THE UNIVERSITY OF TOKYO

数理情報工学特論第一
【機械学習とデータマイニング】
4章：教師なし学習②

かしま　ひさし
鹿島 久嗣
（数理 6 研）
kashima@mist.i.～

東京大学
THE UNIVERSITY OF TOKYO

DEPARTMENT OF MATHEMATICAL INFORMATICS

潜在変数モデルと、潜在変数モデルのそのEM法による最尤推定を学びます

- 多次元正規分布の最尤推定
- 潜在変数モデル
- 混合分布
  - 混合正規分布
- EM法による潜在変数モデルの最尤推定
  - E-ステップとM-ステップ
  - EM法の正当性

多次元正規分布の最尤推定

多次元正規分布のパラメータ（平均と分散共分散行列）を最尤推定します

- 多次元正規分布の確率密度関数：

$$P(\phi(x);\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\phi(x)-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\phi(x)-\boldsymbol{\mu})\right)$$

  - $\boldsymbol{\mu}$：平均ベクトル　$\boldsymbol{\Sigma}$：分散共分散行列

  を、訓練データ集合$\{x^{(i)}\}_{i=1}^{N}$から推定することを考える

- 最尤推定によって得られるパラメータを$\boldsymbol{\mu}^*$および$\boldsymbol{\Sigma}^*$とすると：

$$(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*) \equiv \underset{(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N}\log P(\phi(x^{(i)});\boldsymbol{\mu},\boldsymbol{\Sigma})$$

- 多次元正規分布の場合：

$$(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*) \equiv \underset{(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\operatorname{argmax}} -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2N}\sum_{i=1}^{N}(\phi(x)-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\phi(x)-\boldsymbol{\mu})$$

平均パラメータの最尤推定量は、データの平均と一致します

- 目的関数： $J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2N}\sum_{i=1}^{N}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})$

  を平均パラメータ$\boldsymbol{\mu}$についてこれを偏微分すると

$$\frac{\partial J(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})$$

- 最適な $\boldsymbol{\mu}=\boldsymbol{\mu}^*$ でこれが$\mathbf{0}$ のはずなので、これを$\mathbf{0}$とおいて整理すると：

$$\boldsymbol{\Sigma}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}(x^{(i)})\right) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^*$$

- 両辺の左から$\boldsymbol{\Sigma}$を掛ければ：

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}(x^{(i)})$$

- つまり、多次元正規分布の平均パラメータ$\boldsymbol{\mu}$の最尤推定量は、データ集合に対する特徴ベクトル集合 $\{\boldsymbol{\phi}(x^{(i)})\}_{i=1}^{N}$ の平均となる

分散共分散行列の最尤推定量を求めるため、少し準備します

- 次に分散共分散行列の最尤推定量 $\boldsymbol{\Sigma}^*$を求めるため、$\boldsymbol{\Sigma}$について目的関数を最大化：

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2N}\sum_{i=1}^{N}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})$$

- まずは、以下の等式が成り立つことを確認しておく

$$\begin{aligned}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu}) &= \mathrm{Tr}\left((\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})\right) \\ &= \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^\top\right) \\ &= \mathrm{Tr}\left(\boldsymbol{\Lambda}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^\top\right)\end{aligned}$$

  - 最初の等式が成り立つことはスカラー値のトレースもまたスカラー値であることから
  - 2つ目の等式は行列 $\mathbf{A}$と$\mathbf{B}$ に対し$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ であることから
  - 最後の等式は分散共分散行列 $\boldsymbol{\Sigma}$ と精度行列 $\boldsymbol{\Lambda}$ の関係 $\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Lambda}$ から

分散共分散行列の最尤推定量もデータの分散共分散行列になります

- 先の等式を用いて目的関数$J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$を、$\boldsymbol{\Lambda}$の関数として書き換えれば

$$J(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv \frac{1}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2N} \sum_{i=1}^{N} \mathrm{Tr}\left(\boldsymbol{\Lambda} (\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^{\top} \right)$$

