Xác Suất Thống Kê

(Tác giả: Phạm Nguyễn Mạnh)

1/Giới thiệu về xác suất:

a/Định nghĩa xác suất:

Quan sát các hiện tượng tự nhiên, ta thấy có những hiện tượng thường xảy ra, có những hiện tượng ít xảy ra. Xác suất là một đại lượng thể hiện mức độ xảy ra (thường xuyên hay ít khi) của một biến cố.

Ví dụ:

+Chơi bài cào, có thể sự dụng xác suất để tính được khả năng được 3 con Tây

+Quan sát các ngày trước đó và xem dự báo thời tiết, có thể đoán được gần đúng khả năng hôm nay mưa.

b/Phân biệt xác suất và thống kê:

Vấn đề chính mà xác suất và thống kê giải quyết gần như trái ngược nhau:

- Trong lý thuyết xác suất, chúng ta dựa vào những quy tắc và thông tin đã biết, (cùng với 1 ít ngẫu nhiên) chuyển hóa thành những số liệu để dự đoán chuyện gì đã, đang hoặc sẽ xảy ra.

Ví dụ: Dựa vào kiến thức đã biết về bài cào và bộ bài 52 lá, ta có 7 nút, ta có thể dự đoán khả năng mình thắng

-Trong thống kê, chúng ta dựa vào những gì đã xảy ra, chuyển hóa thành số liệu để dự đoán quy tắc và thông tin nhằm mục đích giải thích sự kiện đó.

Ví dụ: Dựa vào thống kê số người bị bệnh sởi theo độ tuổi, chúng ta có thể dự đoán đối tượng nào dễ mắc bệnh nhất hoặc tác nhân gây bệnh (có liên quan tới đối tượng bị bệnh nhiều)

-Bài giảng này sẽ tập trung vào phần xác suất

c/Luyện tập xác suất căn bản:

Câu 1: Anh Tài là sinh viên đại học Khoa Học Tự Nhiên, chuyên ngành công nghệ thông tin, vẫn còn ế, sống ở Hồ Chí Minh và thích chơi Liên Minh. Trong 2 sự kiện này cái nào có khả năng xảy ra cao hơn: 'Anh Tài 20 tuổi' hay 'Anh Tài 20 tuổi và rất rành về máy vi tính'?

Câu 2: Khả năng để đổ 1 xúc xắc 20 mặt ra 1 số chia hết cho 6 là bao nhiêu?

Câu 3: Trong 3 câu sau câu nào là có cơ sở về xác suất nhất:

-Mình đổ xúc xắc 4 lần số 5 liên tiếp, chắc số tiếp theo không thể nào là 4 đâu!

-Nãy giờ trắc nghiện không có câu D nên những câu tiếp theo chắc sẽ có câu D.

- Thường thì 5h chiều đường Cách Mạng Tháng 8 hay kẹt xe, khả năng cao hôm nay cũng vậy.

Câu 4: Dự đoán thử trong 2 trường hợp sau trường hợp nào có khả năng xảy ra cao hơn và tính thử xấp xỉ khả năng của 2 trường hợp: Khả năng ra 7 mặt ngửa khi tung đồng xu 10 lần và khả năng ra 14 mặt ngửa khi tung đồng xu 20 lần.

Câu 5:Anh Linh và anh Trung tham gia trò chơi sau: 2 người sẽ đoán 1 số nguyên dương bất kì (2 anh không biết kết quả của người còn lại), nếu đoán giống nhau, 2 anh sẽ thắng. Giả sử anh Trung biết số anh Linh thích nhất, và cả 2 đều biết điều này. Việc anh Linh biết số anh Trung thích nhất có làm tăng khả năng thắng của 2 người không?

d/Ứng dụng của xác suất cổ điển trong đời sống:

Ngoài những ứng dụng dễ thấy trong chơi bài, cá cược và 1 vài bộ môn thể thao, xác suất còn được sử dụng trong rất nhiều lĩnh vực khác nhau như Tài chính kinh doanh, machine learning, dự báo thời tiết, tính độ hiệu quả của thuật toán,…

**: Hãy tìm ứng dụng của xác suất trong 3-5 lĩnh vực tùy ý.

2/ Các dạng và kĩ năng xác suất quan trọng:

a/Xác suất rời rạc:

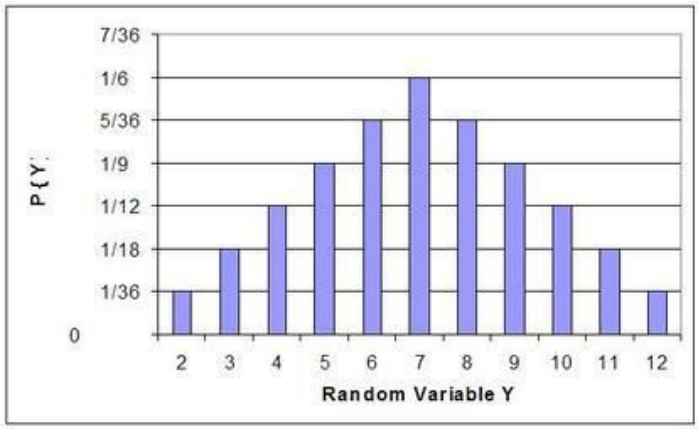
Giả sử ta tung 1 đồng xu 100 lần, ta sẽ tính số lần kết quả là mặt sấp. Số mặt sấp có thể là bất kì số tự nhiên nào từ 0 đến 100 mà không thể là những số thực khác hay số ảo như: 25.5 hay 100i. Ví dụ trên biểu diễn cho sự rời rạc và xác suất rời rạc giải quyết các vấn đề mà các kết quả không liên tiếp nhau trong dãy số.

b/Xác suất liên tục:

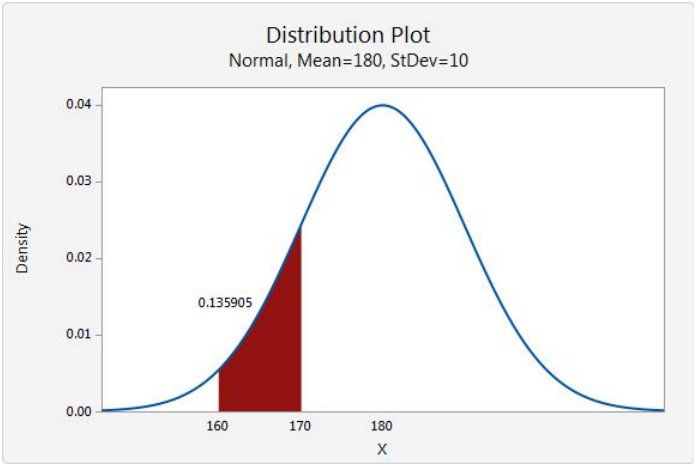
Giả sử 1 tiêu chí làm lính cứu hỏa của trạm cứu hỏa M là cân nặng phải từ 85 đến 120kg thì 1 lính cứu hỏa trong trạm có thể có cân nặng là 1 số thực bất kì liên tục từ 85 đến 120. Ví dụ trên biểu diễn cho sự liên tục và xác suất liên tục giải quyết các vấn đề mà các kết quả có thể chạy từ a đến b (a,b là các giới hạn của kết quả)

(Định nghĩa trên là chưa hoàn chỉnh nhưng hi vọng nó đã giúp mọi người hình dung được rời rạc và liên tục là như thế nào)

c/Mật độ xác suất: (Tham khảo)



Mật độ xác suất rời rạc



Mật độ xác suất liên tục

Hàm mật độ xác suất là hàm biểu diễn xác suất xảy ra của những kết quả khả thi trong thử nghiệm (thường là kết quả khi thực hiện thử nghiệm nhiều lần)

**Tham khảo thêm tại:
https://vi.wikipedia.org/wiki/H%C3%A0m_m%E1%BA%ADt_%C4%91%E1%BB%99_x%C3%A1c_su%E1%BA%A5t

Hoặc trang tiếng Anh:

https://en.wikipedia.org/wiki/Probability_distribution

d/Giá trị kì vọng và 1 số công thức xác suất quan trọng:

-Giá trị kì vọng:

+Khi tung đồng xu, ta quy ước nếu mặt sấp thì ta sẽ được 1 viên kẹo, mặt ngửa thì không được gì cả. Như vậy, giá trị kì vọng của số kẹo ta được sau mỗi lần tung sẽ là: 50%*1 + 50%*0 = 0.5. Vậy giá trị kì vọng của x, số kẹo ta được sau 10 lần tung (kí hiệu là E(x)) sẽ là 0.5*10 = 5 viên kẹo

