ベイズ統計概論

芝浦工業大学　数理科学研究会
BV19083　久保田　静希

2019 年 11 月 1 日

目 次

研究背景

近年, コンピューターの発展により従来ではできなかった複雑な計算を誰もがたやすくできるようになった. これにより統計学は一部の人間のみが扱える高度な学問からパソコンをもっている者なら誰でも扱えるとても利便性の高いツールとなった. その恩恵を多く受けたのが「ベイズ統計学」である. 本発表ではベイズの統計学の基盤となっている「ベイズの定理」から「ベイズ更新」等について, 従来の統計学とどのように異なるのかについて考察を行う.

1　ベイズの定理

1.1　ベイズの定理

定理 1.1. ベイズの定理とは任意の事象 A,B に対し

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

が成り立つことである. ここで $P(A)$ とは事象 A が起こる確率を表し,$P(B|A)$ は A が起きた時に B が起きる条件付き確率を表す.

1.2　ベイズの定理〜証明〜

条件付確率の定義より

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

両辺 $P(B)$ 掛けると

$$\begin{aligned} P(A \cap B) =& P(A|B)P(B) \\ =& P(B|A)P(A) \end{aligned}$$

よって

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

1.3　ベイズの公式の解釈

ベイズの基本公式の $P(H|D)$ は D が起きた時その原因が H であった時の確率であるが, 実際は考えられる原因は一つとは限らない.そこで考えられる原因が n 個あったとしそれぞれ $H_1, H_2, \ldots H_n$ とする.この時 H_1 が原因であった時の確率は $P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)}$ と表せられる.またこれらの原因が全て排反であったとき, $P(D) = P(D \cap H_1) + \cdots + P(D \cap H_n)$ と表せる.ここで「確率の乗法定理」より $P(D) = P(D|H_1)P(H_1) + \cdots + P(D|H_n)P(H_n)$ と表すことができる.これをを先ほどの式に代入し一般化すると

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D|H_1)P(H_1) + \cdots + P(D|H_n)P(H_n)}$$

この式がベイズの統計学において鍵を握ってくる重要な式である.また, 以上のことからベイズの定理とはこの式を指すものとする.

1.4　ベイズの定理の用語と意味

ベイズの定理の一部確率には名前が付けられている.

記号	名称	意味
$P(H_i)$	事前確率	原因 H_i が起こる確率
$P(D\|H_i)$	尤度	原因 H_iが起きた時 D が得られる確率
$P(H_i\|D)$	事後確率	D が得られたとき H_iが原因の確率

1.5　例題 1

ある地域の気象統計では, 4 月 1 日に晴れ, 曇り, 雨の確率はそれぞれ 0.3,0.6,0.1 である.
翌 2 日に雨である確率は前日が晴れなら 0.2, 曇りなら 0.5, 雨なら 0.4 である. この地方で 2 日が雨の時前日が曇りの確率をもとめよ.

回答

以下の様に記号を定義する.

記号	意味
H_1	1 日目が晴れ
H_2	1 日目が曇り
H_3	1 日目が雨
D	二日目が雨

2 日が雨の時前日が曇りの確率はベイズの定理より

$$P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + P(D|H_3)P(H_3)}$$

となる.
問題文より $P(H_1) = 0.3$, $P(H_2) = 0.6$, $P(H_3) = 0.1$ また,
$P(D|H_1) = 0.2$, $P(D|H_2) = 0.5$, $P(D|H_3) = 0.4$ とわかるので代入すると

$$\begin{aligned} P(H_2|D) &= \frac{0.5 \times 0.6}{0.2 \times 0.3 + 0.5 \times 0.6 + 0.4 \times 0.1} \\ &= \frac{3}{4} \end{aligned}$$