  - $|\boldsymbol{\Sigma}| = |\boldsymbol{\Lambda}|^{-1}$であることに注意

- $J(\boldsymbol{\mu}, \boldsymbol{\Lambda})$を$\boldsymbol{\Lambda}$で微分すると：

$$\frac{\partial J(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^{\top}$$

  - ここで行列の微分の公式をもちいた：

$$\frac{\partial \log |\mathbf{A}|}{\partial A} = (\mathbf{A}^{-1})^{\top}, \quad \frac{\partial \mathrm{Tr}(\mathbf{AB})}{\partial A} = \mathbf{B}^{\top}$$

- 最尤推定量 $\boldsymbol{\Sigma}^{*-1} = \boldsymbol{\Lambda}^{*}$において、上式は全ての要素が$\mathbf{0}$の行列と等しいはずであるので、これを解くと、最尤推定量は：

$$\boldsymbol{\Sigma}^{*} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu})^{\top}$$

潜在変数モデル

教師なし学習の主要タスクは4つあるのでした

- 教師なし学習においては通常、データ上の確率分布$P(\phi(x))$を何らかの形で推定することが行われる
- $P(\phi(x))$の使い道としては主に以下の4つが挙げられる
  1. 確率分布そのものを用いた分析
  2. データの確率評価
  3. 未観測値（欠損値）の推定【前回みた】
  4. 潜在変数の推定【今回】

潜在変数の推定
「隠れた情報」を表す仮想的な変数を考え、これを推定します

- 興味のある変数が（訓練）データ集合においても観測されない潜在的な仮想変数であるという点で、未観測値の推定とは異なる
- つまり、xとは別にまったく観測されない変数y（潜在変数）と呼ばれる変数を仮定し、その条件付き分布 $P(y|x)$ を推定することが目的となる
  - y は x のもつ「隠れた情報」を表す
- たとえば、yを各データの属するグループであるとすると、$P(y|x)$ を用いてデータ集合を自動的にグループ分け（**クラスタリング**）することができる

潜在変数とは、データの（観測されない）状態を表す観測されない変数です。これを持つモデルが潜在変数モデルです

- 潜在変数をもつ確率分布は、データの特徴ベクトルに加え、観測されない**潜在変数**をyとして$P(\phi(x),y)$と書く
  - 潜在変数は複数あってよいが、簡単のため1つ（yが1次元）とする
- 潜在変数yとは、データxの潜在的な状態を表す仮想的な変数
  - $\mathcal{X}$を顧客の集合、$\phi(x)$を顧客$x \in \mathcal{X}$をその人の個人情報や購買履歴で表現した特徴ベクトルとする
  - yの取りうる値を$\mathcal{Y} \equiv \{1,2,3\}$の3つであるとすると、$y$は顧客$x$が3つのグループ$\{1,2,3\}$のうち、どのグループに属するかを表す
- yは観測されない変数なので、yのとる値の意味は与えられておらず単に$\mathcal{X}$の各要素の各$y \in \mathcal{Y}$への所属程度$P(y|\phi(x))$だけがわかる
  - 各グループに属する顧客を調べることによって、それぞれのグループの意味を「解釈」することになる（富裕層、若年層、…）

混合分布

潜在変数モデルの代表的モデルは混合分布で、複数の要素確率分布をある割合で混ぜあわせた形をしています

- **混合分布**：潜在変数モデルにおける、もっとも標準的なモデル
  - 複数個（K個）の確率分布を組み合わせたような確率分布
  - 単一の確率分布と比較して、複雑な分布の形を表現できる
- 混合分布における潜在変数の取りうる値は、$\mathcal{Y} \equiv \{1,2,\ldots,K\}$の$K$通り
- 混合分布は2種類の構成要素をもつ：
  - 要素確率分布の混合比を決める確率分布$P(y; \boldsymbol{\theta}^{(0)})$
    - パラメータ$\boldsymbol{\theta}^{(0)} \equiv (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_D^{(0)})$をもち

$$P(y; \boldsymbol{\theta}^{(0)}) \equiv \theta_y^{(0)} \text{ s.t. } \sum_{y=1}^{K} \theta_y^{(0)} = 1, \quad \theta_d^{(0)} \geq 0$$

