N(x_i | \mu_j, \Sigma_j)] \end{aligned}$$

混合分布

確率変数 X と $\theta^{(t)}$ で条件付けた変数 $Z_i = (Z_i^1, \dots, Z_i^k)$ の期待値を求める

なお Z_i^j は2値の確率変数であり

$$E[Z_i^j | x, \theta^{(t)}] = p(Z_i^j = 1 | x, \theta^{(t)}) = \tau_i^{j(t)}$$

この値を $Z_i = (Z_i^1, \dots, Z_i^k)$ に対する最適な推定とすれば、
完全観測の対数尤度の期待値

$$\begin{aligned} \langle l_c(\theta | \{(x_i, Z_i)\}) \rangle_{\theta^{(t)}} &= \left\langle \sum_i \log p(x_i, Z_i | \theta) \right\rangle_{\theta^{(t)}} \\ &= \sum_i \sum_j \langle Z_i^j \rangle_{\theta^{(t)}} \log[\pi_j N(x_i | \mu_j, \Sigma_j)] \\ &= \sum_i \sum_j \tau_i^{j(t)} \log[\pi_j N(x_i | \mu_j, \Sigma_j)] \end{aligned}$$

混合分布

- 前述の **完全観測の対数尤度の期待値** を、その偏微分値を 0 にすることにより

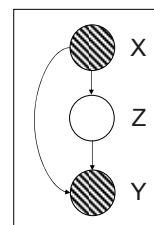
$$\pi_j^{(t+1)} = \frac{1}{N} \sum_i \tau_i^{j(t)},$$

$$\mu_j^{(t+1)} = \frac{1}{\sum_i \tau_i^{j(t)}} \sum_i \tau_i^{j(t)} x_i,$$

$$\Sigma_j^{(t+1)} = \frac{1}{\sum_i \tau_i^{j(t)}} \sum_i \tau_i^{j(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T.$$

条件付混合分布

- Graphical Model



回帰と分類用

変数 X と Z の間の関係を判別分類的にモデル化することができる、e.g. ソフトマックス softmax .

$$\pi_j(x, \xi) = p(Z^j = 1 | x, \xi) = \frac{\exp(\xi_j^T x)}{\sum_j \exp(\xi_j^T x)}$$

隠れ変数 Z 、多峰性を表すもの。
 k 個の値をとる

条件付混合分布

- 変数 Z に関して周辺化すると,

$$p(y|x, \theta) = \sum_j p(Z^j = 1|x, \xi) p(y|Z^j = 1, x, \theta_j)$$

- X は常に観測可能とする. 事後確率 $\tau^j(x, y, \theta)$ は次のように定まる

$$\begin{aligned} \tau^j(x, y, \theta) &= p(Z^j = 1|x, y, \theta) \\ &= \frac{p(Z^j = 1|x, \theta) p(y|Z^j = 1, x, \theta)}{\sum_j p(Z^j = 1|x, \theta) p(y|Z^j = 1, x, \theta)} \\ &= \frac{\pi_j(x, \xi) p(y|Z^j = 1, x, \theta)}{\sum_j \pi_j(x, \xi) p(y|Z^j = 1, x, \theta)} \end{aligned}$$

条件付混合分布

- 各要素分布にはいくつかのとり方がある

- 正規分布

$$p(y|x, \theta) = \sum_j \pi_j(x, \xi) N(y|\beta_j^T x, \sigma_j^2)$$

- ロジスティック分布

$$p(y|x, \theta) = \sum_j \pi_j(x, \xi) \mu(\theta_j^T x)^y (1 - \mu(\theta_j^T x))^{1-y},$$

