概要

前回はデータ点をクラスタに分割する手法である K-means アルゴリズムを学んだ。この手法では、各データをクラスタに割り当てる E ステップと、クラスタの中心であるプロトタイプを変更する M ステップを繰り返していくことで、目的関数である歪み尺度を減少させていた。

9.1.1 では K-means アルゴリズムを画像圧縮に適用し、画像データを圧縮する例を確かめる。各画素をクラスタに割り当てた後に、プロトタイプとなる色（RGB）で置き換えることで、データを最大4%ほどまで圧縮できる。

9.2 節以降では、 K-means を確率的アプローチに拡張した EM アルゴリズムについて学ぶ。特に、生成データが混合ガウス分布に従って得られると仮定したモデリングを行う。混合ガウス分布の表現に、潜在変数を導入することで EM アルゴリズムを簡潔に表現できることを確かめる。

9.1.1 画像分割と画像圧縮

画像分割の目的や前提などは読み合わせとする。各画素の {R,G,B} を3次元データとして扱い K-means 法を適用し、画素をクラスタに割り当てる。クラスタに対応するプロトタイプの {R,G,B} を各画素に置き換えたものが図9.3 である。

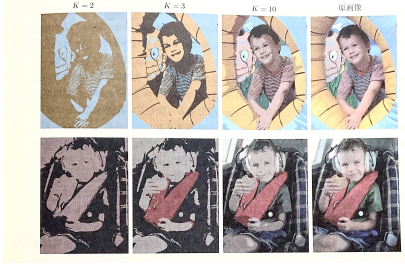
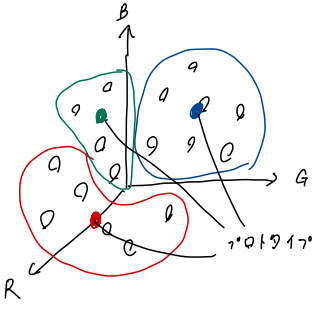


図 9.3



例えば K＝3 のとき{R,G,B}の空間にプロットされる各画素は3つのクラスタに分類される。同じクラスタに属する画素は単一の {R,G,B} の色で表現できる。

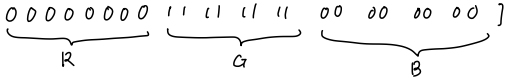
ここで扱った画像分割は、そのまま画像圧縮に適用できる。具体的には、クラスタに割り当てられた代表的なベクトル {R,G,B} を保持しておき、各画素はどの代表ベクトルを参照すべきかのインデックスを保持すると、データ量が削減される。これをベクトル量子化と呼ぶ。

これを下の例で確かめる。この例では、画素数が N の画像をK 個のベクトルで量子化する方法を考える。

圧縮しないケース

{R, G, B} の各色素は通常 0 ~ 255 ($=2^8$) で表される。

例えば緑色は、 [00000000 11 11 11 11 00 00 00 00] の



24ビットで表される。よって画像データ全体は 24N ビットである。

圧縮するケース

まず、K本の代表ベクトルを保持する必要があり、これには 24K ビットを要する

0 : [0000 ⋯ 0100]
1 : [0100 ⋯ 0010]
⋮
K-1 : [1000 ⋯ 1101]

} 24K ビット

また、N個の画素がどのインデックスを参照するかの情報を保持する。K個のインデックスのどれを参照するかは、 $\log_2 K$ ビッドで表現できる。

1ビット ⇒ 0 ~ 1 (2進数) ⇒ $0 \sim 2^1 - 1$
2ビット ⇒ 00 ~ 11 (2進数) ⇒ $0 \sim 2^2 - 1$
⋮
Lビット ⇒ 0…0 ~ 1…1 (2進数) ⇒ $0 \sim 2^L - 1$

$\log_2 K$ ビット ⇒ ⇒ $0 \sim K-1$

K 状態を表現できる

よって全画素で $N \log_2 K$ ビット必要になる。

合計すると、画像圧縮した場合の画像データは $24K + N\log_2 K$ ビットとなる。
実際のデータに適用してみたところ、K=2 で 4.2% 程の圧縮率になる。

9.2 混合ガウス分布

K-means アルゴリズムでは歪み尺度という非確率的な目的関数の最小化を考えていたが、この節では確率的なモデルに拡張する。中でも混合ガウス分布を仮定したEMアルゴリズムによる推定を行う。確率的な扱いをすることのメリットは、各クラスタへの割り当ての事後確率（負担率）を得ることができる点などがある。

混合ガウス分布はすでに上巻で学んだが、ここでは離散的な潜在変数を導入して定式化しなおす。後の EM アルゴリズムで確認するが、この潜在変数を導入することで、EM アルゴリズムの定式化が簡単になる。