  - K個の要素確率分布$\{ P(x \mid y; \boldsymbol{\theta}^{(y)}) \}_{y=1}^{K}$
    - K個の要素確率分布はそれぞれパラメータ$\boldsymbol{\theta}^{(y)}$を持つ

混合モデルにおける潜在変数は「どの要素確率分布を用いてデータを生成したか」に対応します

- 全てのパラメータをまとめて$\Theta \equiv \{\boldsymbol{\theta}^{(k)}\}_{k=0}^{K}$と書くとすると、混合分布は以下のように定義される。

$$P(x, y; \Theta) \equiv P(x|y; \boldsymbol{\theta}^{(y)})P(y; \boldsymbol{\theta}^{(0)}) = P(x|y; \boldsymbol{\theta}^{(y)})\theta_y^{(0)}$$

- これをデータの生成過程という観点から見てみると：
  - まず、$P(y; \boldsymbol{\theta}^{(0)})$ によって、K個の要素確率分布のうちどの確率分布を用いてデータを生成するかを決定する
    - y番目の要素確率分布が選ばれる確率は$\theta_y^{(0)}$
  - 何番目の確率分布が選ばれるかを示す確率変数が、潜在変数$y \in \mathcal{Y} \equiv \{1,2, \ldots, K\}$となる
  - そして、選ばれたy番目の確率分布（パラメータ$\boldsymbol{\theta}^{(y)}$をもつ）を用いて、データ$x$が生成される

要素確率分布として多次元正規分布を用いたものが混合正規分布です

- K個の要素確率分布$\{ P(x \mid y; \boldsymbol{\theta}^{(y)}) \}_{y=1}^{K}$ のそれぞれが、多次元正規分布であるとする
- パラメータ$\boldsymbol{\theta}^{(y)} \equiv (\boldsymbol{\mu}^{(y)}, \Sigma^{(y)})$ をもつ y 番目の多次元正規分布は以下のように書ける：

$$P(x|y; \boldsymbol{\mu}^{(y)}, \boldsymbol{\Sigma}^{(y)}) \equiv \frac{1}{(2\pi)^{D/2}} |\boldsymbol{\Sigma}^{(y)}|^{-1/2} \exp\left((\boldsymbol{\phi}(x) - \boldsymbol{\mu}^{(y)})^\top \boldsymbol{\Sigma}^{(y)^{-1}} (\boldsymbol{\phi}(x) - \boldsymbol{\mu}^{(y)}) \right)$$

  - それぞれの正規分布は異なるパラメータ $(\boldsymbol{\mu}^{(y)}, \Sigma^{(y)})$ をもつ

EM法による潜在変数モデルの最尤推定

潜在変数モデルの最尤推定には、潜在変数について周辺化し「見える部分」のみの確率を用います

- データxと潜在変数yの同時分布を、パラメータΘをもつ確率分布$P(x,y;\Theta)$によってモデル化し、これを（訓練）データ集合$\{x^{(i)}\}$から推定する
- 潜在変数はそもそも観測されない仮想的な変数であるため「観測されたデータの出現確率を最大化する」という最尤推定の方針に従えば、最適なモデルパラメータの最尤推定値Θ^*を得るために解くべき最適化問題は：

$$\Theta^* \equiv \operatorname*{argmax}_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \log \sum_{y \in \mathcal{Y}} P(x^{(i)}, y; \Theta)$$

  - ただし、$P(x) = \sum_{y \in \mathcal{Y}} P(x,y)$であることに注意する

潜在変数モデルの最尤推定を簡単に行うための方法がEM法です

- 最適化問題の目的関数：

$$J(\Theta) \equiv \sum_{i=1}^{N} \log \sum_{y \in \mathcal{Y}} P(x^{(i)}, y; \Theta)$$

