確率への招待　10回目

確率④

ベイズの定理とその応用

1. ベイズの定理(Bayes' Theorem)

事象A, Bに対し、

$P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ で、$P(B|A) = \dfrac{P(A \cap B)}{P(A)}$ だから、

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

これがベイズの定理(単純!)

例えば、A＝{邪馬台国が畿内にあった}
B＝{纏向遺跡で大規模な建物の遺跡が見つかった}

「纏向遺跡で大規模な建物の遺跡が見つかった」という条件の下で、「邪馬台国が畿内にあった確率」が計算できる。

(ただし、そのためには、「そもそも邪馬台国が畿内にあった確率」(事前確率)がないといけない)

※Thomas Bayes:18世紀のイギリスの牧師。ベイズの定理の特別な場合を証明した以外の業績はよく分からない。

ベイズの定理は高校教科書でも「原因の確率」として記述がある。
そのために、ベイズの定理を少し書き換える。

全事象Uが、A_1、A_2、・・・、A_nに分割されているとする。
すなわちA_1、A_2、…、A_nは互いに排反で$A_1 \cup A_2 \cup \cdots \cup A_n = U$。
ベイズの定理の分母P（B）は、

$$
\begin{aligned}
P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B) \\
&= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)
\end{aligned}
$$

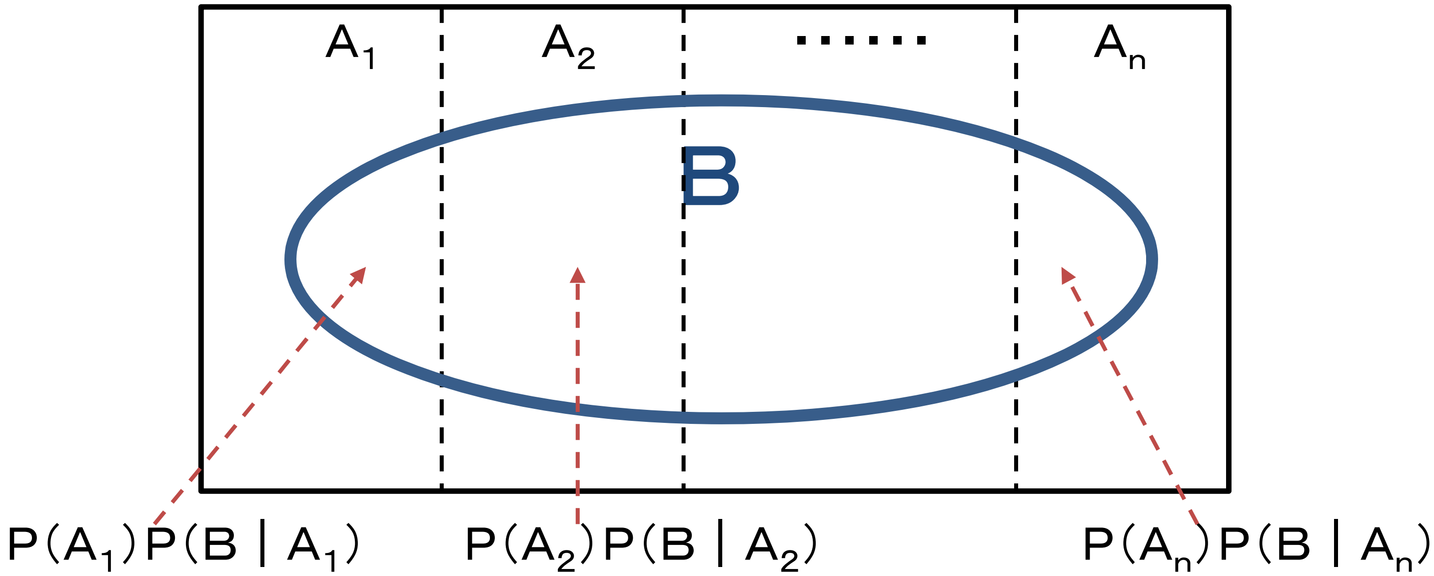
と表せる。よって、

$$
P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}
$$

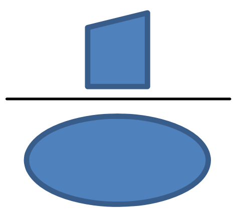
となる。

ビジュアル的に理解すると、

元の全体集合U



$$P(A_i|B) =$$



$$= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}$$

例題１）あるお菓子を製造する工場A，Ｂがあり、Ａ工場のものには３％、Ｂ工場のものには４％の当たりが含まれている。このお菓子のうち、Ａ工場とB工場のものの割合は４：５であるとする。このお菓子を購入し、当たりが出たとする。このとき、このお菓子がA工場のものである確率を求めよ。