既に学んだ混合ガウス分布は、混合係数 π を用いて次のように定式化されていた。

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

ここでのK次元の2値確率変数 z を導入する。細かい設定は読み合わせとするが、確率 π_k で $z_k = 1$ となる 1-of-K 符号化法を適用した確率変数である。

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

この z はいわゆる、「混合分布のどの山に割り当てられるか」を表す潜在変数である。どの山に割り当てられたかが決定した元での観測変数 x は単純なガウス分布で表現できる。

$$p(x \mid z) = \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k}$$

$z_k = 1$ のガウス分布のみ残る。

このように設定した x, z について、x で周辺化すると、正しく混合ガウス分布の表式が得られることを確かめる。

$$\begin{aligned} p(x) &= \sum_{z} p(x, z) \\ &= \sum_{z} p(x \mid z) p(z) \\ &= \sum_{z} \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k} \cdot \prod_{k=1}^{K} \pi_k^{z_k} \\ &= \sum_{z} \prod_{k=1}^{K} \left\{ \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \right\}^{z_k} \end{aligned}$$

$z = (1, 0, \cdots, 0)$
$z = (0, 1, \cdots, 0)$
$\vdots$
$z = (0, 0, \cdots, 1)$
に分解する。

$$
\begin{aligned}
&= \prod_{k=1}^{K} \left\{ \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \right\}^{z_k} \Big|_{z=(1,0,\cdots,0)} \\
&\quad + \prod_{k=1}^{K} \left\{ \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \right\}^{z_k} \Big|_{z=(0,1,\cdots,0)} \\
&\qquad \vdots \\
&\quad + \prod_{k=1}^{K} \left\{ \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \right\}^{z_k} \Big|_{z=(0,0,\cdots,1)}
\end{aligned}
$$

$$
\begin{aligned}
&= \pi_1 \mathcal{N}(x \mid \mu_1, \Sigma_1) \\
&\quad + \pi_2 \mathcal{N}(x \mid \mu_2, \Sigma_2) \\
&\qquad \vdots \\
&\quad + \pi_K \mathcal{N}(x \mid \mu_K, \Sigma_K) \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)
\end{aligned}
$$

↓混合ガウス分布の式と一致

潜在変数を導入した結果、混合分布を得られたものの、このままではただ表現を変えただけでメリットがないように思われる。実は x が得られた元での z の条件付き確率を考えると、これが EM アルゴリズムにおいて大いに役立つ。

$$
\gamma(z_k) = p(z_k = 1 \mid x) = \frac{p(z_k = 1)\, p(x \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)\, p(x \mid z_j = 1)}
$$

$$
= \frac{\pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)} \qquad (9.13)
$$

この式は、混合要素 k が観測データ x を「説明する」度合いとして解釈でき、負担率（responsibility）と呼ばれる。また、この式自体が後に現れる EM アルゴリズムにおける E ステップに相当しており、データがどの混合要素に割り当てられているかの予想（Expectation）となっている。

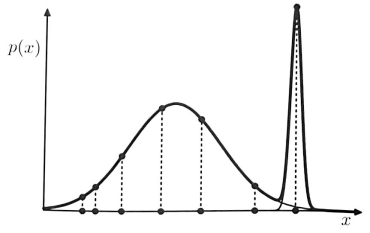
混合ガウス分布からサンプルを生成する方法については読み合わせとする。

9.2.1 最尤推定

潜在変数を導入した混合ガウス分布について整理できたので、観測データ集合に当てはめる問題を考える。事前分布などを考えない単純な推定を考えると、下記の対数尤度関数を最小にするような最尤推定が考えられる。

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\} \qquad (9.14)$$

しかし、この最尤推定を実施しようとすると、対数尤度が無限大に発散してうまく学習できないケースが存在する。これは特異性の存在に起因する問題で、混合分布のうちの特定のガウス分布が、ある単一のデータに適合した結果生ずる。例えば、下図のように、右端のデータに対して一つのガウス分布が適合し、このガウス分布の分散パラメータが0に極限まで近づくことで、対数尤度は無限大に発散してしまう。その結果得られるモデルも有用なものではなくなる。



この問題は適当なヒューリスティックによって解決することができる。例えば、上記の問題のような局所解に陥りそうな場合に、ガウス分布の平均パラメータをランダムな値に、そして分散を大きな値に設定し直して最適化を続ければ解決できる。（感想：だいぶ雑な解決方法に思えたがほんとにこれがメジャーなやり方なのか。。）また、ベイジアンアプローチを適用する場合でもこの問題を避けられるが、この内容は 10.1 節に回す。

