

ベイズ推論

4.1 - 4.2



4. 混合モデルと近似推論.

3章(13)

…事後分布や予測分布を解析的(?)求められるモデル

4章では

複数分布モデル (混合モデル)
推論 (近似推論)

4.1. 混合モデルと事後分布の推論.

○ 単純な確率分布を単体で使うだけでは表現力
が十分でない。

～混合モデル

・各種確率分布をアロットのように組み合わせる

4.1.1. 混合モデルを使う理由.

○ データが複数

…データが複数に分かれ、データを表現するより確率分布を割り当てる方が可能(?)。

○ 確率モデルとの組合せ。

…データの傾向をより的確に表した予測モデルを構築できる。

○ 异常値へ対応

4.1.2. 混合モデルのデータ生成過程.

データを表現するためのモデルを構築するため、

「観測されたデータ点ひとつひとつがどう生まれた過程で生成されるのか」
を具体的に想定する必要がある。

クラスタ構造を持つデータ生成過程

$X = \{x_1, \dots, x_N\}$: データ
 K : クラスター数 ...既知

1. 初期化, クラスターの混合比率 $\pi = (\pi_1, \dots, \pi_K)^T$ が事前分布 $p(\pi)$ から生成される。
 $\dots \pi_k \in (0, 1)$ かつ $\sum_{k=1}^K \pi_k = 1$
2. 初期化, クラスター $k=1, \dots, K$ に対する観測モデルのパラメータ θ_k (平均値, 分散など) が事前分布 $p(\theta_k)$ から生成される。
3. $n=1, \dots, N$ に対して x_n は対応するクラスター割り当て s_n が比率 π によつて選ばれる。
4. $n=1, \dots, N$ に対して s_n によつて選択された k 番目の確率分布 $p(x_n | \theta_k)$ からデータ x_n が生成される。

⇒ 生成モデル

s_n : 隠れ変数 / 潜在変数

・直積観測モデル

・ x_n を発生させた確率分布を潜在的に決めていた r.u.

二、アイデアを使, 実用的なアルゴリズムを構築する(二回).
具体的には確率分布を定義する必要がある.

⇒ 入力, π , θ から s は逆にわかる.

入力, π , θ

最終的に取り出される点 x_n は対応する確率分布を定義.

… K 種類の確率分布を定義

⇒ ここでは, すべてガウス分布を設定する.

$$p(x_n | \Theta_k) = N(x_n | \mu_k, \Sigma_k) \quad \text{for } k=1, \dots, K \quad (4.1)$$

$\cdots \Theta_k = \{\mu_k, \Sigma_k\}$: 観測モデルのパラメータ.

- ガウス分布だけでなくても可能.
- 線形回帰を混合、種類の要因を確率モデルの組合せも可能.

ステップ3

K個の観測モデルを各データ点に12割り当てるまでの手段.

$\cdots S_{n,k}$ は 1 of k 表現 を用いる.

$\rightsquigarrow \pi$ を 10^6 ランダムなカテゴリ分布.

$$p(S_n | \pi) = \text{Cat}(S_n | \pi) \left(= \prod_{k=1}^K \pi_k^{S_{n,k}} \right) \quad (4.2)$$

S_n : K次元ベクトル.

$\exists k, S_{n,k} = 1 \Rightarrow k$ 番目のラストが指定された.

π : 混合比率 (mixed proportion)

\cdots どの k が選ばれるかを決定する.

$$\sum_{k=1}^K \pi_k = 1.$$

ステップ3, 4の分布を組み合わせて, x_n を生成するための確率分布は.

$$p(x_n | S_n, \Theta) = \prod_{k=1}^K p(x_n | \Theta_k)^{S_{n,k}} \quad \begin{array}{l} \rightarrow \text{1つ1つ} \\ \text{ } \\ \pi_k \leftarrow p(x_n | \Theta_k) \propto L_2? \end{array}$$

$$\text{where } \Theta = \{\Theta_1, \dots, \Theta_K\}$$

↓
K個のうじの
観測モデル
 $p(x_n | \Theta_k)$ が
選択される.

