レポート提出票

科目名: 情報工学実験2

実験テーマ: 実験テーマ5 教育システム設計

実施日: 2020年 11月 2日

学籍番号: 4619055

氏名: 辰川力駆

共同実験者:

1　要旨

ベイズの定理は既に習っているが、最尤推定を行う場合の応用のされ方について、演習を通して学習する。今回は、ある大学の男女別や学部別の在籍者数から推定する実験を行う。

2　目的

統計モデルを用いた分析は、例えば商品の推薦や迷惑メールの削除機能など、身近な機能を支える基本的な技術となっている。本実験では、このような統計モデルを用いた分析に欠かせない、パラメタの推定の方法について、基本的な技術を習得することを目的とする。

3　理論

3.1　ベイズの定理

ベイズの定理は以下の式で与えられる。

$$P(A|B) = \frac{P(A)}{P(B)}P(B|A) \tag{1}$$

ここで、$P(A)$ と $P(B)$ はそれぞれ事象 A と B が起こる確率、$P(A|B)$ は事象 B が起こったときに事象 A が起こる確率を表す条件付き確率である。

B が起こったときに A が起こる確率 $P(A|B)$ は B が起こった事象中で A も同時に起こっている事象の割合である。つまり以下のように表すことができる。

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \tag{2}$$

また、これらの関係は A と B を入れ替えても同様に成立するため、以下を得る。

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} \tag{3}$$

これらの 2 つより、ベイズの定理が導出される。

3.2　最尤推定

このベイズの定理を使用することで、ある確率変数の値から別の確率変数の値を推定することが可能である。例えば、ある大学の男女別学部別の在籍者数が表 1 のようであったとする。

表 1: ある大学の学部別、男女別在籍者数

学部	学生数	
	男子学生数	女子学生数
理学部第一部	2230	611
理学部第二部	1271	390
薬学部	492	547
工学部	1771	417
工学部第二部	701	148
理工学部	4289	870
基礎工学部	1028	372
経営学部	950	441
学部合計	12732	3796

ここで「この大学に所属する女子学生の学部」を推定することを考える。この時、「女子学生のそれぞれの学部に所属している確率」が最も高い学部に所属してると考えることが自然である。つまり、以下のような数式を考えることが自然である。

$$\hat{(\text{学部})} = \arg\max_{\text{学部}} P(\text{学部} \mid \text{女子学生}) \tag{4}$$

(この時 $\arg\max_x$ は x を変数とみなし、後に続く式が最も大きくなる x の値を返すことを意味する。) この考え方は、事後分布最大化 (Maximum a Posteriori:MAP) 推定と呼ばれる。それぞれの学部について計算を行うと以下のようになる。

$$\begin{aligned} P(\text{理学部第一部} \mid \text{女子学生}) &= \frac{611}{3796}, \\ P(\text{理学部第二部} \mid \text{女子学生}) &= \frac{390}{3796}, \\ &\vdots \quad \vdots \\ P(\text{経営学部} \mid \text{女子学生}) &= \frac{441}{3796} \end{aligned}$$

また、式 4 はベイズの定理を用いると以下のようにもかける。

$$\begin{aligned} \hat{(\text{学部})} &= \arg\max_{\text{学部}} P(\text{学部} \mid \text{女子学生}) \\ &= \arg\max_{\text{学部}} \frac{P(\text{学部})}{P(\text{女子学生})} P(\text{女子学生} \mid \text{学部}) \end{aligned} \tag{5}$$

これらをそれぞれの学部について計算すると以下のようになる。

$$
\begin{aligned}
\frac{P(\text{理学部第一部})}{P(\text{女子学生})}P(\text{女子学生}\,|\,\text{理学部第一部}) &= \frac{\frac{2230+611}{12732+3796}}{\frac{3796}{12732+3796}} \times \frac{611}{2230+611}, \\
\frac{P(\text{理学部第二部})}{P(\text{女子学生})}P(\text{女子学生}\,|\,\text{理学部第二部}) &= \frac{\frac{1271+390}{12732+3796}}{\frac{3796}{12732+3796}} \times \frac{390}{1271+390}, \\
&\vdots \quad \vdots \\
\frac{P(\text{経営学部})}{P(\text{女子学生})}P(\text{女子学生}\,|\,\text{経営学部}) &= \frac{\frac{950+441}{12732+3796}}{\frac{3796}{12732+3796}} \times \frac{441}{950+441}
\end{aligned}
$$