+Giá trị kì vọng của biến X sẽ được tính theo công thức:

$$E(X) = \sum xP(X = x)$$

Trong đó $P(X = x)$ là khả năng biến X có giá trị x

+Giá trị kì vọng rất hữu dụng trong nhiều trường hợp, ví dụ: 1 vụ đầu tư sẽ được đánh giá (theo tiêu chuẩn thông thường) là lời nếu giá trị kì vọng lớn hơn 0 và lỗ nếu nhỏ hơn 0 (để ý rằng khi thực hiện nhiều lần vụ đầu tư có GTKV lớn hơn 0 thì khả năng rất cao sẽ mang lại lợi nhuận); khi tính độ hiệu quả của 1 thuật toán, ta cũng hay dùng GTKV để tính (dựa trên big-O notation);…

+1 số công thức khác của GTKV (a,b,c là các hằng số thực):

$$E[X+Y] = E[X] + E[Y]$$

$$E[aX] = aE[X]$$

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

**Tham khảo thêm tại: https://en.wikipedia.org/wiki/Expected_value
(không khuyến khích)

+Chỉ số hạnh phúc:

Trong 1 số trường hợp, giả sử như anh A đang có 100000 đồng, anh ấy sẽ không tham gia 1 vụ cá cược 50% được 120000 đồng và 50% mất hết số tiền mình đang có dù giá trị kì vọng cho thấy cá cược này là lời.

Lí do là vì chỉ số hạnh phúc của anh ấy cho thấy vụ cá cược này là không có lợi. Chỉ số hạnh phúc $f(x)$ là chỉ số hài lòng của bản thân dựa trên giá trị x mình đang có (thường là tiền nhưng không nhất thiết)

Quay lại trường hợp của anh A, giả sử ta cho $f(x) = \sqrt{x}$, trong vụ cá cược trên, dựa theo công thức đó sẽ có 50% khả năng anh ấy tăng chỉ số hạnh phúc lên $\sqrt{220000} - \sqrt{100000} \simeq 152,8$ và 50% khả năng giảm xuống $\sqrt{100000} - 0 \simeq 316,2$. Vậy đây sẽ không phải 1 vụ cá cược có lợi cho anh A

Anh A buồn vì ~~ngu~~ thua cược

**Hãy đánh giá chỉ số hạnh phúc của mình và cho ví dụ về 1 vụ cá cược công bằng cho bản thân.

+ 1 điều nên lưu ý về giá trị kì vọng:

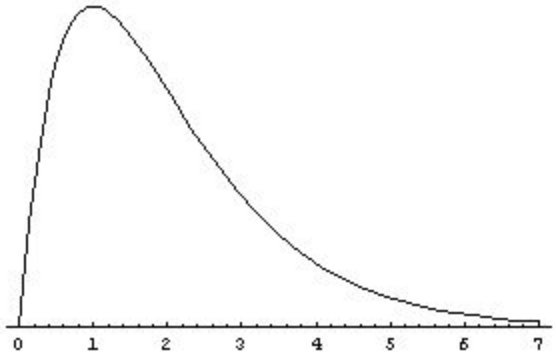
Giả sử tỉ lệ thắng trong 1 vụ cá cược là 1 trên n (n nguyên dương), như vậy theo công thức giá trị kì vọng $E[aX] = aE[X]$ thì thực hiện n vụ cá cược như vậy thì số lần ta kì vọng thắng sẽ là 1.

Bây giờ ta sẽ tính khả năng thực hiện n vụ cá cược mà thắng ít nhất 1 vụ:

Khả năng để thua hết n vụ là (khi n đủ lớn): $(\frac{n-1}{n})^n = \frac{1}{(1+\frac{1}{n-1})^n} \simeq \frac{1}{e} \simeq 36,8\%$

Suy ra tỉ lệ để thắng ít nhất 1 vụ sẽ là khoảng 63,2%, thật sự không cao lắm, nhưng tại sao giá trị kì vọng vẫn là 1? Phải chăng phép toán có vấn đề? Bạn thử suy nghĩ xem.

Câu trả lời rất đơn giản, giá trị kì vọng khác với khả năng xảy ra các giá trị. Giả ta cho biểu đồ tỉ lệ sau (biểu đồ này cũng 'gần giống' biểu đồ tỉ lệ của vấn đề trên):



Giả sử GTKV của sơ đồ trên là 1.5, điều đó không có nghĩa là phần lớn các giá trị có thể có của nó phải lớn hơn 1.5.

