確率について

今日では、いたるところに確率という単語は登場します。
例えば、じゃんけんで勝つ確率は三分の一である。とか、明日の降水確率は２０パーセントである。とか宝くじの当たる確率は・・・といったふうです。

このようにいろいろな場所で登場する確率を詳しく見てみると
1・さいころの出る目が常に 1/6 である。と言ったように前後の影響を全く受けない確率
2・天気予報などのように、特定の条件下である出来事が起こるといった、前後の影響を如実に受ける確率があります。

今回取り上げる、ベイズの定理は後者にあたります。

ベイズの定理と条件付き確率

ベイズの定理と条件付き確率は切っても切れない関係があります。なぜならベイズの定理は条件付き確率を基礎として成り立っています。

１・ベイズの定理

・基本様式

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

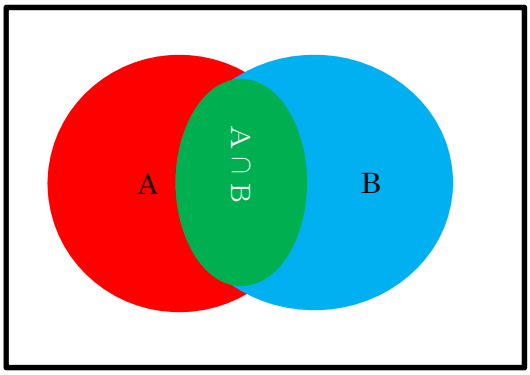
となっていますがこのままではよく意味の分からない式になっています。ここでこの解りにくい記号の意味を説明するために条件付き確率についてふれたいと思います。

２・条件付き確率とその具体的な例
まず確率の記号の定義
P(A)・・事象 A の起こる確率です
次に条件付き確率で定義
$P_A(B)$=P(B|A)・・事象 A が起こった条件の下で事象 B が起こる確率という意味となります。
ですので、
P(A|B)の意味は・・事象 B が起こった条件の下で事象 A が起こる確率という意味になります。

（具体例）
２０回に１回の割合で帽子を忘れる癖のあるＡ君が正月にＡ・Ｂ・Ｃ三軒を年初めに挨拶にいったとします。
この場合、二件目のＢに忘れてきた確率は？？

（A）単純に１９/２０×1/２０　ではなく、この場合は忘れてきたことが起こった条件の下で話が進んでいるので、確率の分母は帽子を忘れた確率、分子はＢの家で帽子を忘れた確率になります。
ですのでP(A)＝帽子を忘れた確率　P(B)＝Ｂの家で忘れた確率とすると
P(A)・・・（帽子を忘れた確率）
＝1/２０＋（１９/２０×1/２０）＋（１９/２０×１９/２０×1/２０）＝１１４１/８０００・・・①
P(B)・・・（Ｂ家で帽子を忘れた確率）
＝１９/２０×1/２０＝１９/４００・・・②
よって、P(B|A)＝①/②×１００＝３３．３％となる。

３・ベイズの定理の証明
２のところで確認した条件付確率の定義を用いて１のベイズの定理を証明を考えていきたいと思います。
$P(A\cap B)=P(A)\times P(B|A)=P(B)\times P(A|B)$
<証明>



全体の事象数はM
Ａの事象数はＰ
Ｂの事象数はQ
Ａ∩Ｂの事象数はR

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

$P(B)\times P(A|B)=\frac{Q}{M}\times\frac{R}{Q}=\frac{R}{M}$

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

よって、$P(A\cap B)=P(A)\times P(B|A)=P(B)\times P(A|B)$となる。

$P(B|A)=\frac{P(B)\times P(A|B)}{P(A)}$　となる。これは、$P(B|A)\propto P(B)\times P(A|B)$をも記している。

このようにして、ベイズの定理は正しい事が証明されました。

つまりこの式を文字として書き表すと
A が起こった条件下で B が起こる確率は、B が起こった条件下で A が起こる確率と B が起こる確率の積になっています。
つまり、今までに B が起こる確率を知っており、事象 B が起こった時に A が起こる確率を知ったならば、新たな A が起こった状況で B が起こるといった確率を得ることができます。よく見ると一見、あべこべで繋がらないように見えますが、これを繋げることをできるのが強みであり、上記の証明が正当性を保証してくれます。

