入門ベイズ統計
第５章　情報検索とベイズ決定

10月10日

田中洸一

項目

- 問題提起
- ５．１　分類子と関連性
- ５．２　文献からのサンプリング

問題提起

図書館などの情報検索のシステムにベイズ決定を用いることができる

↓

情報の関連性の有無は、「確率」が絡む

↓

情報検索システムに応用するための手法を提示

５．１　分類子と関連性

・情報の関連性の有無に二つの基準を定義

――――――――――――――――――――――――――――――――――――

例：図書館での検索１

「オペレーションズ・リサーチ」の手法を「図書館」の「閲覧」の問題に応用するための文献を検索すことを考える。

・以下の四つの概念用語を用意

I1：’閲覧’　I2：’OR’　I3：’図書館’　I4：’閲覧’および’OR’

――――――――――――――――――――――――――――――――――――

・概念用語を文献の関連性の有無の基準とする。

・概念用語を分類子と表現し、m個の分類子$\mathrm{I}_1,\mathrm{I}_2,\mathrm{L}\ ,\mathrm{I}_m$を定義

５．１　分類子と関連性

- 文献$s \in S$に対し、関数$Z_j(s)$

$$Z_j(s) = \begin{cases} 1 & (s\text{に分類子}\mathrm{I}_j\text{が与えられるとき}) \\ 0 & (\text{与えられないとき}) \end{cases} \qquad (5.1.1)$$

$$(j = 1, 2, \mathrm{L}\ , m)$$

例：（1, 0, 0, 1, 1,L 1, 0）（m桁）

文献sにm桁の２進数$\left(Z_1(s), Z_2(s), \mathrm{L}\ , Z_m(s)\right)$が与えられる

どんな文献sも分類子$\mathrm{I}_1, \mathrm{I}_2, \mathrm{L}\ , \mathrm{I}_m$によって$p\ (=2^m - 1)$個の

どれかの部分集合S_i（カテゴリー）に分類され　る

５．１　分類子と関連性

- 関連性の段階的尺度 $\{0,1,2,\mathrm{L}\ ,q\}$　（かなり主観的な基準）

例：ｑ＝１　→　０（関連なし），１（関連あり）

　　ｑ＝２　→　０（関連なし），１（やや関連あり），２（関連あり）

・２つの基準により、各文献ｓに対し、たとえば以下のようなベクトルを表現

$$(1,0,0,1,1,\mathrm{L}\ ,1,0,5)\quad (m+1\text{桁})$$

$$Z_{m+1}(s)=0,1,2,\mathrm{L}\ ,q$$

全文献Ｓの集合　$S=\bigcup_i\bigcup_k S_{ik}$　(集合の直和)　(5.1.3)

$$S_{ik}=\left\{s\middle|\left(Z_1(s),Z_2(s),\mathrm{L}\ ,Z_m(s)\right)=i,Z_{m+1}(s)=k\right\}$$

目的は分類子によるカテゴリーを検索すべきか否かを決定する

$$S_i=\bigcup_k S_{ik}\quad (5.1.4)$$

５．２　文献からのサンプリング

- 集合Sをランダム・サンプリングの問題として考える

各カテゴリーS_{ik}の確率　$\theta_{ik} = P(S_{ik})$　　(5.2.1)

i番目の文献カテゴリ　S_i　の確率$P(S_i) = \sum_{k=0}^{q} \theta_{ik}$　　(5.2.2)

（※$\theta_{ik} \geq 0, \sum_{i=0}^{p} \sum_{k=0}^{q} \theta_{ik} = 1$）

- 標本（ランダムサンプル）n点の文献よりrを表現

$$r = (r_{00}, r_{01}, \mathrm{L}\ , r_{0q}, r_{10}, \mathrm{L}\ , r_{pq}) \qquad (5.2.4)$$

rの確率分布　$f(r|\theta) = \prod_{i=0}^{p} \prod_{k=0}^{q} \frac{n!}{r_{ik}!} (\theta_{ik})^{r_{ik}}$　（二項分布）　(5.2.5)

5．2　文献からのサンプリング

表 5.2　ある図書館の文献構成

文献カテゴリー	閲覧	OR	図書館	閲覧・OR	関連性	確率	標本データ
S_{00}	×	×	×	×	なし	θ_{00}	r_{00}
S_0 S_{01}	×	×	×	×	ややあり	θ_{01}	r_{01}
S_{02}	×	×	×	×	あり	θ_{02}	r_{02}
S_{10}	×	×	×	○	なし	θ_{10}	r_{10}
S_1 S_{11}	×	×	×	○	ややあり	θ_{12}	r_{11}
S_{12}	×	×	×	○	あり	θ_{12}	r_{12}
S_{20}	×	×	○	×	なし	θ_{20}	r_{20}
…	…	…	…	…	…	…	…
$S_{15,2}$	○	○	○	○	あり	$\theta_{15,2}$	$r_{15,2}$
S						1	n（点）

各カテゴリーの確率とランダム・サンプルの結果が与えられている．各カテゴリーの定義は該当（○），非該当（×）で定められている．

５．２　文献からのサンプリング

- 関連性が2段階ならq＝１で

$$f(r\mid\theta)=\prod_{i=0}^{p}\frac{n!}{r_{i0}!\,r_{i1}!}(\theta_{i0})^{r_{i0}}(\theta_{i1})^{r_{i1}} \qquad (5.2.6)$$

- 文献カテゴリーS$_i$の検索基準　（関連性ありθ_{i1},なしθ_{i0}）

$$\frac{\theta_{i1}}{\theta_{i0}}\geq c\text{なら, 文献カテゴリー}S_i\text{を検索}$$

(5. 2. 7)

$$\frac{\theta_{i1}}{\theta_{i0}}<c\text{なら, 文献カテゴリー}S_i\text{を検索しない}$$

（※c:無関係文献を　検索してしまうコストによ　る定数）

すべてのθは値がわかっているとして計算、

そうでなければデータrから推定