GMMの学習法について.

1. GMMとは
2. 学習方法.

1. GMMとは.

混合ガウスモデル, Gaussian Mixture Model.
データが混合ガウス分布で生成されたと考える.
入力されたデータに適合する混合ガウス分布のパラメータを求める.

1.1 混合ガウスモデル

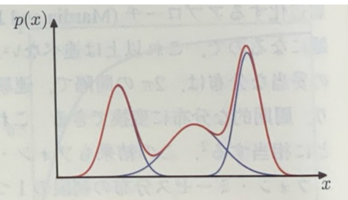


図 2.22 3つのガウス分布（青線，係数に応じた比率で縮小してある）と，これらの和（赤線）を示した，1次元中での混合ガウス分布の例．

混合分布のひとつ.
ある分布を, ガウス分布の線形結合として表現する.

1.1.1 混合分布

データ,分布の形状によっては,それを1つの分布でうまく表現できないことがある.
そういった分布を,複数の分布を重ね合わせる(線形結合する)ことで表現するもの.

十分な数のガウス分布を用い, 各分布の重みやパラメータ(平均,共分散)を調節すれば,
ほぼ任意の連続密度関数を,任意の精度で近似できる.

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

ここで,

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

は混合要素(mixture component)といい, それぞれ平均μ_k, 共分散Σ_kをパラメータとしてもつ.

また, パラメータπ_kは混合係数といい,

$$\sum_{k=1}^{K} \pi_k = 1, \quad 0 \le \pi_k \le 1$$

である.

分布のパラメータ

$\pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$,
$\mu = \{\mu_1, \mu_2, \cdots, \mu_K\}$,
$\Sigma = \{\Sigma_1, \Sigma_2, \cdots, \Sigma_K\}$

の求め方は, 最尤推定, EMアルゴリズムがある
→ch2.

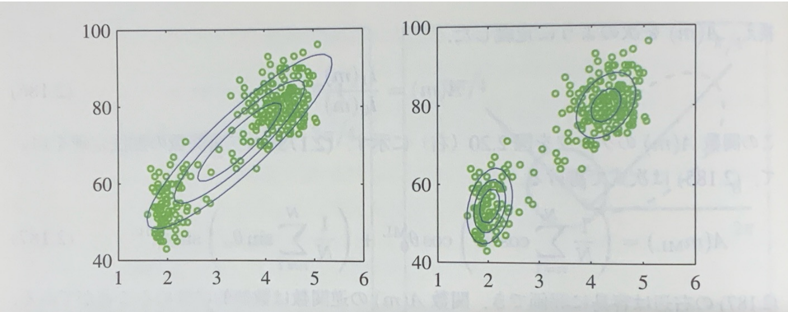


図 2.21 「Old Faithful 間欠泉」データの散布図を，確率密度の等高線（青）と共に示す．左側は，単一のガウス分布の下，最尤推定でフィッティングしたもの．ここで，このガウス分布ではデータの2つのかたまりをうまく捉えられていないことに注目する．2つのかたまりの間の中央部分は，実際にはデータが比較的疎だが，この部分に大きな確率を割り当ててしまっている．右側は，9章の手法を用いて，最尤推定でデータにフィッティングした，2つのガウス分布の線形結合である分布，データをより適切に表現できている．

2. 学習方法

2.1 分布のパラメータを推定する
- 2.1.1 最尤推定
- 2.1.2 ガウス分布の最尤推定
- 2.1.3 混合ガウス分布と最尤推定

2.2 EMアルゴリズム

2.1 分布のパラメータを推定する

分布の形が分かっているとき，入力のデータに最もフィットするように，分布のパラメータを決めたい
パラメータの決め方は，最尤推定とEMアルゴリズムがある。

2.1.1 最尤推定

尤度を最大化する．尤度は積の形であるため，最大値を考えにくい（微分がダルい）
したがって，尤度の対数をとった，対数尤度を最大化する問題で解く。
解き方はいろいろ．ラグランジュとか．

2.1.1.1 尤度

確率変数 X のサンプルデータ $D=\{x_1, x_2, \ldots, x_n\}$ について，x_i が独立に同一の確率分布に従うとき，
D が生成される確率 $P(D)$ は，

$$P(D) = \prod_{x_i \in D} p(x_i)$$

と書ける．この $P(D)$ を，尤度という．
尤度の対数をとったものは，対数尤度といい，

$$\log P(D) = \log \prod_{x_i \in D} p(x_i)$$

$$= \sum_{x_i \in D} \log p(x_i)$$

できる．
尤度の最大化と，対数尤度の最大化は等価である．

2.1.2 ガウス分布の最尤推定

ある多変量ガウス分布から，観測値 $\{x_n\}$ が独立に得られたと仮定したデータ集合 $X = (x_1, x_2, \cdots, x_N)^T$
があるとき，この分布のパラメータは最尤推定で求められる．
対数尤度は

$$\ln p(X \mid \mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T \Sigma^{-1}(x_n - \mu)$$

