第５章　情報検索とベイズ決定

§5.3 ベイズ検索
§5.4 ロジット分析

05T4074A
久保田　敦

目次

- §5.3　ベイズ検索
  - ベイズ検索について
  - 検索例
- §5.4　ロジット分析
  - ロジット分析
  - 対数オッズ
  - 対数例
- まとめ

§5.3 ベイズ検索（1）

- 各カテゴリーS_{ik}の確率θ_{ik}を推定する

母数 $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$

θ の事前確率分布：
$$w(\theta) \propto \prod_{i=0}^{p} (\theta_{i0})^{\alpha_{i0}-1} (\theta_{i1})^{\alpha_{i1}-1}$$

データrの確率分布：
$$f(r \mid \theta) = \prod_{i=0}^{p} \frac{n!}{r_{i0}!\, r_{i1}!} (\theta_{i0})^{r_{i0}} (\theta_{i1})^{r_{i1}}$$

これらの積から（1.7.3）のようにθ の事後確率分布は

$$w(\theta \mid r) \propto \prod_{i=0}^{p} (\theta_{i0})^{\alpha_{i0} r_{i0} - 1} (\theta_{i1})^{\alpha_{i1} r_{i1} - 1}$$

よって母数は

$$\alpha \rightarrow \alpha + r$$

§5.3　ベイズ検索(2)

事後確率分布と(5.2.7)より

$$E\left(\frac{\theta_{i1}}{\theta_{i0}}\right) = \frac{\alpha_{i1} + r_{i1}}{\alpha_{i0} + r_{i0} - 1} \begin{cases} \geqq c & \text{:文献カテゴリーSiを検索} \\ < c & \text{:検索しない} \end{cases}$$

$$c = \frac{1}{k}$$

k:関連文献1点を得るために誤って調べることが許される無関連文献の数

§5.3（例） 図書館での検索２

- 文献カテゴリーS1を検索するか判別する

α10=2,α11=1,k=5

$$\frac{\alpha_{i1}+r_{i1}}{\alpha_{i0}+r_{i0}-1}$$ の式に代入

$$\frac{1+r_{11}}{1+r_{10}} \geq \frac{1}{5}$$

$$r_{10} \leq 5r_{11}+4$$

よって、無関連文献点数≦5×関連文献点数+4
が成立するときS1を検索する。

§5.4 ロジット分析

- 確率比の変換　（e_{ik}は誤差項）

$$\log\frac{\theta_{ik}}{\theta_{i0}} = \sum_{j=1}^{m} \beta_{jk} Z_{ij}(s) + e_{ik}$$

ウエイト付けの回帰モデルにすることで、Aの指定に条件がかかり事前分布の指定が容易になる。

オッズ： $\dfrac{\theta_{ik}}{\theta_{i0}}$

対数オッズ： $\lambda_{ik} = \log\dfrac{\theta_{ik}}{\theta_{i0}}$

対数オッズ

- 確率比の行列表現

$$\boldsymbol{\lambda} = \mathbf{Z}\beta + \mathbf{E}$$

- 期待値

$$E(\boldsymbol{\lambda}) = \mathbf{Z}E(\beta)$$

$$E(\boldsymbol{\lambda} \mid \mathbf{r}) = \mathbf{Z}E(\beta \mid \mathbf{r}) + E(\mathbf{E} \mid \mathbf{r})$$

最小二乗法より正規方程式を得る。

$$\mathbf{Z}'\mathbf{Z}E(\beta \mid \mathbf{r}) = \mathbf{Z}'E(\boldsymbol{\lambda} \mid \mathbf{r})$$

すると、ウエイトβの推定は事後的期待値として求まる。

$$E(\beta \mid \mathbf{r}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E(\boldsymbol{\lambda} \mid \mathbf{r})$$

$$\boxed{E(\boldsymbol{\lambda})} \Rightarrow \boxed{\alpha} \Rightarrow \boxed{E(\boldsymbol{\lambda} \mid \mathbf{r})}$$

対数オッズ(例)(1)