解）Ａ＝{購入したお菓子がA工場のもの}
Ｂ＝{購入したお菓子がB工場のもの}
Ｅ＝{購入したお菓子が当たり}とする。

$$P(A)=\frac{4}{9},P(B)=\frac{5}{9},P(E\mid A)=\frac{3}{100},P(E\mid B)=\frac{4}{100}$$

$$P(A\mid E)=\frac{P(E\mid A)P(A)}{P(E\mid A)P(A)+P(E\mid B)P(B)}=\frac{\frac{3}{100}\times\frac{4}{9}}{\frac{3}{100}\times\frac{4}{9}+\frac{4}{100}\times\frac{5}{9}}=\frac{3}{8}$$

ベイズの定理は、数学的には何でもない定理（定義に毛の生えた程度）なのだが、解釈のしようによっては応用範囲が極めて広い。

最初は何だかよくわからない状態（A_1、A_2、・・・、A_nのどれが起きているのかよく分からない）だったのが、いろいろ観測を行って情報を追加していくことにより、A_1、A_2、・・・、A_nのどれだったのか、絞り込んでいくイメージ。

このような、ベイズの定理を基礎にした統計学を**ベイズ統計**という（現代の主流）。

ただし、世の中にはベイズ統計を嫌っている人もいる。

・「邪馬台国が畿内にあったか否か」は、歴史的にはyes/noが決まっていることなので、そんなものは「確率」とは認めない！

・ベイズ統計では「事前分布」が必要。（ベイズ派の人は「とりあえず何でもいいから事前分布を仮定して、」というのだが）それは恣意的、主観的すぎる！

モンティ・ホール問題

（Monty Hallはアメリカのテレビ番組の司会者）

　プレーヤーの前に閉まった3つのドアがあって、1つのドアの後ろには景品の新車が、2つのドアの後ろには、はずれを意味するヤギがいる。プレーヤーは新車のドアを当てると新車がもらえる。
　プレーヤーが1つのドアを選択した後、司会のモンティが残りのドアのうちヤギがいるドアを開けてヤギを見せる。
　ここでプレーヤーは、最初に選んだドアを、残っている開けられていないドアに変更してもよいと言われる。プレーヤーはドアを変更すべきだろうか？

・最初は、どのドアも当たる確率は公平で1/3のはず。
・司会者がハズレのドアを開けると、残りは2つ→確率1/2？
・司会者はいずれにせよドアを開けるのだから、それで確率が変わるのはおかしい→選択を変更する意味なし？
1990年代にアメリカで大論争を巻き起こした。