よって 75 %

2　ベイズ更新

2.1　ベイズ更新とは

ベイズ更新とは従来の母数を重要視するために新たなデータが与えられたとき, 計算をやり直す統計とは違い与えられた複数のデータを 1 つずつ処理していくベイズ統計学の醍醐味である.具体的には, あるデータ D_1, D_2とその原因 H_1, H_2が与えられたとする.このとき, 2 つのデータ

が与えられた原因がH_1である確率を求めたいとする.このような場合まずはD_1のみを用いてベイズの定理を利用し事後確率 $P(H_1|D_1)$ を求める.次に D_2を用いてベイズの定理を使うのだがこの時, 事前確率 $P(H_1)$ の代わりに先ほどの事後確率 $P(H_1|D_1)$ を用いて計算する.このような計算方法を行うことによって複数のデータに対し 1 つずつ処理を行うことができる.この性質はデータ処理をするソフトウェアを作成するときに大変便利である.

2.2 理由不十分の原則

実際に日常でベイズ更新を用いようとするとき, 一番最初の事前確率についての条件がないことが多々ある. しかし問題が不厳密であるからと言って問題を投げ出すことはできない. よってベイズの統計学では「特に条件がなければ確率は同等だろう」という常識を用いて計算を始める. ここが他の統計学との大きな差である.

2.3 ベイズ更新における逐次合理性

基本的にベイズの定理では複数のデータが与えられたとき, そのデータをどのようなが順番並べ替えてもデータが同じならば解析順序はこだわらないという性質がある. これをベイズの理論の「逐次合理性」という.

2.4 ベイズ更新における逐次合理性～証明～

任意の二つのデータ D_1, D_2 に対し事後確率を

$$P(H|D_1) = \frac{P(D_1|H_i)P(H_i)}{P(D_1)}$$
$$P(H|D_2) = \frac{P(D_2|H_i)P'(H_i)}{P(D_2)}$$

とする.
D_1, D_2 の順にデータが得られたとき

$$\begin{aligned} P(H|D_2) &= \frac{P(D_2|H)P'(H)}{P(D_2)} \\ &= \frac{P(D_2|H)P(H|D_1)}{P(D_2)} \\ &= \frac{P(D_2|H)}{P(D_2)} \times \frac{P(D_1|H)P(H)}{P(D_1)} \\ &= \frac{P(D_2|H)P(D_1|H)P(H)}{P(D_2)P(D_1)} \end{aligned}$$

D_2, D_1 の順にデータが得られたとき

$$\begin{aligned} P(H|D_1) &= \frac{P(D_1|H)P(H)}{P(D_1)} \\ &= \frac{P(D_1|H)P(H|D_2)}{P(D_1)} \\ &= \frac{P(D_1|H)}{P(D_1)} \times \frac{P(D_2|H)P'(H)}{P(D_2)} \\ &= \frac{P(D_2|)P(D_1|H)P'(H)}{P(D_2)P(D_1)} \end{aligned}$$

ここで $P(H)$ と $P'(H)$ はともに事前確率なので

$$P(H|D_1) = P(H|D_2)$$

となる. よって逐次合理性がなりたつ.

2.5 事前確率の重要性について

ベイズ理論においても，日常においても事前確率は重要なものになっている．例えば，ある事故で「軽傷者が 10 名死亡，重傷者が 7 名死亡」という文が流れたとき，重傷者よりも軽傷者のほうが死者数が多い. この一文だけでは不思議に思う方がいるかもしれないがそもそもの軽傷者と重傷者の数が 100 人と 10 人だった場合，軽傷者の死亡率は 0.1，重傷者の死亡率は 0.7 となる．事前確率的な情報をないがしろにすると本質を見失ってしまうかもしれない.

2.6 例題２

ある宝箱を売る会社 A,B がある.
会社 A では真珠とガラス玉の割合が 3 : 1, B 社では 1 : 3 の割合で宝箱が売られている. 外見では区別はつかない.
いまここに A 社製か B 社製か不明な宝箱がある. この中には十分に玉が入っているとする. この中から続けて球を 3 つ取り出したところ, 順に真珠, 真珠, ガラス玉の割合であった. この時この箱が A 社製である確率を求めよ.