表5.5　標本の要約

関連性	文献No.	分類子
0	1	2,3
0	3	3
0	9	3
0	10	2
1	4	3
1	5	2
1	6	2,3
2	2	1,2,3,4
2	7	1,2,3
2	8	2,3

表5.3　関連性のオッズ及びウエイトの期待値

分類子		段階1		段階2	
		オッズ	ウエイト	オッズ	ウエイト
[1]	閲覧	3	1.1	5	1.6
[2]	OR	5	1.6	2	0.7
[3]	図書館	5	1.6	5	1.6
[4]	閲覧･OR	2	0.7	10	2.3

表5.6　可能な文献のカテゴリー

文献カテゴリー	閲覧	OR	図書館	閲覧･OR
1	0	1	0	0
2	0	0	1	0
3	0	1	1	0
4	1	1	1	0
5	1	1	1	1

対数オッズ(例)(2)

◆母数

$$\alpha = \begin{array}{cl} \text{関連性} & \\ \begin{pmatrix} 1 & 6 & 3 \\ 1 & 6 & 6 \\ 1 & 13 & 9 \\ 1 & 17 & 15 \\ 1 & 20 & 24 \end{pmatrix} & \begin{array}{l} (\text{カテゴリー}1) \\ (\text{カテゴリー}2) \\ (\text{カテゴリー}3) \\ (\text{カテゴリー}4) \\ (\text{カテゴリー}5) \end{array} \\ (0)\ (1)\ (2) & \end{array}$$

◆標本のカテゴリー分け(データr)

文献カテゴリー	関連性(段階)		
	段階0	段階1	段階2
1	1[10]	1[5]	0
2	2[3,9]	1[4]	0
3	1[1]	1[6]	1[2]
4	0	0	1[7]
5	0	0	1[8]

対数オッズ(例)(3)

◆ベイズ更新後のオッズの期待値

オッズ：$\dfrac{\theta_{ik}}{\theta_{i0}}$

(k=1,2 i=1,2,・・・,5)

文献カテゴリー	関連性のオッズ(期待値)	
	段階1	段階2
1	1.5	0.5
2	0.9	0.8
3	2.2	1.8
4	3.4	3.3
5	3.5	3.8

よって、$\mathrm{E}(\lambda \mid \mathrm{r})$が求まった。

よって、各分類子から
関連性を指定するウエイト
が推定できる。

$$
E(\beta \mid r) = \begin{array}{c|l}
\text{関連性} & \\
\begin{pmatrix} 1.1 & 1.7 \\ 1.4 & 0.7 \\ 0.9 & 1.0 \\ 0.2 & 0.5 \end{pmatrix} & \begin{array}{l} \text{(閲覧)} \\ \text{(OR)} \\ \text{(図書館)} \\ \text{(閲覧・OR)} \end{array} \\
\text{(0) (1)} &
\end{array}
$$

対数オッズ(例)(4) -分類子の追加-

◆追加する分類子

追加分類子		関連性	
		段階1	段階2
I5	‘単行本’	0.6	0.7
I6	‘重複’	0	0.7
I7	‘有効性指標’	0	0.7

関連性のオッズは

$$\log\left(\frac{\theta_{i1}}{\theta_{i0}}\right) = 1.1Z_{i1}(s) + 1.4Z_{i2}(s) + 0.9Z_{i3}(s) + 0.2Z_{i4}(s) + 0.6Z_{i5}(s)$$

$$\log\left(\frac{\theta_{i2}}{\theta_{i0}}\right) = 1.7Z_{i1}(s) + 0.7Z_{i2}(s) + 1.0Z_{i3}(s)$$

$$+ 0.5Z_{i4}(s) + 0.7Z_{i5}(s) + 0.7Z_{i6}(s) + 0.7Z_{i7}(s)$$

で推定できる。

まとめ

- 関連性のオッズが求まれば、後は(5.3.4)の式で検索ができる。
- ロジット分析によって文献カテゴリーそのものの関連性が自動的に判定できるので文献検索を大幅に自動化できる。