ステップ 2

ステップ 4 で定義した観測モデルのパラメータ Θ_k に関する事前分布 $p(\Theta_k)$ を定義。

~ $p(x_n | \Theta_k)$ に対する共役事前分布を設定するが一般的。

(例) $p(x_k | \Theta_k)$ がホーアண分布
 \Rightarrow ガニマ分布を設定。

ステップ 1

混合比率πについて

~ 分量、事前分布を観測するまでは未知。

~ 事前分布をもたらすデータから予測させた方が良い。

π : カテゴリ分布のパラメータ

~ 共役事前分布は K 次元のディリクレ分布

$$p(\pi) = \text{Dir}(\pi | \alpha) \left(= C_0(\alpha) \prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \quad (4.4)$$

α : K 次元ベクトル、ディリクレ分布、 $(\lambda_1 \pi_1 + \dots + \lambda_K \pi_K)$ 。

ここで、確率分布を用いる。同時分布は π より (2) など。

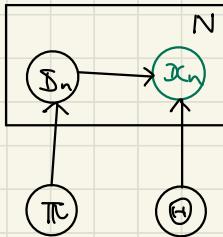
$$p(X, S, \Theta, \pi) = \underbrace{p(X | S, \Theta)}_{(4.3)} \underbrace{p(S | \pi) p(\Theta) p(\pi)}_{(4.2)}$$

Xを生成する
ための確率分布
K個、観測
モデルをデータ点
に割り当てる
確率分布

$$= \left\{ \prod_{n=1}^N p(x_n | s_n, \Theta) p(s_n | \pi) \right\} \left\{ \prod_{k=1}^K p(\Theta_k) \right\} p(\pi) \quad (4.5)$$

where $S = \{s_1, \dots, s_N\}$

(4.5) に対応するグラフ、カレモデルは以下.



今ま、後半では、混合分布における $p(x_n | \theta_k)$ と対応する事前分布 $p(\theta_k)$ を定義するとして、具体的な例を作っていく。

4.1.3 混合モデルの事後分布.

混合モデルにおける推論問題

…手元にある観測データ X にモデルを与えて

背後にあら未知の変数 S, π, θ の事後分布を計算する。

$$\sim p(S, \pi, \theta | X) = \frac{p(X, S, \pi, \theta)}{p(X)} \quad (4.6)$$

を求める。

データ X が属するクラス S の推定は、

$$p(S | X) = \sum_{\pi} p(S, \pi | X) d\pi \quad (4.7)$$

を計算すればよい。but. 理論的に計算できない。

→ 近似化項 $p(X)$ の計算が困難。

$$\begin{aligned} p(X) &= \sum_S \int_{\pi} p(X, S, \pi) d\pi \\ &= \sum_S p(X, S) \end{aligned} \quad \begin{matrix} \swarrow \text{事前分布} \\ \text{構成的積分除去可能。} \end{matrix} \quad (4.8)$$

S の分布を調べる

組合せに対して $p(X, S)$ を評価 $\rightarrow k^N$ 回の計算

4.2 破壊率分布の近似手法

近似手法論

- } キーパスサニ γ^0 リニ γ^*
} 土工場近似による変分法論.

がミニ γ^0 ILでよく利用されてます。

- ③ ベイズ学習によるアルゴリズム構成
 } データを表現するモデル
 } +
 } 対応する近似法論手法の組合せ

4.2.1. キーパスサニ γ^0 リニ γ^* .