【ベイズの定理を使った証明】

３つのドアをA，B，Cとし、ドアAに商品がある事象をA，ドアBに商品がある事象をB，ドアCに商品がある事象をCとする。

事前確率は、P（A）＝P（B）＝P（C）＝１／３。

プレーヤーが選んだドアがAであったとしても一般性を失わないので、プレーヤーがドアAを選んだとする。

このとき、司会者モンティがドアBを開けた事象をMとすると、

P（M｜A）＝１／２（司会者はB，Cどちらを開けてもいいから）

P（M｜B）＝０　　（Bが当たりなので司会者がBを開けない）

P（M｜C）＝１　　（Cは当たりなので、Bしか開けられない）

$$P(A|M)=\frac{P(A)P(M\mid A)}{P(A)P(M\mid A)+P(B)P(M\mid B)+P(C)P(M\mid C)}$$

$$=\frac{\frac{1}{3}\times\frac{1}{2}}{\frac{1}{3}\times\frac{1}{2}+\frac{1}{3}\times 0+\frac{1}{3}\times 1}=\frac{1}{1+0+2}=\frac{1}{3}$$

$$P(B \mid M) = \frac{0}{1+0+2} = 0$$

$$P(C \mid M) = \frac{2}{1+0+2} = \frac{2}{3}$$

よって、ドアCに選択を変えた方が当選確率は高くなる。（証終）

直観的には、次のように考えると分かるかも。。。。

・最初は、どのドアが正解か分からないから、ドアAを選んだ時に当選の確率は1/3

・次に、司会者はどれかのドアを開けなくてはならない（しかも外れドア）と分かっているので、司会者がドアBを開けたとしても、ドアAが当たりである確率が変わるわけがない。
⇒引き続き1/3

・一方、司会者がドアBを開けたということは、Bが当たりの可能性は消えた→残りの２／３の確率がCに凝縮された！

【別証明】
最初に選んだドアが当たりである事象をA、
司会者モンティがはずれのドアを開けた後に残っているドアに
変更して当たりとなる事象をBとする。

最初に選んだドアが当たりとなる確率は$P(A) = 1/3$であり、
はずれである確率は$P(\bar{A}) = 2/3$である。

次に、モンティがはずれのドアを開けた後について考える。
もし最初に当たりのドアを選んでいれば、その後ドアを変更するとはずれとなるので$P(B|A) = 0$である。一方、はずれのドアを選んでいれば、残ったドアはあたりとなるので、$P(B|\bar{A}) = 1$である。

以上より、

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$$

となる。

もう一つ、有名なパズルの問題

例題２）３枚のカード①、②、③があって、
　　　カード①は両面が赤、
　　　カード②は両面が白、
　　　カード③は赤と白が１面ずつ、となっている。
　この３枚のカードから１枚を引いたところ、オモテが赤であった。
　このカードの裏が赤である確率を求めよ。

３枚のカードから１枚引いたところオモテが赤だった
→引いたカードは①か③のどちらかだ。
→①なら裏は赤、③なら裏は白なので、確率1/2　？？？

解）根元事象を書き出すと、

	表向き	裏向き
カード①	（赤、赤）	（赤、赤）
カード②	（白、白）	（白、白）
カード③	（赤、白）	（白、赤）

オモテが赤の場合の数は3で、うち裏も赤なのは2とおり
→求める確率は2/3。

「カードを選んだ」というのがマチガイ。
「カードと、その表面か裏面か、を選んだ」ということ。

検診における偽陽性（NHK「ためしてガッテン」でやっていた）

例題3）女性の乳がんの罹患率は約0.3%である。

乳がん検査であるマンモグラフィーは、がんの人をほぼ100%「がん」と判断するが、本当はがんでない人に対し９%を間違って「がん」と判断する（偽陽性）。

ある人がマンモグラフィーで「がん」と判定されたとき、その人が本当にがんである確率はいくらか。

答）A＝{がんである}、M＝{検査でがんと判断された}とする。

$$P(A)=\frac{3}{1000},\ P(\overline{A})=\frac{997}{1000},\ P(M\mid A)=1,\ P(M\mid\overline{A})=\frac{9}{100},$$

$$P(A\mid M)=\frac{P(A)P(M\mid A)}{P(A)P(M\mid A)+P(\overline{A})P(M\mid\overline{A})}=\frac{\frac{3}{1000}\times 1}{\frac{3}{1000}\times 1+\frac{997}{1000}\times\frac{9}{100}}$$