回答

以下の記号を定義する.

記号	意味
H_A	取り出した一個の玉が A 社製の宝箱からである
H_B	取り出した一個の玉が B 社製の宝箱からである
S	取り出した一個の玉が真珠である
G	取り出した一個の玉がガラス玉である

以下の尤度を定義する.

記号	意味
$P(H_A\|S)$	真珠が取り出されたときそれが宝箱 A からの確率
$P(H_B\|S)$	真珠が取り出されたときそれが宝箱 B からの確率
$P(H_A\|G)$	ガラス玉が取り出されたときそれが宝箱 A からの確率
$P(H_B\|G)$	ガラス玉が取り出されたときそれが宝箱 B からの確率

尤度を算出する.「会社 A では真珠とガラス玉の割合が 3 : 1,B 社では 1 : 3 の割合で宝箱が売られている.」という文から

記号	確率
$P(H_A\|S)$	$\frac{3}{4}$
$P(H_B\|S)$	$\frac{1}{4}$
$P(H_A\|G)$	$\frac{1}{4}$
$P(H_B\|G)$	$\frac{3}{4}$

・一回目の玉取り出し.

最初に取り出された球は真珠だったのでその事後確率 $P(H_A|S)$, $P(H_B|S)$ をそれぞれ求める.
ベイズの定理より事後確率を求める.

$$P(H_A|S) = \frac{P(S|H_A)P(H_A)}{P(S|H_A)P(H_A) + P(S|H_B)P(H_B)}$$

$$P(H_B|S) = \frac{P(S|H_B)P(H_B)}{P(S|H_A)P(H_A) + P(S|H_B)P(H_B)}$$

このとき, 事前確率 $P(H_A)$, $P(H_B)$ は理由不十分の原則より等しいとする.
先ほど求めた尤度などを代入すると

$$P(H_A|S) = \frac{\frac{3}{4} \times \frac{1}{2}}{\frac{3}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2}} = \frac{3}{4}$$

$$P(H_B|S) = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{3}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2}} = \frac{1}{4}$$

事後確率がわかる.

・二回目の玉取り出し

一回目同様真珠だったので, ベイズの定理を用いる

$$P(H_A|S) = \frac{P(S|H_A)P(H_A)}{P(S|H_A)P(H_A) + P(S|H_B)P(H_B)}$$

$$P(H_B|S) = \frac{P(S|H_B)P(H_B)}{P(S|H_A)P(H_A) + P(S|H_B)P(H_B)}$$

ただし今回, 事前確率は一回目の事後確率を用いる.
事前確率に前回の事後確率を用いると

$$P(H_A|S) = \frac{\frac{3}{4} \times \frac{3}{4}}{\frac{3}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{4}} = \frac{9}{10}$$

$$P(H_B|S) = \frac{\frac{1}{4} \times \frac{1}{4}}{\frac{3}{4} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{4}} = \frac{1}{10}$$

新たな事後確率がわかる.

・三回目の玉取り出し.

三回目はガラス玉を取り出したので同じくベイズの定理を用いる.

$$P(H_A|G) = \frac{P(G|H_A)P(H_A)}{P(G|H_A)P(H_A) + P(G|H_B)P(H_B)}$$

$$P(H_B|G) = \frac{P(G|H_B)P(H_B)}{P(G|H_A)P(H_A) + P(G|H_B)P(H_B)}$$

$$P(H_A|G) = \frac{\frac{1}{4} \times \frac{9}{10}}{\frac{1}{4} \times \frac{9}{10} + \frac{3}{4} \times \frac{1}{10}} = \frac{3}{4}$$

$$P(H_B|G) = \frac{\frac{3}{4} \times \frac{1}{10}}{\frac{1}{4} \times \frac{9}{10} + \frac{3}{4} \times \frac{1}{10}} = \frac{1}{4}$$

よってA社製である確率が$\frac{3}{4}$であることが分かった.