で与えられる．整理すると，尤度関数はデータに対して

$$\sum_{n=1}^{N} x_n, \quad \sum_{n=1}^{N} x_n x_n^T$$

のみに依存していると分かる．この2つを，ガウス分布の十分統計量という．
対数尤度の μ についての導関数は，

$$\frac{\partial}{\partial \mu}\ln p(X \mid \mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu)$$

であり，これを0とおくて，平均の最尤値が得られる．

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad -①$$

ML: Maximum Likelihood

共分散については，Magnus and Neudecker (1999) より，

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T \quad -②$$

である．μ_{ML} が含まれているのは，この式が μ，Σ について同時に最大化したものであるから．
①より，μ_{ML} は Σ_{ML} へ依存しないので，①で μ_{ML} を求めてから②で Σ_{ML} を求める．

真の分布下で μ_{ML}，Σ_{ML} の値を評価すると，

$$\mathbb{E}[\mu_{ML}] = \mu$$
$$\mathbb{E}[\Sigma_{ML}] = \frac{N-1}{N}\Sigma$$

となる. $\Sigma_{ML} < \Sigma$ であるが, これは②式を次のように補正することで, 真の値が得られる.

$$\tilde{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

2.1.3 混合ガウス分布と最尤推定

混合ガウス分布は, K個のガウス分布の線形結合で, パラメータは

$\pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$

$\mu = \{\mu_1, \mu_2, \cdots, \mu_K\}$

$\Sigma = \{\Sigma_1, \Sigma_2, \cdots, \Sigma_K\}$

である. 定義は

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

であるから, $X = \{x_1, x_2, \cdots, x_N\}$ についての対数尤度関数は

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

となる. 対数の内部にkについての和があり, 複雑であり. 閉形式の解析解では最尤推定解が得られない.

尤度関数を最大化するアプローチは, 繰り返し的な数値最適化や, EMアルゴリズムがある.

2.2 EMアルゴリズム

EMアルゴリズムは, 有名なくり返し的なパラメータ推定手法.

GMMのような潜在変数をもつモデルの最尤パラメータを求めるエレガントな方法.

2.2.1 潜在変数

これまで, 入力されるデータは

$X = \{x_n : n = 1 \cdots N\}$

としていた. GMMでは各データ点 x_n がどのガウス分布から生成されたものか分からない.

ここで, 潜在変数 z を導入する.

$Z = \{z_{nk} : n = 1, \cdots, N \text{ and } k = 1, \cdots, K\}$

$$z_{nk} = \begin{cases} 1 & (\text{if } x_n \text{ is generated by } k\text{ th Gaussian}) \\ 0 & (\text{else}) \end{cases}$$

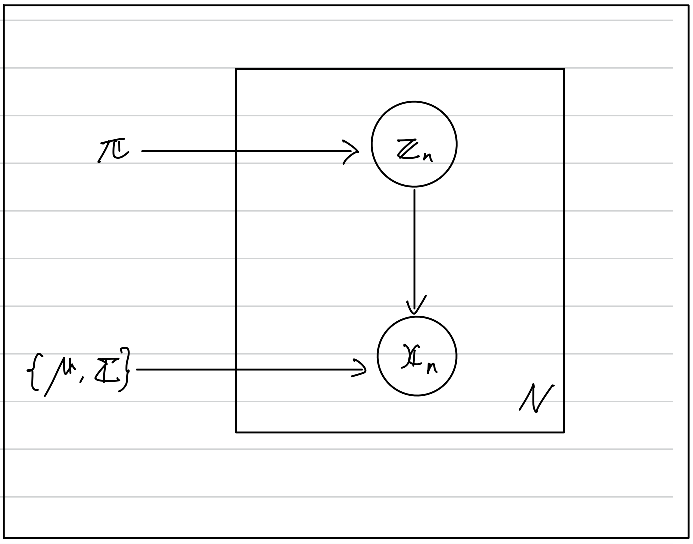
このとき, 混合分布は右図のように表せる.

図からも分かるように, x_n は z_n に依存している.

したがって, 周辺尤度 $p(x_n)$ は, 潜在変数 z_n marginalizing out することで得られる

$$p(x_n) = \sum_{z_n} p(z_n)p(x_n|z_n)$$
$$= \sum_{k=1}^{K} p(z_{nk} = 1)p(x_n|z_{nk} = 1)$$
$$= \sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k, \Sigma_k)$$

これはすなわち, 混合ガウス分布は, その潜在変数を陽に含んだ形で記述できることを示す.



$n = 1, \cdots, N$ について,

$z_n = [z_{n1}, \cdots, z_{nK}]^T$ が潜在変数,

x_n が観測ベクトル

であることを示す.