- この目的関数をニュートン法や最急勾配法によって直接Θについて最大化することも可能だが煩雑
- **EM (Expectation-Maximization; 期待値最大化)法**：
  - 目的関数の下界を作ることで潜在変数モデルの推定を非常に見通し良く簡単に行える方法

EM法は対数尤度関数の下界を最大化します

- 最適化問題の目的関数：$J(\Theta) \equiv \sum_{i=1}^{N} \log \sum_{y \in \mathcal{Y}} P(x^{(i)}, y; \Theta)$
- EM法では、まず、目的関数 $J(\Theta)$の下界 $J'(\Theta, \mathbf{z})$ を作る

$$J(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \log \sum_{y \in \mathcal{Y}} P(x^{(i)}, y; \Theta) \frac{P(y|x^{(i)}; \mathbf{z})}{P(y|x^{(i)}; \mathbf{z})}$$

$$\geq \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log \frac{P(x^{(i)}, y; \Theta)}{P(y|x^{(i)}; \mathbf{z})} \equiv J'(\Theta, \mathbf{z})$$

  - 若干人工的ではあるが、データxが与えられた時の潜在変数yの条件付き確率分布$P(y|\, x; \mathbf{z})$を導入する
    - $P(y|\, x; \mathbf{z})$はパラメータ$\mathbf{z}$ をもつ
  - 不等式は、以下のイェンセン(Jensen)の不等式（次頁）による

目的関数の下界を作るためにイェンセン(Jensen)の不等式を用います

- 目的関数 $J(\Theta)$ の下界 $J'(\Theta, \mathbf{z})$ ：

$$J(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \log \sum_{y\in\mathcal{Y}} P(x^{(i)}, y; \Theta) \frac{P(y|x^{(i)}; \mathbf{z})}{P(y|x^{(i)}; \mathbf{z})}$$

$$\geq \frac{1}{N}\sum_{i=1}^{N} \sum_{y\in\mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log \frac{P(x^{(i)}, y; \Theta)}{P(y|x^{(i)}; \mathbf{z})} \equiv J'(\Theta, \mathbf{z})$$

をつくるためにイェンセン(Jensen)の不等式を利用できる

$$\log \sum_{y\in\mathcal{Y}} f(y)g(y) \geq \sum_{y\in\mathcal{Y}} f(y) \log g(y) \text{ s.t. } \sum_{y\in\mathcal{Y}} f(y) = 1,\ f(y) \geq 0,\ g(y) \geq 0$$

- $f(y) = P(y|x^{(i)}; \mathbf{z}), \quad g(y) = \frac{P(x^{(i)}, y; \Theta)}{P(y|x^{(i)}; \mathbf{z})}$ とおけばよい
- なお、この形式は、対数関数に対するイェンセンの不等式
- 一般に上に凸関数に対して上式のような不等式が成立する

EM法は、目的関数の下界のパラメータを交互に最適化することを繰り返すアルゴリズムです

- EM法では、直接最適化をすることが面倒な目的関数の代わりに、不等式を用いて作った下界 $J'(\Theta, \mathbf{z})$ を最大化する

$$J'(\Theta, \mathbf{z}) \equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log \frac{P(x^{(i)}, y; \Theta)}{P(y|x^{(i)}; \mathbf{z})}$$

- 元々の目的関数のパラメータはΘのみであったのに対し、下界 $J'(\Theta, \mathbf{z})$では新たなパラメータ$\mathbf{z}$が加わっていることに注意する
- EM法では、Θと $\mathbf{z}$ を交互に最適化を繰り返すことで、解を逐次的に更新していく

EM法はE-ステップとM-ステップを収束するまで繰り返します

- EM法では、E-ステップとM-ステップの2つのステップを繰り返すことで、解を逐次的に更新していく：
  - E-ステップ：Θ を現在の値 $\Theta=\Theta^{\mathrm{OLD}}$ に固定し、$J'(\Theta^{\mathrm{OLD}}, \mathbf{z})$を$\mathbf{z}$について最大化する

$$\mathbf{z}^{\mathrm{NEW}} \equiv \underset{z}{\operatorname{argmax}} \; J'(\Theta^{\mathrm{OLD}}, \mathbf{z})$$