ただし $\mu(\theta_j^T x)$ はロジスティック関数:

$$\mu(\theta_j^T x) = \frac{1}{1 + \exp(\theta_j^T x)}$$

条件付混合分布

- EM を用いたパラメータ推定

完全観測の対数尤度:

$$\begin{aligned} l_c(\theta|\{(x_i, y_i, z_i)\}) &= \log \prod_j p(y_i, z_i|x_i, \theta_j) \\ &= \log \prod_i \prod_j [\pi_j(x_i, \xi) p(y_i|Z_i^j = 1, x_i, \theta_j)]^{z_i} \\ &= \sum_i \sum_j z_i^j \log[\pi_j(x_i, \xi) p(y_i|Z_i^j = 1, x_i, \theta_j)] \end{aligned}$$

期待値を最適な推定と考え

$$\begin{aligned} \langle Z_i^j \rangle_{\theta^{(t)}} &= p(Z_i^j = 1|x_i, y_i, \theta^{(t)}) \\ &= \tau_i^j(x_i, y_i, \theta^{(t)}) \end{aligned}$$

条件付混合

- そうすると **完全観測の対数尤度** は

$$\begin{aligned} l(\theta|\{(x_i, y_i, z_i)\}) \\ &= \sum_i \sum_j \tau_i^{j(t)} \log[\pi_j(x_i, \xi) p(y_i|Z_i^j = 1, x_i, \theta_j)] \end{aligned}$$

- 偏微分をとり, それらを 0 とおけば, EM の更新公式が得られる

条件付混合

条件付混合モデルに対する EM algorithm のまとめ

- (E step): 事後確率 $\tau_i^{j(t)}$ を計算する
- (M step): 例えば, IRLS algorithm を用い, データ対 $(x_i, \tau_i^{j(t)})$ の元で, パラメータ ξ を更新する.
- (M step): 荷重付き IRLS algorithm を用い, データ点 (x_i, y_i) と荷重 $\tau_i^{j(t)}$ のもとで, パラメータ θ_j を更新する

IRLS: iterative reweighted least square

一般公式

- X - 全観測可能変数
- Z - 全隠れ変数
- θ - 全パラメータ

完全観測の対数尤度

仮に Z が観測されたとき, その ML 推定量は

$$\theta = \arg \max_{\theta} l_c(\theta; x, z) = \arg \max_{\theta} \log p(x, z|\theta)$$

しかしながら, Z は実際には観測されないので

$$l(\theta; x) = \log p(x|\theta) = \log \sum_z p(x, z|\theta)$$

不完全観測の対数尤度

一般公式

- 仮に $p(x, z | \theta)$ が積に分解できたとすると、完全観測の対数尤度は

$$l_c(\theta; x, z) = \sum_z f(z | x, \theta_z) \log p(x, z | \theta)$$

- ここで $f(z | x, \theta_z)$ は未知であるので、この ML 推定量の求め方は不明である。しかしながら、もし $f(z | x, \theta_z)$ の確率変数の上で平均化すれば

$$q(z | x) = \sum_{\theta_z} f(z | x, \theta_z)$$

一般公式

- $q(z | x)$ を $f(z | x, \theta_z)$ の推定量とすれば、完全観測の対数尤度は完全観測の対数尤度の期待値となり

$$\langle l_c(\theta; x, z) \rangle_q = \sum_z q(z | x) \log p(x, z | \theta)$$

- この完全観測の対数尤度の期待値はとけて、期待としては、それが完全観測の対数尤度をなぜか改善する。(EM の基本的アイデア。)

一般公式

- EM は不完全観測の対数尤度を最大化する

$$l(\theta; x) = \log p(x | \theta)$$

$$= \log \sum_z p(x, z | \theta)$$

$$= \log \sum_z q(z | x) \frac{p(x, z | \theta)}{q(z | x)}$$

Jensen の
不等式

$$\geq \sum_z q(z | x) \log \frac{p(x, z | \theta)}{q(z | x)} \xrightarrow{\Delta} L(q, \theta)$$

補助関数

一般公式

- $q(z | x)$ が所与のもと、 $L(q, \theta)$ の最大化は、完全観測対数尤度の期待値を最大化することに等しい

$$\begin{aligned} L(q, \theta) &= \sum_z q(z | x) \log \frac{p(x, z | \theta)}{q(z | x)} \\ &= \sum_z q(z | x) \log p(x, z | \theta) - \sum_z q(z | x) \log q(z | x) \\ &= \langle l_c(\theta; x, z) \rangle_q - \sum_z q(z | x) \log q(z | x) \end{aligned}$$