特異性の問題の他に、識別可能性の問題がある。これは K 個の混合要素の順番の入れ替えにより、同等な解が複数存在するという問題であるが、良いモデルを得る目的においては問題にならないらしい。（感想：上巻285頁ではニューラルネットの重み対称性を考慮してエビデンスを補正していたので、今回もモデル選択の枠組みでは考慮する必要があるのかも。）

9.2.2 混合ガウス分布のEMアルゴリズム

最尤推定によって混合ガウス分布のあてはめを行うときの注意点を見てきたが、ここでは実際に最尤推定を行う手続きを学ぶ。 EM アルゴリズムは潜在変数を持つモデルの最尤解を求めるための方法であり、今回はこの EM アルゴリズムを適用して混合ガウス分布の最尤推定を行う。

EM アルゴリズムは、K-means アルゴリズムと同様に E ステップと M ステップを繰り返すことによって最尤解を得る手法である。 E ステップでは各データ点がどの混合要素に属するかの確率（負担率）を更新し、 M ステップでは負担率を固定した元で、尤度を減らすようパラメータを更新する。K-means アルゴリズムと EM アルゴリズムの違いを比較すると次のようになる。

	K-means	EM アルゴリズム
目的関数	歪み尺度 $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\|x_n - \mu_k\|\|^2$	対数尤度 $\ln p(X) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x\|\mu_k, \Sigma_k)\right\}$ (9.14)
E ステップで更新する値	r_{nk}	$\{\gamma(z_k)\}_{k=1}^{K}$ (9.13)
M ステップで更新するパラメータ	μ_k	μ_k, Σ_k, π_k

まず、M ステップについて考える。Mステップで更新すべきパラメータは3つで、負担率 $\{\gamma(z_k)\}_{k=1}^{K}$ を固定した元で、対数尤度関数の導関数を0とおくことでパラメータを更新する。以降では、それぞれのパラメータの更新後の値を求めてみる。

μ についてのMステップ

対数尤度関数を μ_k で微分し、0とおくことで次の式を得る。

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \mu_k} \ln p(X \mid \pi, \mu, \Sigma) \\
&= \frac{\partial}{\partial \mu_k} \left[\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\} \right] \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \mu_k} \left\{ \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\}}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)} \frac{\partial}{\partial \mu_k} \left\{ (x_n - \mu_k)^{\top} \Sigma_k^{-1} (x_n - \mu_k) \right\} \\
&\qquad \text{（} = \gamma(z_{nk}) \text{）} \qquad \text{（ガウス分布のexpの中身だけ出てくる。）} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k^{-1} (x_n - \mu_k) \\
&\qquad \text{両辺に左から } \Sigma_k \text{ をかけて消去する。} \\
\Leftrightarrow \quad \mu_k &= \frac{1}{\sum_{n=1}^{N} \gamma(z_{nk})} \left(\sum_{n=1}^{N} \gamma(z_{nk}) x_n \right) \\
&\qquad \sum_{n=1}^{N} \gamma(z_{nk}) = N_k \text{ とおく。} \\
&= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n
\end{aligned}
$$

この更新式の右辺は、データ $\{x_n\}_n$ を負担率 $\{\gamma(z_{nk})\}_n$ で重み付き平均をとった値であり、混合要素 k におけるガウス分布の平均の更新という直観と一致している。

Σ についてのMステップ

μ と同様に対数尤度関数に関する導関数を求めれば良い。途中でガウス分布に対する Σ の微分の式を使う。

$$
\begin{aligned}
\frac{\partial \mathcal{N}}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma}\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\} \\
&= \frac{1}{(2\pi)^{D/2}}\left(\frac{\partial}{\partial \Sigma}\frac{1}{|\Sigma|^{1/2}}\right)\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\} \\
&\quad + \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\frac{\partial}{\partial \Sigma}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\} \quad ※3 \\
&= \frac{1}{(2\pi)^{D/2}}\left(-\frac{1}{2}\right)|\Sigma|^{-1/2}\Sigma^{-1}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\} \quad ※4 \\
&\quad + \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\frac{1}{2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\}\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1} \quad ※5 \\
&= -\frac{1}{2}\Sigma^{-1}\mathcal{N}(\mathbf{x}|\mu,\Sigma) + \frac{1}{2}\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}\mathcal{N}(\mathbf{x}|\mu,\Sigma) \\
&= -\frac{1}{2}\left\{\Sigma^{-1}-\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}\right\}\mathcal{N}(\mathbf{x}|\mu,\Sigma)
\end{aligned}
$$

参照：http://sioramen.sub.jp/prml_wiki/doku.php/%E3%82%AC%E3%82%A6%E3%82%B9%E5%88%86%E5%B8%83%E3%81%AE_mu_sigma_%E3%81%AB%E3%81%A4%E3%81%84%E3%81%A6%E3%81%AE%E5%BE%AE%E5%88%86