**Dựa vào đó, khi thực hiện 1 vụ cá cược, ngoài việc tính GTKV, ta cũng có thể tính khả năng chúng ta sẽ được lãi.

**Hãy tìm công thức tính GTKV của 1 sự kiện có xác suất liên tục (gợi ý: sử dụng tích phân)

-Công thức tính xác suất quan trọng:

$P(A \mid B) = P(A \cap B) / P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Sự kiện A và B không phụ thuộc vào nhau khi và chỉ khi:

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

$F_X(x) = P(X \leq x)$ (phương trình mật độ xác suất)

$$E(n) = \sum_{k=1}^{\infty} P(n \geq k)(n \in Z^+)$$

Trong đó $P(A \cap B)$ là khả năng A xảy ra khi B đã xảy ra, $P(A \cap B)$ là khả năng A và B đều xảy ra, $P(A \cup B)$ là khả năng A hoặc B xảy ra

e/Phương pháp Monte Carlo:

Phương pháp Monte Carlo là phương pháp và chúng ta sử dụng thử nghiệm kiểm tra biến ngẫu nhiên (thường là trên máy tính) để kiểm tra kết quả tổng quát.

Phương pháp này tùy theo yêu cầu đề bài mà sẽ có sự khác nhau nhưng thường sẽ đi theo quy luật sau:

1/Tạo vùng giới hạn cho các biến

2/Tạo nên các biến ngẫu nhiên theo tỉ lệ cho trước trên vùng giới hạn đó

3/Tính số các biến ở những vị trí cho trước

4/Tính kết quả dựa trên công thức đã tạo và thông tin từ bước 3

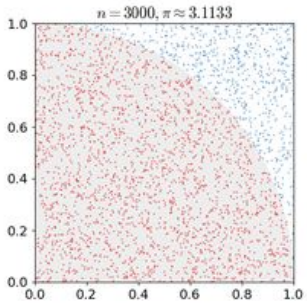
**Ví dụ: Ta muốn tính xấp xỉ giá trị của số pi:

1/Vẽ 1 hình vuông rồi vẽ hình tròn bên trong nó có cùng tâm với hình vuông (tiếp xúc với cạnh hình vuông)

2/Chọn ngẫu nhiên (1000 điểm, 1000000 điểm,…) các điểm trong vùng giới hạn hình vuông, khả năng ở mọi vị trí có điểm ngẫu nhiên là như nhau

3/Tính số điểm nằm trên và trong hình tròn

4/ Tính số pi bằng cách lấy số điểm trên hình tròn chia cho tổng số điểm rồi nhân 4



Ưu điểm của phương pháp này:

-Có thể tính toán bằng máy tính mà không cần phải tốn quá nhiều công sức.

-Với đủ số biến khả năng cao sẽ cho ra kết quả gần đúng chấp nhận được.

Nhược điểm:

-Không thể đảm bảo kết quả trong giới hạn chấp nhận được

**Hãy thử chứng minh khi cho càng nhiều biến ngẫu nhiên thì khả năng cho ra kết quả đúng của bài trên càng cao

f/Phương sai, độ lệch chuẩn:

-Phương sai là 1 công thức thống kê dùng để đo độ lệch về giá trị của các biến ngẫu nhiên. Độ lệch được hiểu là mức độ chênh lệch của các biến so với giá trị kì vọng của chúng.

-Phương sai được tính theo công thức sau:

$\text{var}(X) = E((X - E(X))^2)$ hay $\text{var}(X) = E(X^2) - E^2(X)$ theo công thức GTKV

Do GTKV chẳng khác gì giá trị trung bình trong trường hợp các biến có khả năng xảy ra như nhau. Khi đó ta có:

$$\text{var}(X) = \frac{\sum_{i=1}^{n}(x - \bar{X})^2}{n}$$ ($\bar{X}$ là giá trị trung bình của các biến)

-Độ lệch chuẩn cũng có công dụng tương tự phương sai, thậm chí độ lệch chuẩn có giá trị bằng căn bậc 2 của phương sai:

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{X})^2}{n}}$$

-Sự khác biệt của 2 đại lượng này là độ lệch chuẩn sẽ có bậc 1, bằng với bậc của các biến ngẫu nhiên, điều này rất quan trọng trong nhiều lĩnh vực, điểm hình là tài chính kinh doanh.

**Tại sao chúng ta không sử dụng công thức

$$\sigma(X) = \frac{\sum_{i=1}^{n} | x - \bar{X} |}{n}$$

Để thay cho độ lệch chuẩn?