$p(z_1, z_2, z_3)$: 破壊率分布が何かしらの実現値を得たも

~) $p(z_1, z_2, z_3)$ が z_1, z_2, z_3 の実現値を複数
 サニ γ^0 ILす。

$$z_1^{(i)}, z_2^{(i)}, z_3^{(i)} \sim p(z_1, z_2, z_3) \text{ for } i=1, 2, \dots \quad (4.9)$$

.. 分布が複雑な形状をもつ場合、可変への対応を
 困難(=サニ γ^0 リニ γ^* する)と思案。

~) キーパスサニ γ^0 リニ γ^*
 -- MCMC の手法の1つ、

$$z_1^{(i)} \sim p(z_1 \mid \underbrace{z_2^{(i-1)}, z_3^{(i-1)}}_{\text{}})$$

$$z_2^{(i)} \sim p(z_2 \mid \underbrace{z_1^{(i)}, z_3^{(i-1)}}_{\text{}}) \quad (4.10)$$

$$z_3^{(i)} \sim p(z_3 \mid \underbrace{z_1^{(i)}, z_2^{(i)}}_{\text{}})$$

すでにサニ γ^0 ILは各値の分布を条件づけ。

① $\pi = \pi^0$ (し収が十分多い場合、真の事後分布が $\pi = \pi^0$ とみなせる)。

② 実装における留意点

$Z = 1, \pi = \pi^0$ で π^0 を得る方法は 2 種類。 $Z = 0$ の他の値を設定する必要がある。

... $Z^{(1)}$ から $\pi = \pi^0$ を得た場合に、 $Z_1^{(0)}$ と $Z_2^{(0)}$ は $Z_1^{(1)}$ と $Z_2^{(1)}$ に対して等価な値を取る。

(4.10) を一般化してアルゴリズムの形式で表す。

③ 1番目、変数から $\pi = \pi^0$ で行うか、3つ目の方法を使う。

Algorithm : Gibbs

Require : Initial Sample $Z^{(0)} = (Z_1^{(0)}, \dots, Z_D^{(0)})^\top$
The number of Iteration N

Conditional distribution $p(Z_i | Z_{-i})$

Ensure : Samples $Z = \{Z^{(0)}, Z^{(1)}, \dots\}$

$Z = \{Z^{(0)}\}, T \leftarrow Z^{(0)}$

count ← 1
For $i = 1$ to N do

For $j = 1$ to D do

$s \leftarrow p(Z_j | T_{-j})$

$t_j \leftarrow s$

$Z^{(count)} \leftarrow T; Z \leftarrow Z \cup \{Z^{(count)}\}$

count ← count + 1

End For

End For

Return Z

④ キーボード入力 = $\gamma^0 \gamma^1 \dots \gamma^n$ 肉透式
複数次事後分布、相論に對して、 γ^i の事後分布 $\gamma^i = \gamma^0 \gamma^1 \dots \gamma^{i-1}$
が明確でない。

$\gamma^0, \gamma^1, \dots, \gamma^n$ の複数次事後分布を同一時間 $\gamma^0 = \gamma^1 = \gamma^n$

$$\begin{aligned} \bar{z}_1^{(i)} &\sim p(z_1 | \bar{z}_2^{(i-1)}, \bar{z}_3^{(i-1)}) \\ \bar{z}_2^{(i)}, \bar{z}_3^{(i)} &\sim p(z_2, z_3 | \bar{z}_1^{(i-1)}) \end{aligned} \quad (4.11)$$

計算コスト、面倒
+ 分量 $\gamma^0 = \gamma^1 = \gamma^n$
で $\gamma^3 = \dots = \gamma^n$ 。

崩壊型キーボード入力 = $\gamma^0 \gamma^1 \dots \gamma^n$
いくつかの変数を固定するので、モデルから除去。

$$p(z_1, z_2) = \int p(z_1, z_2, z_3) dz_3 \quad (4.12)$$

以後、固定化された分布 $p(z_1, z_2)$ に対して、通常の
キーボード入力 = $\gamma^0 \gamma^1 \dots \gamma^n$ を適用。

$$\begin{aligned} z_1^{(i)} &\sim p(z_1 | \bar{z}_2^{(i-1)}) \\ z_2^{(i)} &\sim p(z_2 | z_1^{(i)}) \end{aligned} \quad (4.13)$$

標準的なキーボード入力 = $\gamma^0 \gamma^1 \dots \gamma^n$ と比べて、サ = γ^0 にするべき。
変数の種類が少なくて、
高速に所望のサ = $\gamma^0 \gamma^1 \dots \gamma^n$ が得られることが期待される。

実用的実験状況

- (4.12) の積分計算が解析的に行える。
- 残りの各変数 $\gamma^0 \gamma^1 \dots \gamma^n$ が容易に分かる。

4.2.2. 变分推論.