言い換えると、P(B｜A) 事後確率＝P(A|B) 尤度×P(B) 事前確率
となります。

さて、先ほどのベイズの定理の分母には P(A)がありました。
P(A)は規格化の意味を持ちます。

（ベイズの定理の実用例）
今、空港にいるとして自分の乗るべき飛行機が運航を中止しています。
この時、同時に二つ以上のシステムが故障することはないと考える
ある便が故障のために運転中止になった時・・
原因が動力であった確率は？

i	システムの故障個所 Hi	システムの故障確率 P(Hi)	システムが故障した時の運航中止 P(A|Hi)
1	機体	0.307	0.008
2	ローター	0.156	0.048
3	電気	0.129	0.040
4	計器	0.130	0.052
5	動力	0.080	0.100
6	通信・制御	0.030	0.151
7	その他	0.171	0.014

P(A)は飛行機が運航中止)

$$P(H5|A)=\frac{P(H5)P(A|H5)}{\sum_{i=1}^{7}P(Hi)P(A|Hi)}=\frac{0.080\times0,100}{0.0366}=0.219=(21.9\%)$$

となる。

４・ベイズの定理の生まれ

１７４０年代のイングランドでアマチュア数学者、トーマス・ベイズによって発見される。アマチュアといってもこの時期のイングランドでは宗教対立により非国教徒は大学から締め出されていたので業績ある数学者の多くがアマチュアであった。
ベイズ自身はあまりこの発見に関心をしるさなかったが、ベイズの友人であるリチャード・プライスがベイズの法則を用いた論文を公表したことで今日にベイズの名を残した。また今日のベイズの法則の基になっているのは１７７４年に大量の天文のデータを処理する方法を考えた、数学者のラプラスの見つけた方法の応用となっている
その後は第二次世界大戦においてエニグマの解読に役立ち、現在ではスパムメールのブロックや沈んでいる潜水艦の発見などに役立っている。

５・現在社会におけるベイズの定理
　迷惑メールをいかにしてブロックするか
スパムメールを判定するには・・
単純に・・・
① ヘッダを含むメールの中身から、スパムか否かを判定する方法。
② スパム発信者の通信パターンから、スパムを遮断する方法がある。

② の方法では判別の方法は通信パターンのみで非常に簡易であるが、通信パターンが必ず必要になるので予測はおろか、かなりの回数でスパムメールが届くアドレスしか検出することはできない。つまり未然に防ぐことができるのは少ないと考えられる。

一方、①でならどうするかというと、送られてきたメールを全て名詞単位で分割することで、メールを名詞のたくさん入った袋の様に認識することができます（これを bag-of-words といいます。）そうして作った袋の中を覗いてみて、怪しい単語の方が多そうならばスパムメール、善良な単語が多ければ安全なメールとすることができます

ここにベイズの定理を用いてみようと考えます。

$$P(\text{スパム} \mid \text{メールの特徴 E}) = \frac{(P(\text{メールの特徴E} \mid \text{スパム}) \times P(\text{スパム}))}{(P(\text{メールの特徴E})}$$

メールの特徴 E は bag-of-words を用いる今、判断したいメール
P（スパム）は、あるメールがスパムである確率（事前確率）・・これは今まで手元に届いたスパムメールの数（事前情報がなければ２分の１）
P（スパム｜メールの特徴 E）はある特徴 E を見たときにスパムかどうか（事後確率）
P（メールの特徴 E｜スパム）はスパムにメールの特徴 E が現れる確率（尤度）
P(メールの特徴 E)はスパムと非スパムを合わせた全メールの中にメールの特徴 E を持ったメールが出現する確率

この式からわかる事
特徴 E をもつメールがスパムである確率が上がるのは（式の左辺）
1・メール全体でスパムの割合が増える時
2・特徴 E をもつスパムが増える時となります。

もし、ベイズの定理を使わなければ世界中の全てのメールから同じ特徴 E を持つメールがスパムか非スパムであるかを調べて確率を計算しなければならないので非常に煩雑であり自分の家のパソコンではとても可能ではありません。

この様にして、ベイズは生活の中でも活躍しています。