対数尤度の導関数を0とおくと次の式が得られる。

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \Sigma_k}\ln p(X|\pi,\mu,\Sigma) \\
&= \text{<} \mu_k \text{と同じ操作をおこなった答え>} \\
&= \sum_{n=1}^{N}\frac{\frac{\partial}{\partial \Sigma_k}\left\{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right\}}{\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)} \\
&= \sum_{n=1}^{N}\frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}\left(-\frac{1}{2}\right)\left\{\Sigma^{-1}-\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^{\prime}\Sigma_k^{-1}\right\} \\
&\qquad \text{左右から } \Sigma_k \text{ をかける} \\
\Leftrightarrow\ 0 &= \sum_{n=1}^{N}\gamma(z_{nk})\left\{\Sigma_k-(x_n-\mu_k)(x_n-\mu_k)^{\prime}\right\} \\
\Leftrightarrow\ \Sigma_k &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^{\prime}
\end{aligned}
$$

この更新式も、共分散の式を負担率で重み付けた形式となっていることがわかる。

π_k についてのMステップ

最後に混合係数 π_k について最適化を行う。 π_k は各パラメータの総和が1である制約条件を考慮し、ラグランジュの未定係数法を用いる。未定係数 λ を導入し、次のラグランジュ関数の停留点を求める。

$$L(\pi, \lambda) = \ln p(X|\pi, \mu, \Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

これを π_k で微分して、0とおくと

$$0 = \frac{\partial}{\partial \pi_k} L(\pi, \lambda) = \quad <\mu_k と同じ操作なので略>$$

$$= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \pi_k}\left\{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right\}}{\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} + \lambda$$

$$= \sum_{n=1}^{N} \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} + \lambda$$

$$= \frac{1}{\pi_k}\sum_{n=1}^{N}\gamma(z_{nk}) + \lambda$$

$$\Leftrightarrow \quad \pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N}\gamma(z_{nk})$$

これを k について総和をとると、

$$\sum_{k=1}^{K}\pi_k = -\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk})$$

$$\Leftrightarrow \quad 1 = -\frac{1}{\lambda}\sum_{n=1}^{N}\underbrace{\sum_{k=1}^{K} p(z_{nk}=1)}_{=1}$$

$$\Leftrightarrow \quad 1 = -\frac{1}{\lambda}N$$

$$\Leftrightarrow \quad \lambda = -N$$

よって

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk})$$

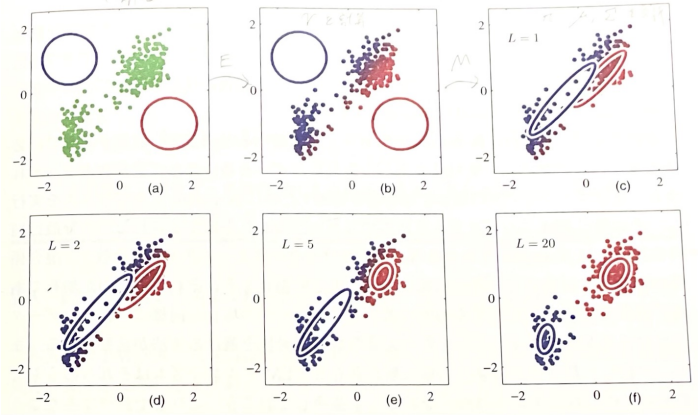
を得る。

以上より各パラメータの M ステップにおける更新式を得られた。各更新式は陽な解を与えていないことに注意する。というのも、負担率 $\gamma(z_{nk})$ には更新対象のパラメータ自身も含まれるためである。そこで、M ステップでパラメータ更新と E ステップの負担率の更新を交互に実施する必要があるのである。

E ステップでは、更新されたパラメータに基づいて負担率を（9.13）式で更新する。

$$\gamma(z_{nk}) = \frac{\pi_k \cdot \mathcal{N}(x \mid \mu_k \cdot \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k \cdot \Sigma_k)} \qquad (9.13)$$

ここまでで得られた EM アルゴリズムを間欠泉データに適用したのが下の図である。2つのガウス分布の混合分布を用いて EM アルゴリズムを適用すると、およそ 20 回の繰り返す数で収束することがわかる。



EM アルゴリズムは K-means アルゴリズムと比べると収束するまでに必要な繰り返し数と、一度の繰り返しの計算量が多い。そのため、混合分布の適切な初期値を決めるために K-means アルゴリズムを予め実行することがある（らしい）。

また、特異点により特定のガウス分布が一点に潰れる問題は EM アルゴリズムでも生じるため、既に述べたような対策を施す必要がある。すなわち、単一のデータに過剰に適合するガウス分布がある場合、平均パラメータをランダムな値に変更し、分散パラメータを大きな値に修正するなどのヒューリスティックが有効である。