EMにかかわるとき、$\{\mathbb{X}, \mathbb{Z}\}$ は完全データセット、$\mathbb{X}$ は不完全データセットとよばれる。
EMアルゴリズムでは $\log p(\mathbb{X}|\Lambda)$ を最大化する代わりに、$\log p(\mathbb{X}, \mathbb{Z}|\Lambda)$ を最大化する。

（各ガウスのパラメータ (μ_k, Σ_k) を $\Lambda = \{\mu_k, \Sigma_k | i=1,\dots,K\}$ としている）

重要なのは、$\log p(\mathbb{X}, \mathbb{Z}|\Lambda)$ の最大化が簡単で、各イテレーションで閉形式の解が得られること。
しかし、実際は $\mathbb{Z}$ を知り得ないため、$\log p(\mathbb{X}, \mathbb{Z}|\Lambda)$ を計算できない。幸い、この事後分布 $(P(\mathbb{Z}|\mathbb{X},\Lambda))$ はベイズの定理から導ける

2.2.2 負担率

事後分布 $P(\mathbb{Z}|\mathbb{X},\Lambda)$、特に、$x_n$ についての事後確率は、

$$\gamma(z_{nk}) = P(z_{nk}=1 | x_n, \Lambda)$$
$$= \frac{P(z_{nk}=1|\Lambda)\, p(x_n | z_{nk}=1, \Lambda)}{P(x_n|\Lambda)}$$
$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

と書ける。これは、コンポーネント k が x の観測を説明する度合いを表す「負担率」としても解釈できる。

2.2.3. Q関数

$\gamma(z_{nk})$ とこれまでに推定したパラメータ Λ^{old} から、
$\mathbb{Z}$ の事後分布の下で $\log p(\mathbb{X}, \mathbb{Z}|\Lambda)$ の期待値を計算することで新しいパラメータの推定値 Λ を計算できる。

$$Q(\Lambda|\Lambda^{old}) = \mathbb{E}_{\mathbb{Z}}\{\log p(\mathbb{X}, \mathbb{Z}|\Lambda) | \mathbb{X}, \Lambda^{old}\}$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} P(z_{nk}=1 | x_n, \Lambda^{old}) \log p(x_n, z_{nk}=1|\Lambda)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n, z_{nk}=1|\Lambda)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n | z_{nk}=1, \Lambda) P(z_{nk}=1|\Lambda)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log [\mathcal{N}(x_n|\mu_k, \Sigma_k)\pi_k]$$

この関数は、補助関数、あるいはQ関数とよばれる。

Λ について、

$$\frac{\partial Q(\Lambda|\Lambda^{old})}{\partial \Lambda} = 0$$

として、$Q(\Lambda|\Lambda^{old})$ を最大化すればよい。この条件から、

音声認識では、コンポーネントの事後分布を計算することを、アラインメントという。これは、音声ベクトル x_n が、どのコンポーネントに近いかを求めるため。

$$\begin{cases} \mu_k = \dfrac{\sum_{k=1}^{K} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \\ \Sigma_k = \dfrac{\sum_{k=1}^{K} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \quad -① \\ \pi_c = \dfrac{1}{K}\sum_{n=1}^{N} \gamma(z_{nk}) \end{cases}$$

を得る。ここで、

$$\begin{cases} n_k = \sum_{n=1}^{N} \gamma(z_{nk}) \\ f_k = \sum_{n=1}^{N} \gamma(z_{nk}) x_n \end{cases}$$

$$\mathbb{S}_k = \sum_{n=1}^{N} \gamma(z_{nk}) x_n x_n^T$$

とおくと、①は

$$\begin{cases} \mu_k = \dfrac{1}{n_k} \mathbb{f}_k \\ \Sigma_k = \dfrac{1}{n_k} \mathbb{S}_k - \mu_k \mu_k^T \\ \pi_k = \dfrac{1}{N} n_k \end{cases}$$

とかける。

この値を用いて、$Q(\Lambda|\Lambda^{old})$ の値を求めることができる。

イテレーションを通じてパラメータを更新し、Q関数の値の変化量がしきい値を下回ったら終了。

2.2.4 E-M

Init.

- XからランダムにK個のデータをえらぶ。
- えらんだデータを$\{\mu_k | k=1, \cdots, K\}$にあてがう。
- $\pi_k = \frac{1}{K}$, $\Sigma_k = I$ にそれぞれ設定する。$(k=1, \cdots, K)$.

E-STEP

- Q関数がμ_k, Σ_k, π_k (for $k=1, \cdots, K$) で表せるように、$\gamma(z_{nk})$を、全学習サンプルについて求める。

M-STEP

- E-stepで求めた$\gamma(z_{nk})$を用いて、新しいパラメータを求める。
- $\gamma(z_{nk})$と求めたパラメータから、Q関数の値を求める。
- 前試行時のQ関数の値と、今回求めたQ関数の値を比較する。
    - 値の増加が止まっていれば、アルゴリズム終了。
    - そうでなければ、goto E-step.