2.7 例題３

ある病気を発見する検査Tに関して，次のことが知られている.

- 病気にかかっている人に検査Tを適用すると，98%の確率で病気であると正しく判定される.
- 病気にかかってない人に検査Tを適用すると，5%の確率で誤って病気にかかっていると判定される.
- 人全体では，病気にかかっている人とかかっていない人の割合はそれぞれ3%, 97%である.

母集団より無造作に抽出された1人に検査Tを適用し病気にかかっていると判定されたとき，この人が実際に病気にかかっている確率を求めよ.

回答

以下の記号を定義する.

記号	意味
H_1	実際に病気にかかっている.
H_2	実際に病気にかかっていない.
D	検査で陽性と判定される

ベイズの定理より

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

問題文より尤度はそれぞれ

$$P(D|H_1) = 0.98$$

$$P(D|H_2) = 0.05$$

事前確率は,

$$P(H_1) = 0.03$$

$$P(H_2) = 0.97$$

代入すると

$$
\begin{aligned}
P(H_1|D) &= \frac{0.98 \times 0.03}{0.98 \times 0.03 + 0.05 \times 0.97} \\
&= \frac{0.0294}{0.0779} \\
&\fallingdotseq 33.7\%
\end{aligned}
$$

よって検査Tで陽性と判断されても本当に病気の確率は33.7%である．

3　ナイーブベイズフィルター

3.1　ベイズ分類

ベイズ分類とはベイズ理論を利用し与えられたデータを分類する方法である．複数のデータをひとつずつ処理できるベイズ理論の本領が発揮する．

3.2　ナイーブベイズフィルター

ベイズ分類で有名な応用の一つがナイーブベイズフィルターである．ベイズ理論を利用して迷惑メールと通常のメールを仕分けすることができる．

対象のメールの文中の単語をすべて独立のデータとして扱いメールをふるい分ける.「メールの文中の単語をすべて独立」というのは苦しい仮説であるが，この仮説を利用したナイーブベイズフィルターは実用上大いに有効であることが知られている．

3.3　ナイーブベイズフィルター～判定方法～

ナイーブベイズフィルターを用いて迷惑メールを判定する際，迷惑メールとそうでないメールによく含まれる単語をに着目し，それらの単語が迷惑メールと通常メールに含まれる確率を調べる．

最初は複数のデータをまとめて考えたとき，以下の記号を定義する．

記号	意味
H_1	受信したメールが迷惑メールである．
H_2	受信したメールが迷惑メールでない．
D	ある複数の単語がメールから検出された．

このとき，メールが迷惑メールである確率 $P(H_1|D)$ とそうでない確率 $P(H_2|D)$ はベイズの定理より

$$
P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}
$$

$$
P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}
$$

とわかる．判定はこの二つの確率の大小関係で以下の様に考える．

判定	判定条件
迷惑メール	$P(H_1\|D) > P(H_2\|D)$
通常メール	$P(H_1\|D) < P(H_2\|D)$

このとき $P(H_1|D)$ と $P(H_2|D)$ の分母は同じなので分子だけで大小関係が求まる．よって

判定	判定条件
迷惑メール	$P(D\|H_1)P(H_1) > P(D\|H_2)P(H_2)$
通常メール	$P(D\|H_1)P(H_1) < P(D\|H_2)P(H_2)$

次に複数個のデータを別々に考える．具体的に，今 D をある複数の単語がメールから検出されたとしているが，この D を D_1，D_2，D_3 という事象に分けて表す．これらの記号の意味は以下のとおりである．

記号	意味
D_1	ある一つの単語 A がメールから検出された．
D_2	ある一つの単語 B がメールから検出された．
D_3	ある一つの単語 C がメールから検出された．