-Phương sai và độ lệch chuẩn có rất nhiều công dụng, nổi bật nhất là dùng để đánh giá độ ổn định của số liệu. Ví dụ như: chị Tiên có thử nghiệm chất lượng món ăn của mình bằng 2 phương pháp khác nhau. Phương pháp thứ nhất luôn đạt điểm 8 trong khi phương pháp còn lại dao động từ điểm 6 tới điểm 10. Chị khả năng cao sẽ chọn phương pháp 1 nếu muốn bán loại món ăn này vì nó ổn định hơn.

**Tìm những trường hợp mà 2 kết quả có GTKV như nhau thì kết quả có phương sai cao hơn có lợi hơn, phương sai thấp hơn có lợi hơn.

**Tham khảo thêm tại
https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php

<u>g/ 1 số lỗi thường gặp trong xác suất:</u>

-Trong tư tưởng:

+ Không đánh giá được rằng sự kiện a,b,c cùng xảy ra có khả năng thấp hơn sự kiện a,b cùng xảy ra (vd: câu 1 trong phần 1c)

+Không đánh giá được dữ liệu cung cấp có làm tăng khả năng sự kiện xảy ra hay không (vd: 3 lần tung đồng xu ra sấp liên tiếp không làm tăng khả năng lần tung thứ 4 là ngửa)

-Trong tính toán:

+ Không phân biệt được sự kiện có phụ thuộc vào nhau hay không, hoặc tính khả năng ngược lại nhưng không trừ đi lại (Vd: cần tính $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ nhưng ta chỉ tính $P(A \cap B)$, quên tính phần còn lại)

+Nhầm lẫn trong việc tính toán số lượng còn lại sau mỗi trường hợp (vd: khả năng lấy 2 bi đỏ từ rổ 5 bi xanh 5 đỏ là (1/2)*(4/9) chứ k phải 1/4)

+Nhầm lẫn xác suất bình thường và xác suất Bayesian (Sẽ được đọc ở chương sau)

h/Luyện tập xác suất nâng cao:

(Khuyến khích sử dụng Python để xác định đáp số)

1/ Anh Mạnh muốn mua nhẫn tặng chị Y, anh không biết chọn loại nào giữa 20 loại khác nhau nên anh nghĩ ra cách sau: chọn k chiếc nhẫn bất kì (biết rằng mỗi loại có đúng 2 chiếc nhẫn) và lấy 2 chiếc nhẫn cùng loại bất kì (nếu có) trong k chiếc đó. Hỏi anh nên chọn bao nhiêu chiếc nhẫn thì:

a/ Khả năng cao sẽ có ít nhất 1 cặp nhẫn và không phải chọn quá nhiều.

b/ *Khả năng có đúng 1 cặp nhẫn là cao nhất(đáp số k =9; 47,2%)

2/ A và B chơi đổ xúc xắc 20 mặt, khả năng để A cao điểm hơn B là bao nhiêu?

3/ Cổ phiếu của 1 công ti cứ 1 tuần có 50% khả năng tăng 1% so với giá trị đầu tư từ thời điểm đầu tiên và 50% khả năng giảm 1%. Hỏi sau 8 lần khả năng sinh lời là bao nhiêu?

4/Bạn cần 8 điểm cho môn toán trắc nghiệm để đỗ đại học, hỏi bạn cần chắc chắn làm đúng bao nhiêu câu để có khả năng cao đạt được 8 điểm, nếu đề toán có 50 câu? 100 câu?

(Bạn có thể chọn trên 60%, 70%, 80% hoặc 90%)

5/*Khả năng công ty TA có ít nhất 1 thương vụ lỗ mỗi năm (biết rằng mỗi tháng khả năng như nhau) là 95%, hỏi khả năng công ty này có ít nhất 1 thương vụ lỗ mỗi tháng là bao nhiêu?

6/a/Trong 3 lượt, mỗi lượt anh Nhật chọn 1 số bất kì từ 1 đến 100, giá trị kì vọng tích của các số anh đã chọn là bao nhiêu?