$$=\frac{300}{9273}=3.2\%$$

スパムメール対策におけるベイズフィルター

スパムメールの検出にも、ベイズの定理が応用される。

例題4）スパムメールに特徴的な用語、例えば「無料」について、

・スパムメールに「無料」という単語が含まれる確率　60%

・スパムでないメールに「無料」という単語が含まれる確率　10%

とする。あるメールに「無料」という単語が含まれている場合、そのメールがスパムである確率を求めよ。

解）ベイズの定理を使うためには、事前確率が必要。今の場合、何の事前情報もないので、とりあえず1/2としておく。

Ｓ＝{スパムメール}、Ｍ＝{「無料」の語が使われている}として、

$$P(S \mid M) = \frac{\frac{1}{2} \times 0.6}{\frac{1}{2} \times 0.6 + \frac{1}{2} \times 0.1} = \frac{6}{7}$$

○モデルの推定

経済分析やもっと一般にデータ分析では、モデルを構築して、そのモデルのパラメーターを推定。

モデル（ブラックボックス）

入力　　　　出力

（$x_1,x_2,\cdots,x_n$）　$y = a_1x_1+a_2x_2+\cdots+a_nx_n$＋誤差項　　y

例えば、　入力　　　　出力

税率、補助金、・・・　　GDPの増加

顔の特徴、歩き方、・・・　　不審人物か否か

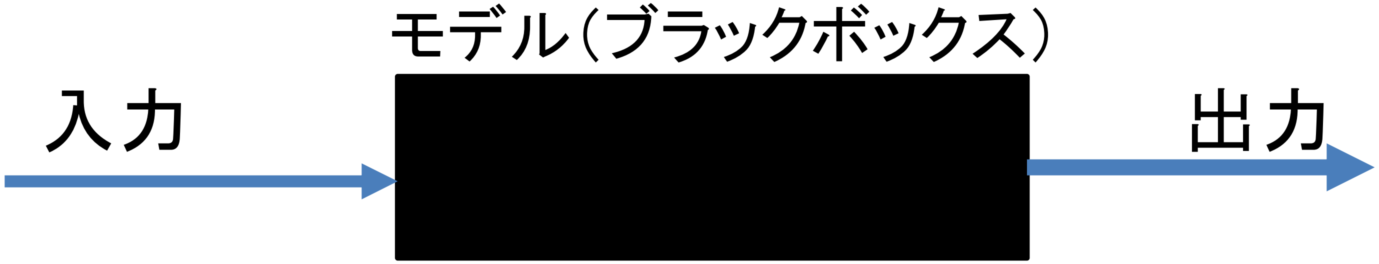
地盤の変化　　地震の発生　　など

（$x_1,x_2,\cdots x_n$）とyを観測して、モデルを推定する。

パラメータ A_i＝（$a_1,a_2\cdots$）に何らかの値を代入すると、ベイズの定理の右辺P(B｜A_i)が決まる。ここで A_i をいろいろ変えてシミュレーションをする（モンテカルロシミュレーション）をすると、ベイズの定理の左辺P（Ai｜B)が求められる。

○モデルの推定

経済分析やもっと一般にデータ分析では、モデルを構築して、そのモデルのパラメーターを推定。



$(x_1, x_2, \cdots, x_n)$　$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ ＋誤差項　y

例えば、　入力　　　　出力

税率、補助金、・・・　　GDPの増加

顔の特徴、歩き方、・・・　　不審人物か否か

地盤の変化　　地震の発生　　など

$(x_1, x_2, \cdots x_n)$とyを観測して、モデルを推定する。

パラメータAi＝$(a_1, a_2 \cdots)$に何らかの値を代入すると、ベイズの定理の右辺P(B｜Ai)が決まる。ここでAiをいろいろ変えてシミュレーションをする（モンテカルロシミュレーション）をすると、ベイズの定理の左辺P（Ai｜B)が求められる。