  - M-ステップ：$\mathbf{z}$を現在の値$\mathbf{z}=\mathbf{z}^{\mathrm{OLD}}$に固定し、$J'(\Theta, \mathbf{z}^{\mathrm{OLD}})$を$\Theta$について最大化する

$$\Theta^{\mathrm{NEW}} \equiv \underset{\Theta}{\operatorname{argmax}} \; J'(\Theta, \mathbf{z}^{\mathrm{OLD}})$$

- 目的関数$J'(\Theta, \mathbf{z})$は各ステップにおいて徐々に改善され、目的関数が有限であれば、いつかは目的関数値がこれ以上改善しなくなるため繰り返しは終了する
- 本来の目的関数 $J(\Theta)$ が改善されていることは後ほど確認する

E-ステップの最適化問題は（一定の条件のもと）閉じた解が得られます

- 目的関数の下界は：

$$J'(\Theta, \mathbf{z}) \equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log \frac{P(y|x^{(i)}; \Theta) P(x^{(i)}; \Theta)}{P(y|x^{(i)}; \mathbf{z})}$$

- ここで、仮に $P(y|\, x^{(i)}; \mathbf{z}) = P(y|\, x^{(i)}; \Theta)$と取れば：
  （等号が成立できるほどにP$(y|\, x^{(i)}; \mathbf{z})$が精緻であれば）

$$J'(\Theta, \mathbf{z}) \equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log P(x^{(i)}; \Theta) = \frac{1}{N} \sum_{i=1}^{N} \log P(x^{(i)}; \Theta) = J(\Theta)$$

  となり、下界 $J'(\Theta, \mathbf{z})$ は上界 $J(\Theta)$ に一致する。

- この等号を達成するzは、E-ステップの最大化問題の解となっている
  - 上界 $J(\Theta)$ は $\mathbf{z}$ に依存しないため、zについては定数とみなせる
- $P(y|\, x^{(i)}; \mathbf{z}) = P(y|\, x^{(i)}; \Theta)$を満たすzが存在するならば、E-ステップの最大化問題を解かなくとも、直接閉じた形の最適解が得られる

モデルが混合分布の場合には、E-ステップはモデルの確率評価のみで実行できます

- $P(y|\, x^{(i)}; \mathbf{z}) = P(y|\, x^{(i)}; \Theta)$にベイズの定理を適用すると：

$$P(y|x^{(i)}; \mathbf{z}) = \frac{P(x^{(i)}|y; \Theta)P(y; \Theta)}{P(x^{(i)}; \Theta)} = \frac{P(x^{(i)}|y; \Theta)P(y; \Theta)}{\sum_{y \in \mathcal{Y}} P(x^{(i)}|y; \Theta)P(y; \Theta)}$$

- $P(x^{(i)}, y; \Theta)$が混合分布の場合には：

$$P(x^{(i)}, y; \Theta) \equiv P(x|y; \boldsymbol{\theta}^{(y)})\theta_y^{(0)}$$

  - $\theta_y^{(0)}$：y番目の要素確率分布を選ぶ確率
  - $\mathrm{P}(x|y; \boldsymbol{\theta}^{(y)})$：$y$番目の要素確率分布
    - 多次元正規分布のときには前述の $P(x \mid y; \boldsymbol{\mu}^{(y)}, \Sigma^{(y)})$

- これを用いると$P(y|\, x^{(i)}; \mathbf{z})$は：

$$P(y|x^{(i)}; \mathbf{z}) = \frac{P(x|y; \boldsymbol{\theta}^{(y)})\theta_y^{(0)}}{\sum_{y \in \mathcal{Y}} P(x|y; \boldsymbol{\theta}^{(y)})\theta_y^{(0)}}$$