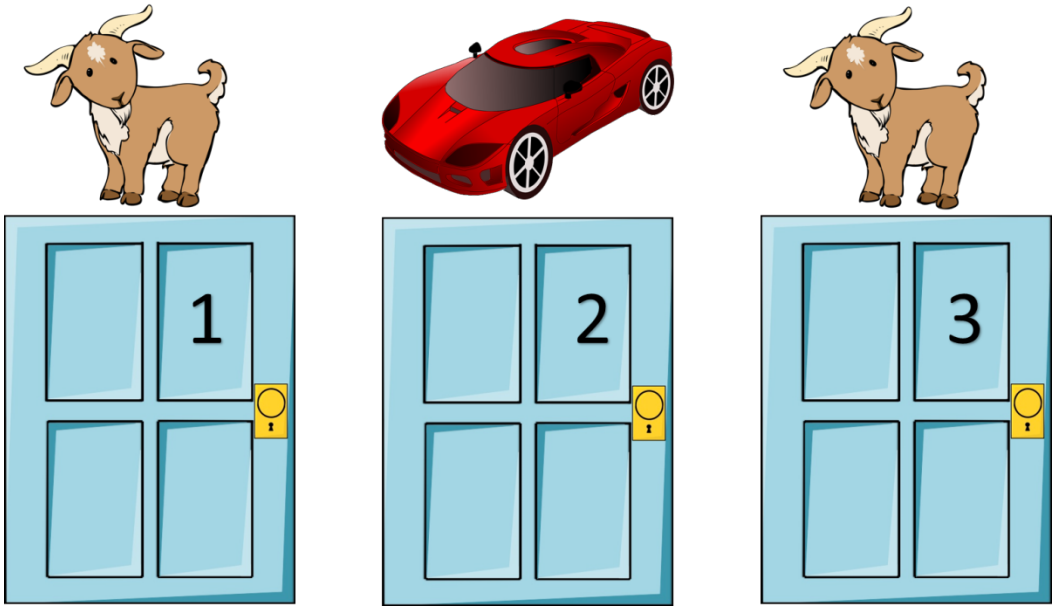
b/Nếu anh ấy chọn 1 số bất kì rồi áp dụng tiếp cho 2 lượt còn lại thì GTKV sẽ là bao nhiêu?

3/Công thức Bayesian:

a/Bài toán Monty Hall: (Từng được sử dụng trong chương trình truyền hình có thật ở Mĩ)

Trò chơi như sau: Có 3 cánh cửa, trong đó 1 cảnh cửa đằng sau là xe hơi và 2 cửa còn lại đằng sau là con dê. Người chơi được chọn 1 cánh cửa và người dẫn chương trình (biết trước cửa nào có xe hơi) sẽ mở 1 trong 2 cánh cửa người chơi không chọn, cánh cửa đó chắc chắn mở ra có con dê ở phía sau. Người chơi được quyết định có thay đổi lựa chọn của mình sang cửa còn lại hay không để tìm được cửa có xe hơi.

**Nếu là người chơi bạn sẽ quyết định như thế nào?



b/Công thức Bayesian:

Trong phần trước chúng ta đã biết công thức $P(A \mid B) = P(A \cap B) / P(B)$. Như vậy, chúng ta cũng có $P(B \mid A) = P(A \cap B) / P(A)$ trong đó A,B là 2 sự kiện bất kì. Kết hợp 2 công thức trên lại ta có:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \text{hay:} \quad P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

(H,E là hypothesis và evidence, giả thuyết và bằng chứng)

Công thức trên chính là xác suất có điều kiện, nó giúp ta tính được khả năng sự kiện H xảy ra khi đã có dữ liệu E.

Quay lại bài toán Monty Hall, 1 ví dụ kinh điển cho công thức này. Giả sử bạn chọn cửa số 1 và người dẫn chương trình mở cho xem cửa số 3, khi đó E là sự kiện MC mở cửa số 3 (anh ta biết rằng cửa có con dê là cửa số 3 và 1 cửa khác), vậy P(E) = 0.5 do khả năng anh ta mở cửa 3 hoặc mở cửa còn lại là như nhau. Nếu người chơi không thay đổi đáp án, gọi H là khả năng xe hơi ở cửa số 1, vậy P(E|H) sẽ bằng 0.5 do MC có thể chọn cửa 2 hoặc cửa 3, theo công thức Bayesian, P(H|E) sẽ bằng P(H) = 1/3 và khả năng người chơi thắng là 1/3.

Nếu người chơi quyết định thay đổi,gọi H khả năng xe hơi ở cửa 2 thì P(E|H)=1 và P(H|E) = 2/3, khả năng người chơi thắng là 2/3.