变分推論 / 变分近似入

... 確率分布, 近似.

... 最適化問題を解くことによつて, 未知の確率分布の近似分布を表現を得る二つが目標.

③ 考え方.

$p(z_1, z_2, z_3)$: 確率分布 (表現 LTU, 複雑) を,

$g(z_1, z_2, z_3)$: 近似分布 (簡単) で表現.

~ KLダイバー-ジエニスを用いて最小化問題.

$$g_{\text{opt.}}(z_1, z_2, z_3) = \underset{g}{\operatorname{arg\,min}} \text{KL}\left[g(z_1, z_2, z_3) \mid p(z_1, z_2, z_3) \right] \quad (4.14)$$

... 制約条件 $\sum z_i = 1$ と $g_{\text{opt.}}(z_1, z_2, z_3) = p(z_1, z_2, z_3)$

(2つ, 7つまで).

~ g の表現能力を固定.

真の事後分布 p も, ともかく分布を最適化によつて探す.

④ 平均場近似に基づく变分推論.

近似分布 g は p の、各確率変数の独立性を仮定する.

$$g(z_1, z_2, z_3) = g(z_1) g(z_2) g(z_3) \quad (4.15)$$

... $p(z_1, z_2, z_3)$ の各変数の間に因縁関係をもつ.

⇒ 分離された近似分布 $g(z_1), g(z_2), g(z_3)$ を

KLダイバー-ジエニスが最小となるよう(?) 修正.

137p(7).

$g(z_2), g(z_3) = \text{given}$ かつて最適な $g(z_1)$ を求める

つまし。

$$g_{\text{opt.}}(\bar{z}_1) = \underset{g(\bar{z}_1)}{\arg\min} KL \left[g(\bar{z}_1) g(\bar{z}_2) g(\bar{z}_3) \parallel p(\bar{z}_1, \bar{z}_2, \bar{z}_3) \right] \quad (4.16)$$

にあたり、 $g(\bar{z}_2)$ 、 $g(\bar{z}_3)$ を固定した場合の(4.16)式を(4.17)

$$KL \left[g(\bar{z}_1) g(\bar{z}_2) g(\bar{z}_3) \parallel p(\bar{z}_1, \bar{z}_2, \bar{z}_3) \right] \quad (4.17)$$

$$= - \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \ln \frac{p(\bar{z}_1, \bar{z}_2, \bar{z}_3)}{g(\bar{z}_1) g(\bar{z}_2) g(\bar{z}_3)} g(\bar{z}_1) g(\bar{z}_2) g(\bar{z}_3) d\bar{z}_1 d\bar{z}_2 d\bar{z}_3 \quad (4.18)$$

(\because definition)

$$= - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \int_{\mathbb{R}} \ln \frac{p(\bar{z}_1, \bar{z}_2, \bar{z}_3)}{g(\bar{z}_1) g(\bar{z}_2) g(\bar{z}_3)} g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3 \right) g(\bar{z}_1) d\bar{z}_1 \quad (4.19)$$

$$= - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \int_{\mathbb{R}} \ln p(\bar{z}_1, \bar{z}_2, \bar{z}_3) g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3 \right.$$

$$- \ln g(\bar{z}_1) \int_{\mathbb{R}} \int_{\mathbb{R}} g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3$$

$$- \underbrace{\int_{\mathbb{R}} \int_{\mathbb{R}} \ln g(\bar{z}_2) g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3}_{\text{定数}} \quad (4.20)$$