一般公式

- θ が所与のもと、 $q^{(t+1)}(z | x) = p(z | x, \theta^{(t)})$ とすれば $L(q, \theta)$ の最大化が計られる

$$L(q^{(t+1)}(z | x), \theta^{(t)}) = \sum_z p(z | x, \theta^{(t)}) \log \frac{p(x, z | \theta^{(t)})}{p(z | x, \theta^{(t)})}$$

$$= \sum_z p(z | x, \theta^{(t)}) \log p(x | \theta^{(t)})$$

$$= \log p(x | \theta^{(t)})$$

$$= l(\theta^{(t)}; x) \quad \text{注: } L(q, \theta^{(t)}) \text{ は } l(\theta^{(t)}; x) \text{ の上界となる}$$

一般公式

- 前記において、EMの各ステップで、 $L(q, \theta)$ を最大化した

- しかしながら、最後に最大化した $L(q, \theta)$ が同時に不完全観測対数尤度 $l(\theta; x)$ をも最大化すると、どうして分かるのか?

一般公式

- $l(\theta; x)$ と $L(q, \theta)$ との差は

$$\begin{aligned} l(\theta; x) - L(q, \theta) &= l(\theta; x) - \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)} \\ &= \sum_z q(z|x) \log p(x|\theta) - \sum_z p(z|x) \log \frac{p(x, z|\theta)}{q(z|x)} \\ &= \sum_z q(z|x) \log p(x|\theta) + \log \frac{q(z|x)}{p(z|x, \theta) p(x|\theta)} \\ &= \sum_z q(z|x) \log \frac{q(z|x)}{p(z|x, \theta)} = D(q(z|x) \| p(z|x, \theta)) \end{aligned}$$

非負であり $q(z|x) = p(z|x, \theta)$ でユニークに最小化する \rightarrow KL divergence

一般公式

- EM と交互最小化について
 - 尤度の最大化はモデルと経験分布との間の KL divergence を最小化することと同値であった。
 - 隠れ変数 Z を含めることにより、KL divergence は (x, z) の結合分布(それぞれの)間の完全観測の KL divergence となる

一般公式

$$\begin{aligned} D(\tilde{p}(x) \| p(x|\theta)) &= -\sum_x \tilde{p}(x) \log p(x|\theta) + \sum_x \tilde{p}(x) \log \tilde{p}(x) \\ &\leq -\sum_x \tilde{p}(x) L(q, \theta) + \sum_x \tilde{p}(x) \log \tilde{p}(x) \\ &= -\sum_x \tilde{p}(x) \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)} + \sum_x \tilde{p}(x) \log \tilde{p}(x) \\ &= \sum_x \tilde{p}(x) \sum_z q(z|x) \log \frac{\tilde{p}(x) q(z|x)}{p(x, z|\theta)} \\ &= D(\tilde{p}(x) q(z|x) \| p(x, z|\theta)) \end{aligned}$$

一般公式

- EM algorithm の再定式化
 - (E step) $q^{(t+1)}(z|x) = p(z|x, \theta^{(t)})$
 $D(q(z|x) \| p(z|x, \theta))$
 - (M step) $q^{(t+1)}(z|x) = \arg \min_q D(q \| \theta^{(t)})$
 $\theta^{(t+1)} = \arg \min_{\theta} D(q^{(t+1)} \| \theta)$ \leftarrow 交互最小化

まとめ(一般的な説明)

- (条件なし)混合分布
 - [Graphic model](#)
 - [EM algorithm](#)
- 条件付混合分布
 - [Graphic model](#)
 - [EM algorithm](#)
- EM algorithm の一般的な表現
 - 補助関数の最大化
 - 完全観測の KL divergence の最小化