Bài toán trên đã cho thấy sự lầm tưởng có thể xảy ra khi không cập nhật xác xuất dựa trên dữ liệu (nghĩ rằng đổi hay không đổi đều như nhau)

Chúng ta hãy thử tìm hiểu 1 bài toán khác nếu bạn chưa tin về khả năng bị mắc bẫy bởi công thức Bayesian (nếu tin cũng cứ làm cho vui):

Gia đình X có 2 con, trong đó ít nhất 1 đứa là con trai, khả năng có con trai hay con gái là như nhau (chỉ có 2 giới tính thôi), hỏi khả năng gia đình trên có 2 đứa con trai là bao nhiêu?

Các bạn có thể nghĩ đến là 50%?

Chúng ta hãy dựa vào công thức Bayesian để tính. Ta đặt H là giả thuyết gia đình có 2 con và E là thông tin gia đình có ít nhất 1 đứa con trai. Ta có: P(H) = ¼, P(E|H) = 1, còn P(E)? Ta có 4 khả năng: gia đình có 2 gái, chị gái em trai, anh trai em gái hoặc 2 trai, vậy P(E) = ¾, ta rút ra kết luận P(H|E) = 1/3, khả năng có 2 con trai chỉ là 1/3.

**Lưu ý, khi đặt H,E cần phải tính toán kĩ lưỡng trước khi đặt và E phải là sự kiện đã diễn ra. Bài toán Monty Hall sẽ ra kết quả sai là 50-50 nếu ta đặt H và E sai.

c/ Ứng dụng của công thức Bayesian trong đời sống:

Công thức Bayesian rất quan trọng trong đời sống. Dựa theo nguyên lí xác suất, giả thuyết sẽ thay đổi khi có thêm dữ liệu cập nhật.

Công thức này có ứng dụng trong nhiều lĩnh vực khác nhau: Y học, Kinh tế, Tài chính-Kinh doanh, Thể thao (nhất là cá cược), Dự báo Thời tiết, Máy học (Machine Learning),…

Công thức Bayesian đóng vai trò quan trọng trong nhiều nghiên cứu y học, ~~điển hình như công ti Plague Inc.~~

Chúng ta sẽ thử tìm hiểu ứng dụng củ công thức Bayesian trong y học qua bài toán sau:

Các nhà khoa học đã nghiên cứu thành công máy chuẩn đoán bệnh Trumpf, 1 bệnh cực kì dễ lây nhiễm nhưng không thể phát hiện người bệnh bằng mắt thường. Hiện tại họ dự đoán có khoảng 10% dân số ở thành phố A bị bệnh Trumpf và máy chuẩn đoán bệnh có tỉ lệ chính xác là 90% (tức một người bị bệnh đi chuẩn đoán thì 90% máy cho ra kết quả dương tính, 10% âm tính và ngược lại) Anh Tài

đi chuẩn đoán và ra kết quả dương tính, hỏi khả năng anh ấy thật sự bị bệnh là bao nhiêu?

Ta đặt H là giả thuyết anh ấy bị bệnh Trumpf, E là kết quả dương tính. Ta có: P(E|H) = 90%, P(H) = 10% (bỏ qua các yếu tố phụ như vị trí, độ tuổi,...), P(E)= 10%.90% +90%.10% (dù không bị bệnh vẫn có 10% anh Tài ra kết quả dương tính) = 18%. Vậy P(H|E) = 50%! Tuy máy chuẩn đoán này có vẻ rất hiệu quả, thật sự áp dụng vào thực tế lại không đem lại nhiều hiệu quả, thậm chí còn gây hoang mang cho những người không bị bệnh.

**Hãy nghĩ phương án để cải thiện vấn đề trên.

<u>d/Ứng dụng của công thức Bayesian nâng cao trong đời sống và luyện tập (tham khảo):</u>

-Luyện tập:

1/Hãy tính tỉ lệ thật sự bị bệnh Trumpf ở phần trên nếu anh Hoàng chuẩn đoán 3 lần dương tính? 2 lần đầu dương tính và lần cuối âm tính?

2/Giả sử 1 đứa trẻ có khả năng được sinh ra là nam hoặc nữ, sinh ra vào 1 trong 12 tháng là như nhau. Gọi p là khả năng 1 gia đình 2 con có 2 đứa con gái, biết rằng có ít nhất 1 đứa là con gái. Gọi q là khả năng 1 gia đình 2 con có 2 đứa con gái, biết rằng có ít nhất 1 đứa con gái sinh vào tháng 4. So sánh p và q.