ただし、訓練データの数や潜在変数の取りうる値の数が膨大になると、最適化問題を解く必要がある場合があります

- $P(y|\, x^{(i)};\, \mathbf{z}) = P(y|\, x^{(i)};\, \boldsymbol{\Theta})$を用いて $P(y|\, x^{(i)};\, \mathbf{z})$ を求める場合、実際に $J'(\boldsymbol{\Theta}^{\mathrm{OLD}},\, \mathbf{z})$ を$\mathbf{z}$について最大化する問題を解く必要はない
- 従って、通常はE-ステップの最適化問題は存在せず、$P(y|\, x^{(i)};\, \boldsymbol{\Theta})$を用いて、全ての$(i,\, y)$の組み合わせについて$P(y|\, x^{(i)};\, \mathbf{z})$を計算すればよい
- これは、あたかも $\mathrm{z}_{i,\, y} \equiv P(y|\, x^{(i)};\, \mathbf{z})$ という「変数」が存在し、E-ステップはこの値を更新しているかのように見ることができる
- 一方、訓練データの数や潜在変数の取りうる値の数が膨大なときには、全ての$(i,\, y)$に対して $\mathrm{z}_{i,\, y}$ を管理することが難しくなる
- そのような場合にはパラメータ$\mathbf{z}$を持つような確率モデル$P(y|\, x^{(i)};\, \mathbf{z})$を考えることで、$\mathrm{z}_{i,\, y}$ を「圧縮」する必要が出てくる

M-ステップは重みつきの（完全データに対する）最尤推定になります

- M-ステップでは目的関数の下界 $J'(\Theta, \mathbf{z})$ をΘについて最大化する
- 目的関数の下界 $J'(\Theta, \mathbf{z})$ は：

$$J'(\Theta, \mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log P(x^{(i)}, y; \Theta) - \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}) \log P(y|x^{(i)}; \mathbf{z})$$

- Θに関係のある部分は1項目のみであるので、これだけを取り出して考えると、M-ステップの最大化問題は以下のようになる

$$\Theta^{\mathrm{NEW}} \equiv \operatorname*{argmax}_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}}) \log P(x^{(i)}, y; \Theta)$$

- データxと潜在変数yの両方が既知の場合の最尤推定となっている
  - ただし、潜在変数のとりうる値はそれぞれ $P(y|\, x^{(i)}; \mathbf{z}^{\mathrm{OLD}})$ で重みづけられているような形となる

モデルが混合分布の場合のM-ステップの最適化問題を解いてみます

- $P(x^{(i)}, y; \Theta)$が混合分布の場合には、最大化問題はさらに：

$$\Theta^{\mathrm{NEW}} \equiv \underset{\{\boldsymbol{\theta}^{(y)}\}_{y=1}^{K}, \boldsymbol{\theta}^{(0)}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}}) \log P(x|y; \boldsymbol{\theta}^{(y)}) \theta_y^{(0)}$$

$$= \underset{\{\boldsymbol{\theta}^{(y)}\}_{y=1}^{K}, \boldsymbol{\theta}^{(0)}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}}) \log P(x|y; \boldsymbol{\theta}^{(y)})$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}}) \log \theta_y^{(0)}$$

$$\text{s.t.} \sum_{y=1}^{K} \theta_y^{(0)} = 1, \quad \theta_d^{(0)} \geq 0$$

まずは混合比パラメータを求めてみます

- まずは、これを$\boldsymbol{\theta}^{(0)}$について最大化する
- この最適化問題には制約 $\Sigma_{y=1}{}^{K}\,\theta_y{}^{(0)} = 1$ があるため、ラグランジュ関数 $L(\boldsymbol{\theta}^{(0)})$を作る

$$L(\boldsymbol{\theta}^{(0)}) \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{y\in\mathcal{Y}} P(y|x^{(i)};\mathbf{z}^{\mathrm{OLD}})\log\theta_y^{(0)} - \lambda\left(\sum_{y=1}^{K}\theta_y^{(0)} - 1\right)$$