こちらの式は計算の結果が全く変わらないことがわかる。

ここで、式 5 をよく見れば、$P(\text{女子学生})$ は学部に関係のない定数となっているため、以下のように考えても推定される学部は変化しない。

$$
\frac{P(\text{学部})}{P(\text{女子学生})}P(\text{女子学生}\,|\,\text{学部}) \propto P(\text{学部})P(\text{女子学生}\,|\,\text{学部}) \tag{6}
$$

したがって、式 4 は以下のようになる。

$$
\begin{aligned}
\hat{(\text{学部})} &= \arg\max_{\text{学部}} P(\text{学部}\,|\,\text{女子学生}) \\
&= \arg\max_{\text{学部}} P(\text{学部})P(\text{女子学生}\,|\,\text{学部}) \qquad (7)
\end{aligned}
$$

このように、確率ではないが、確率に比例するスコア：尤度を用いて推定を行うこともできる。この尤度が最大のものを推定値として採用する推定法が最尤推定 (Maximum likelihood estimation:MLE) である。

4 課題

4.1 課題 1-1

この大学にある女子学生がいた場合、その学部を推定する。

表 2: 各学部ごとの推定

学部	頻度による推定	尤度推定
	P(学部 | 女子学生)	P(学部)P(女子学生 | 学部)
理学部第一部	0.161	0.037
理学部第二部	0.103	0.024
薬学部	0.144	0.033
工学部	0.110	0.025
工学部第二部	0.039	0.009
理工学部	0.229	0.053
基礎工学部	0.098	0.023
経営学部	0.116	0.027

ある女子学生の学部を推定するために、まずは女子学生のそれぞれの学部に所属している確率を求めた。結果は表 2 にまとめた。縦の合計が 1 になっていないのは、小数第 4 位を四捨五入しているからである。また、尤度推定も行った。計算式は表 2 の通りである。

したがって表 2 より、確率が最も大きい学部は理工学部である。よって、この大学にある女子学生がいた場合理工学部である可能性が高い。

今回は、尤度推定と頻度による推定の両方が行えたが、これらの違いは使っている情報が違うことである。尤度推定は、例えば P(学部) と P(女子学生 | 学部) しか分かっていない時に有効である。

4.2 課題 1-2

この大学にある経営学部学生がいた場合、その性別を推定する。

課題 1-1 と同様に考えて、

$$\begin{aligned} \hat{(\text{性別})} &= \arg\max_{\text{性別}} P(\text{性別} \mid \text{経営学部生}) \\ &= \arg\max_{\text{性別}} \frac{P(\text{性別})}{P(\text{経営学部生})} P(\text{経営学部生} \mid \text{性別}) \end{aligned} \tag{8}$$

となることから、それぞれの性別について計算する。

$$\frac{P(\text{男子学生})}{P(\text{経営学部生})}P(\text{経営学部生} \mid \text{男子学生}) = 0.683$$
$$\frac{P(\text{女子学生})}{P(\text{経営学部生})}P(\text{経営学部生} \mid \text{女子学生}) = 0.317$$

したがって、男子学生である可能性が高い。

4.3 課題 1-3

公開された問題を 4 題以上ランダムに選び解答する。

表 3: 解いた問題とその正誤判定

問題番号	正誤
3	○
9	○
21	○
24	○
30	○
35	○
40	○
51	○

解答した結果を表 3 にまとめた。次に、解いた思考過程について説明する。

問 3

条件式で、音楽の話をしているのは「読書が好きでない人は音楽が好きである」しか存在しない。対偶を取ると「音楽が好きでない人は読書が好きである」となる。よって、音楽が好きな人の話をしている条件は存在しないので、確実には言えないのは「音楽が好きな人はスポーツが好きではない」と分かる。