このとき，$P(D|H_1)$, $P(D|H_2)$ は以下の様に書きかえられる．

$$P(D|H_1) = P(D_3|H_1)P(D_2|H_1)P(D_1|H_1)P(H_1)$$

$$P(D|H_2) = P(D_3|H_2)P(D_2|H_2)P(D_1|H_2)P(H_2)$$

よって判定は

判定	判定条件
迷惑メール	$P(D_3\|H_1)P(D_2\|H_1)P(hD_1\|H_1)P(H_1) > P(D_3\|H_2)P(D_2\|H_2)P(D_1\|H_2)P(H_2)$
通常メール	$P(D_3\|H_1)P(D_2\|H_1)P(D_1\|H_1)P(H_1) < P(D_3\|H_2)P(D_2\|H_2)P(D_1\|H_2)P(H_2)$

となる．このことより各尤度つまりは各単語の出現確率と事前確率を掛け合わせた値を見比べるだけで判定ができる．

3.4 例題４

いまここに迷惑メールか通常メールかわからないメールが一通ある．このメールから次の順でプレゼント，無料，経済という単語が検出された．このときこのメールが迷惑メールか通常メールか調べよ．ただし，受信メールの中で迷惑メールと通常メールの比率は 7 : 3 の割合でこれらの単語は以下の確率で迷惑メールと通常メールに含まれていることが調べられている．

検出語	迷惑メールの確率	通常メールの確率
プレゼント	0.6	0.1
無料	0.5	0.3
経済	0.05	0.5

回答

以下の記号を定義する．

記号	意味
D_1	単語「プレゼント」がメールから検出された．
D_2	単語「無料」がメールから検出された．
D_3	単語「経済」がメールから検出された．
H_1	メールが迷惑メールである．
H_2	メールが通常メールである．

よって

記号	意味	確率
$P(D_1\|H_1)$	単語「プレゼント」がメールから検出されたときそのメールが迷惑メールの確率．	0.6
$P(D_2\|H_1)$	単語「無料」がメールから検出されたときそのメールが迷惑メールの確率	0.5
$P(D_3\|H_1)$	単語「経済」がメールから検出されたときそのメールが迷惑メールの確率	0.05
$P(D_1\|H_2)$	単語「プレゼント」がメールから検出されたときそのメールが通常メールの確率	0.1
$P(D_2\|H_2)$	単語「無料」がメールから検出されたときそのメールが通常メールの確率	0.3
$P(D_3\|H_2)$	単語「経済」がメールから検出されたときそのメールが通常メールの確率	0.5
$P(H_1)$	迷惑メールの確率	0.7
$P(H_1)$	通常メールの確率	0.3

となるので

$$P(D_3|H_1)P(D_2|H_1)P(hD_1|H_1)P(H_1), P(D_3|H_2)P(D_2|H_2)P(D_1|H_2)P(H_2)$$

$$0.05 \times 0.5 \times 0.6 \times 0.7, 0.1 \times 0.3 \times 0.5 \times 0.3$$

$$0.000105 < 0.00045$$

この比較により，通常メールの確率のほうが高い．

4　まとめ

ベイズの統計学について学び、従来の統計学との違いを調べた。理由不十分の原則等, 従来の統計学との大きな差である．ベイズの統計学が注目を浴びてこなかった欠点の 1 つとして計算量が多くなってしまう点が挙げられる．そのためベイズの統計学を最大限利用するためにはコンピューターを利用した計算方法を学ぶことが必要であると考える．今後はベイズの統計学とそのプログラミングについても学んでいきたい．

参考文献

[1] 涌井良幸, 涌井貞美, 史上最強図解これならわかる！統計学, ナツメ社, 2010.

[2] 涌井良幸, 涌井貞美, 史上最強図解これならわかる！ベイズ統計学, ナツメ社, 2012.