テキストデータの統計科学入門

第8章　テキストにおける確率モデリング

茨城大学
國井慎也

概要

- テキスト分析のための確率の基本的な概念と主に使用される確率分布、その確率分布を用いた簡単なモデリングを紹介する。

確率

- 標本空間

  -起こりうるすべての事象の集合

  Ω＝{A1,A2,･･･,An}

- 確率

  N:試行の回数

  xi:N回の試行で事象Aiが起こる回数

$$P\{A_i\} \fallingdotseq \frac{x_i}{N}$$

確率の主な定理

- 乗法定理

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

P(A∩B):事象A,Bが同時に起きる確率

P（A|B）:Bが起きた条件の下でのAが起きる確率

- ベイズの定理

-乗法定理から導かれる定理

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

P(A):事象Aの事前確率

P(A|B)：事象Aの事後確率

確率分布

- 確率変数

  -試行の標本空間の事象の変数

- 確率分布

  -確率変数とその変数に対応する確率のセット

$$P\left(X=x_i\right)=P\left(A_{xi}\right)$$

$$\sum_i P\left(X=x_i\right)=1$$

確率分布の例（6面のサイコロ）

確率変数	1	2	3	4	5	6
確率	1／6	1／6	1／6	1／6	1／6	1／6

主な分布（1）

- 二項分布

  -成功と失敗、男性と女性など試行の結果が2通りある問題をモデル化するときに用いられる

$$P(X=x)={}_nC_x\,p^x(1-p)^{n-x}$$

$$E(X)=np\ \ ,\ Var(X)=np(1-p)$$

- 幾何分布

  -独立な試行を繰り返して、x回失敗してx+1回目に成功する確率を考えた分布

$$P(X=x)=pq^x\ \ (x=0,1,2,\cdots,\ 0<p<1\ ,\ q=1-p)$$

$$E(X)=\frac{1-p}{p}\ \ ,\ Var(X)=\frac{1-p}{p^2}$$

主な分布(2)

- ポアソン分布

  -一定期間内で、平均λで起きる事象がx回起きる確率を考えた分布

$$P(X=x)=e^{-\lambda}\frac{\lambda^{x}}{x!}$$

- 正規分布

  -平均値に集積するようなデータの分布に関する確率分布

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

特にμ＝0、σ＝1の場合を標準正規分布と呼ぶ

主な分布(3)

- 対数正規分布

  -正規分布の確率変数に対数を取ったY=log(X)がN(μ,σ)に従うとき、確率変数Xが従う分布

$$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}e^{\frac{-(\log(x)-\mu)^2}{2\sigma^2}}$$

$$E(X)=e^{\mu+\frac{\sigma^2}{2}}\ ,\ Var(X)=e^{2\mu+\sigma^2}-e^{2\mu+\sigma^2}$$

モデリングの例(1)

- 英文の手紙における単語を、音節を単位として調べ、単語の長さの頻度分布を幾何分布に当てはめる

単語の長さの度数表

音節xi	度数fi	相対度数pi
1	3978	0.7613
2	831	0.1587
3	281	0.0537
4	121	0.0231
5	15	0.0029
6	6	0.0004
合計	5237	1

幾何分布　$P(X=x)=pq^{x}$

1.実測値の平均を求める

$$\bar{x}=\sum_{i=1}^{n} x_i p_i = 1.3491$$

2.幾何分布のp,qを求める

$$p=\frac{1}{\bar{x}}=\frac{1}{1.3491}=0.7412\ ,\ q=1-0.712=0.2588$$

モデリングの例(2)

3.音節の推測相対度数を求める

$$P_1 = p(1-p)^{1-1} = 0.7412 \times 0.2588^0 = 0.7412$$

$$P_2 = p(1-p)^{2-1} = 0.7412 \times 0.2588^1 \fallingdotseq 0.1918$$

4. 推測度数を求める

$$NP_1 = 5237 \times 0.7412 \fallingdotseq 3882$$

$$NP_2 = 5237 \times 0.1918 \fallingdotseq 1002$$

音節xi	実測度数fi	推測度数NPi	推測相対度数Pi
1	3987	3882	0.7412
2	831	1002	0.1918
3	281	260	0.0496
4	121	76	0.0128
5	15	17	0.0033
6	2	5	0.0009
合計	5237	5235	1