  - λ：ラグランジュ乗数
- $L(\boldsymbol{\theta}^{(0)})$を$\theta_y{}^{(0)}$で偏微分すると：$\dfrac{\partial L(\boldsymbol{\theta}^{(0)})}{\partial\theta_y^{(0)}} = \dfrac{1}{N}\sum_{i=1}^{N} P(y|x^{(i)};\mathbf{z}^{\mathrm{OLD}})\dfrac{1}{\theta_y^{(0)}} - \lambda$
- これを0とおいて解くと：$\theta_y^{(0)} = \dfrac{1}{\lambda}\dfrac{1}{N}\sum_{i=1}^{N} P(y|x^{(i)};\mathbf{z}^{\mathrm{OLD}})$
- 制約 $\Sigma_{y=1}{}^{K}\,\theta_y{}^{(0)} = 1$ が満たされるようにλを決めれば：

$$\theta_y^{(0)} = \frac{\sum_{i=1}^{N} P(y|x^{(i)};\mathbf{z}^{\mathrm{OLD}})}{\sum_{y\in\mathcal{Y}}\sum_{i=1}^{N} P(y|x^{(i)};\mathbf{z}^{\mathrm{OLD}})}$$

要素確率分布が多次元正規分布の場合のパラメータは、重みつき最尤推定になります

- $\boldsymbol{\theta}^{(y)}$ についての最大化問題は：

$$\boldsymbol{\theta}^{(y)\mathrm{NEW}} \equiv \underset{\boldsymbol{\theta}^{(y)}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}}) \log P(x|y; \boldsymbol{\theta}^{(y)})$$

- 要素確率分布が多次元正規分布の場合を考えると、目的関数：

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2N}\sum_{i=1}^{N} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu}^{(y)})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu}^{(y)})$$

  - 多次元正規分布の最尤推定と殆ど同じ形：
    ただし i 番目の訓練データが重み$P(y|\, x^{(i)}; \mathbf{z}^{\mathrm{OLD}})$を与えられている

- 従って、同様に最尤推定量を求めると：

$$\boldsymbol{\mu}^{(y)\mathrm{NEW}} = \sum_{i=1}^{N} \frac{P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}})}{\sum_{i=1}^{N} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}})} \boldsymbol{\phi}(x^{(i)})$$

$$\boldsymbol{\Sigma}^{(y)\mathrm{NEW}} = \sum_{i=1}^{N} \frac{P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}})}{\sum_{i=1}^{N} P(y|x^{(i)}; \mathbf{z}^{\mathrm{OLD}})} (\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu}^{(y)})(\boldsymbol{\phi}(x^{(i)}) - \boldsymbol{\mu}^{(y)})^{\top}$$

重み$P(y|\, x^{(i)}; \mathbf{z}^{\mathrm{OLD}})$での重みつき平均

EM法の正当性：もともとの目的関数が改善されます

- E-ステップにおいて、条件$P(y|\, x^{(i)}; \mathbf{z}) = P(y|\, x^{(i)}; \boldsymbol{\Theta})$を満たす$\mathbf{z}$を見つけることができれば、E-ステップが終了した時点で、下界 $J'(\Theta, \mathbf{z})$ は本来の目的関数 $J(\Theta)$ は一致しているのだった
- この状態で、M-ステップを実行すると、 $J'(\Theta, \mathbf{z})$ が大きくなる
- 不等式 $J'(\Theta, \mathbf{z}) \leqq J(\Theta)$は必ず成立することから、M-ステップでは同時に$J(\Theta)$も大きくなることになる
- つまり、EM法は、本来の目的関数 $J(\Theta)$のかわりにその下界$J'(\Theta, \mathbf{z})$を逐次的に改善していくが、それは同時に本来の目的関数$J(\Theta)$も改善しているということがわかる
- ただし、$P(y|\, x^{(i)}; \mathbf{z}) = P(y|\, x^{(i)}; \boldsymbol{\Theta})$ が満たせない場合はこの限りでない
  - このようなケースは「一般化」EM法と呼ばれる