問 9

まずは話に多く上がっている人について 100 点を取っているかどうかを考えた。C が 100 点だと仮定すると、B,C,D,F の 4 人が本当のことを言っていることとなるので不適である。次に、H が 100 点だと仮定すると、B,D,G の 3 人が本当のことを言っていることとなるので正しい。よって、H が 100 点である。

問 21

「歴史」は「180511091909」、「物理」は「022120211809」と表されていることから数字 4 つごとに区切って前の 2 つと後の 2 つでアルファベットの ABC 順の番号に対応していると分かる。例えば、「歴史」の数字の先頭 4 つ「1805」は ABC 順で 18 番目は「r」で、5 番目は「e」であることから「れ」である。

これを用いて、「05041510090401091415202107 09」を解読すると、「江戸時代の次」となるので、答えは明治時代である。

問 24

表と「DICTIONARY」は「JNFXNVRAQY」、「EIGHT」は「ONEIX」と表されることから、180 度回転している場所も同じような動きをすることが分かる。例えば、「I」は「N」なので「Q」は「L」である。

これを用いて「LUCKY」を変換すると「HKFDY」となる。

問 30

「小春日和」は「CYFMVSLEOYVE」と表されることから、アルファベットの T を中心に対称に変換されていることが分かる。

これを用いて「FMVSZYZMZMCSUMFM」を変換すると、「HARUNONANAKUSAHA」となるので春の七草のなずなが正解である。

問 35

すべての条件から当てはまる部屋の割り当ては

$$\begin{pmatrix} F & E & \text{空} \\ \text{空} & A & C \\ B & \text{空} & D \end{pmatrix}, \begin{pmatrix} F & E & \text{空} \\ \text{空} & A & D \\ B & \text{空} & C \end{pmatrix}$$

のどちらかである。よって、確実にいえるのは「F は 301 に住んでいる」である。

問 40

これは問 35 をさらに複雑化しただけである。条件より、B と E に関しては入れ替えても成り立つので、α を B または E とすると

$$\begin{pmatrix} \alpha & C & \text{空} & I \\ A & \text{空} & \alpha & K \\ \text{空} & H & D & L \\ G & J & F & \text{空} \end{pmatrix}, \begin{pmatrix} \alpha & A & \text{空} & I \\ C & \text{空} & \alpha & K \\ \text{空} & G & D & L \\ H & J & F & \text{空} \end{pmatrix}, \begin{pmatrix} C & \text{空} & I & \alpha \\ \text{空} & \alpha & K & A \\ H & D & L & \text{空} \\ J & F & \text{空} & G \end{pmatrix}$$

のどれかとなる。よって、正しくいえることは「I の左隣は空室である」である。

問 51

まず、予選 1 位を決めるためのトーナメントは $400-1=399$ 試合行われる。次に、予選 2 位を決めるためのトーナメントは $399-1=398$ 試合行われる。最後に決勝戦で 1 試合するので、合計 $399+398+1=798$ であるから、798 試合である。

4.4 課題 1-4

課題 1-3 で解いた問題群を、1 問あたり (100/問題数) 点の 100 点満点のテストとし、自分の偏差値と順位を求める。(平均点 60、標準偏差 20、1000 人中の順位で、1000 人の得点分布は正規分布に従っているとする。)

偏差値は、

$$S_i = \frac{10(x_i - u_x)}{\sigma_x} + 50$$

と表せるので、自分の偏差値は、

$$\frac{10(100-60)}{20} + 50 = 70$$

である。

また、得点分布は正規分布に従っているので、自分の順位は NORM.DIST 関数を用いて計算すると 22.75... となった。小数点は切り捨てるので、順位は 22 位である。

4.5　課題 1-5

世の中にいる「あゆみ」という名前の人の男女比が 1:9 だとする。この大学に「あゆみ」という人が 1 名所属していることがわかっているとき、「あゆみ」さんの学部と性別を上位 3 つまで推定する。

5　まとめ

参考文献