$$\left. - \underbrace{\int_{\mathbb{R}} \int_{\mathbb{R}} \ln g(\bar{z}_3) g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3}_{\text{定数}} \right) g(\bar{z}_1) d\bar{z}_1 \quad (4.20)$$

$$= - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \int_{\mathbb{R}} \ln p(\bar{z}_1, \bar{z}_2, \bar{z}_3) g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3 \right. \\ \left. - \ln g(\bar{z}_1) \right) g(\bar{z}_1) d\bar{z}_1 + \text{const.} \quad (4.21)$$

$$= - \int_{\mathbb{R}} \ln \frac{\exp \left[\int_{\mathbb{R}} \int_{\mathbb{R}} \ln p(\bar{z}_1, \bar{z}_2, \bar{z}_3) g(\bar{z}_2) d\bar{z}_2 g(\bar{z}_3) d\bar{z}_3 \right]}{g(\bar{z}_1)} g(\bar{z}_1) d\bar{z}_1$$

$$+ \text{const.} \quad (4.22)$$

$$= KL \left[g(\bar{z}_1) \parallel \exp \left\{ \langle \ln p(\bar{z}_1, \bar{z}_2, \bar{z}_3) \rangle_{\bar{z}_2, \bar{z}_3} \right\} \right] + \text{const.} \quad (4.23)$$

$g(z_2), g(z_3)$ が "S23RT2T-9 (4.23)" の最小値.

$$\ln g(z_i) = \langle \ln p(z_1, z_2, z_3) \rangle_{g(z_2) g(z_3)} + \text{const.} \quad (4.24)$$

$\propto T^2$.

... 共役事前分布と上手く使った。右辺の計算結果と事前分布と同じ分布の形式は似通っている。

... 角度角差が得られた可能性がある。

$g(z_2), g(z_3)$ に対する最適化も同じ主義論。

~ 2次元の2入力分布に対する更新と結果を巡回させて、全体のKLダイバー化が最小化。

Algorithm 平均場近似による变分推論

$g(z_2), g(z_3)$ を初期化。

For $i = 1$ to MAXITER : do

$$\ln g(z_i) = \langle \ln p(z_1, z_2, z_3) \rangle_{z_2, z_3} + \text{const.}$$

$$\ln g(z_i) = \langle \ln p(z_1, z_2, z_3) \rangle_{z_1, z_3} + \text{const.}$$

$$\ln g(z_i) = \langle \ln p(z_1, z_2, z_3) \rangle_{z_1, z_2} + \text{const.}$$

End For

① 観測データ D が与えられた一般的な確率モデル $p(D, z_1, \dots, z_m)$ の事後分布(2対象3変数)の公式

D : データ集合

z_1, \dots, z_m : 未観測の変数(潜在変数, $(\theta, \gamma, \alpha, \dots)$)

$p(z_1, \dots, z_m | D)$ が2入力
 \Rightarrow 固定分布の期待値

$$p(z_1, \dots, z_m | D) = \frac{p(D, z_1, \dots, z_m)}{p(D)}$$

$g(z_{-i})$ が固定されたときに。

$$\ln g(z_i) = \langle \ln p(z_1, \dots, z_m | D) \rangle_{g(z_{-i})} + \text{const.}$$

$p(D) g(z_i)$

$$= \langle \ln p(D, z_1, \dots, z_m) \rangle_{g(z_{-i})} + \text{const.} \quad (4.25)$$

- ・変分推論では、分解を仮定して複数の粗肉が混ざる場合がある。
 - ～) 素質が十分収束しても、複数の粗肉をもつよければ分布には複数の包、この2つをもつてくなる。
- ・変分推論、ほうがギガ入力で70%以上も高速。
 - ～) 大規模な確率モデルには複数の優れた収束性能。

④構造化変分推論

(例)

$$p(z_1, z_2, z_3) \approx q(z_1) q(z_2, z_3)$$

…独立性を仮定してより簡単に計算する結果。

…×MRI、計測コントローラー