

ベイズ統計学概論

久保田静希

芝浦工業大学 数理科学研究会

2019/11/03

ベイズの統計学とは従来の統計学とは違い, 与えられたデータに対してまとめではなく一つずつ処理をしていくことのできる統計学である.

任意の事象 A, B に対して

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

但し, $P(A)$ を事象 A が起こる確率, $P(A|B)$ を事象 B が起きた時事象 A が起こる条件付確率とする.

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

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

ベイズの定理の A を原因もしくは仮定 (Hypothesis) , B をそれらからえられた結果 (Data) と置き換える

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

このように置き換えることによって左辺の式の意味はある結果 D が原因 H によって起こされた確率ととらえることができる。

与えられたデータに対して考えられる原因は一つとは限らない. よって考えられる原因を H_i とするとベイズの定理は

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)}$$

と書き換えられる.

また原因 H がすべて独立ならば $P(D) = P(D \cap H_1) + \dots + P(D \cap H_i)$ と置き換えることができる.

更に条件付確率の定義より $P(D \cap H) = P(D|H)P(H)$ となるので代入すると

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

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。

例題 1

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.

例題 1

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

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

ベイズの定理の一部の確率には名前がある

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

$P(H_i)$ を事前確率, $P(D|H_i)$ を尤度, $P(H_i|D)$ を事後確率という.

ベイズ更新とは従来の統計とは違い与えられた複数のデータを一つずつ処理していくベイズ統計学の醍醐味である. 具体的には一度出した事後確率を事前確率として用いる.

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、ここが従来の統計学との大きな差である。

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. 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.

例題 2

以下の記号を定義する.

記号	意味
H_A	取り出した一つの玉が A 社製の宝箱からである
H_B	取り出した一つの玉が B 社製の宝箱からである
S	取り出した一つの玉が真珠である
G	取り出した一つの玉がガラス玉である

例題 2

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

例題 2

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

例題 2

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.

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

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

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.

例題 2

・ 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.

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

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

例題 2

・ 三回目の玉取り出し.

今回の事前確率には前回出した事後確率を用いる.

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

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

例題 2

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.

先ほどの例題では真珠, 真珠, ガラス玉の順に玉が出ましたがこの順番が違っていても同じ値が出てくる.

基本的にベイズの定理では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.

任意の二つのデータ D_1, D_2 に対し事後確率を

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

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

とする.

D_1, D_2 の順にデータが得られたとき

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

D_2, D_1 の順にデータが得られたとき

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

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。

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。

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。
- 病気にかかっていない人に検査 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。

例題 3

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

例題 3

問題文より尤度はそれぞれ

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

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

例題 3

事前確率は,

$$P(H_1) = 0.03$$

$$P(H_2) = 0.97$$

代入すると

$$\begin{aligned} P(H_1|D) &= \frac{0.98 \times 0.03}{0.98 \times 0.03 + 0.05 \times 0.97} \\ &= \frac{0.0294}{0.0779} \\ &\doteq 0.337 \end{aligned}$$

例題 3

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.

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。

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。

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。

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、次の記号を定義する。

ナイーブベイズフィルター

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次の様に考える。

ナニーブベイズフィルター

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．

ナイーブベイズフィルター

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	$P(D_3 H_1)P(D_2 H_1)P(hD_1 H_1)P(H_1) > P(D_3 H_2)P(D_2 H_2)P(D_1 H_2)$
通常メール	$P(D_3 H_1)P(D_2 H_1)P(D_1 H_1)P(H_1) < P(D_3 H_2)P(D_2 H_2)P(D_1 H_2)$

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。

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

例題 4

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.

例題 4

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

例題 4




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

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

$$0.000105 < 0.00045$$

この比較により，通常メールの確率のほうが高い。

- 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.
- 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た.
- 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.

-  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.
-  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.
-  一石賢, まずはこの一冊から意味がわかるベイズ統計学, ベレ出版,2016.