3/Hưng thi Phổ Thông Năng Khiếu không được tốt lắm và rất lo về kết quả. Theo nguồn tin nội bộ, bạn của Hưng đã biết kết quả của 1 số bạn đã đậu vào trường nhưng Hưng từ chối hỏi thông tin vì sợ nó sẽ làm giảm khả năng của mình nếu không có anh trong đó. Anh Hưng có đúng hay không? Giải thích.

-Công thức Bayesian nâng cao:

1 ứng dụng khác của công thức Bayesian là để lọc thư spam trong e-mail, chúng ta hãy cùng xem ví dụ sau (Đáp số là thứ quan trọng nhất trong ví dụ nên nếu không đọc lời giải cũng không sao):

	Thư bình thường	Thư rác	Tổng cộng
Thư mẫu	400	600	1000

Có từ “miễn phí”	100	300	400
Có từ “thông báo”	10	90	100

Trong ví dụ trên, ta sẽ đặt R là khả năng thư đến là thư rác, B là là khả năng thư đến là thư bình thường, M là khả năng thư đến có từ “miễn phí”, T là khả năng thư đến có từ “thông báo”. Ta có:

$$P(R \mid M) = \frac{P(M \mid R)P(R)}{P(M)} = \frac{0,5.0,6}{0,4} = 75\%$$

$$P(R \mid T) = \frac{P(T \mid R)P(R)}{P(T)} = \frac{0,15.0,6}{0,1} = 90\%$$

Câu hỏi đặt ra bây giờ là : khả năng thư đến là thư rác là bao nhiêu nếu nó có cả 2 từ trên? Kí hiệu là P(R|M^T)

Hiển nhiên ta có:

$$P(R \mid M \wedge T) = \frac{P(M \wedge T \mid R)P(R)}{P(M \wedge T)}$$

Theo công thức xác suất, nếu 2 sự kiện X,Y không phụ thuộc vào nhau thì :

P(X^Y|Z) = P(X|Z).P(Y|Z)

Do đó, ta sẽ giả sử 2 sự kiện M,T không phụ thuộc vào nhau (có thể điều giả sử này sai nhưng sai số chấp nhận được so với việc đếm tổng số lần 2 từ này cùng xuất hiện)

(**Chấp nhận sai số nhỏ là điều rất quan trọng trong toán ứng dụng)

Do đó ta có :

P(M^T|R) = P(M|R).P(T|R)

P(M^T|B) = P(M|B).P(T|R)

Ta cũng có :

$$P(R \mid M \wedge T) = \frac{P(M \wedge T \mid R).P(R)}{P(M \wedge T)}$$

$$P(B \mid M \wedge T) = \frac{P(M \wedge T \mid B).P(B)}{P(M \wedge T)}$$

Do 1 thư chỉ có thể là thư bình thường hoặc thư rác :

$$1 = \frac{P(M \wedge T \mid R).P(R)}{P(M \wedge T)} + \frac{P(M \wedge T \mid B).P(B)}{P(M \wedge T)}$$

$$\Rightarrow P(M \wedge T) = P(M \mid R).P(T \mid R).P(R) + P(M \mid B).P(T \mid B).P(B)$$

Vậy ta đưa ra được kết luận :

$$P(R \mid M \wedge T) = \frac{P(M \mid R).P(T \mid R).P(R)}{P(M \mid R).P(T \mid R).P(R) + P(M \mid B).P(T \mid B).P(B)}$$

**Lưu ý : Để ý rằng đây chỉ là kết quả xấp xỉ

Thử ghép vào câu hỏi ban đầu, ta có được tỉ lệ thư có 2 từ “miễn phí” và “thông báo” là thư rác sẽ bằng :

$$\frac{0.6*0.5*0.15}{0.6*0.5*0.15+0.4*0.25*0.025}=95\%$$

Theo phương pháp trên ta rút ra được công thức tổng quát :

$$P(R\mid \wedge_{i=1}^{n} thongtin_i)=\frac{P(R)\prod P(thongtin_i\mid R)}{P(R)\prod P(thongtin_i\mid R)+P(B)\prod P(thongtin_i\mid B)}$$

(Nhớ rằng các thông tin không quá phụ thuộc vào nhau)

**Có thể tham khảo thêm các ứng dụng nâng cao của côn g thức Bayesian tại http://mlg.eng.cam.ac.uk/zoubin/talks/lect1bayes.pdf (không khuyến khích)

Hết

Nếu có câu hỏi về chủ đề này, có thể liên lạc thêm qua: manhphamnguyen2810@gmail.com