Bài 51: Gia nhập trường phái Bayes

Để bạn làm quen với trường phái phân tích dữ liệu Bayes ([25]) thì bài này sẽ giúp bạn ôn lại khái niệm Xác suất có điều kiện và làm rõ định lý Bayes.

Xác suất có điều kiện (conditional probability)

Cho 2 biến cố X và Y, xác suất có điều kiện của Y nếu X xảy ra được viết là:

$$P(Y \mid X)$$

Định lý Bayes

Định lý Bayes phát biểu như sau:

Xác suất Y xảy ra nếu X xảy ra bằng **tích của xác suất của X xảy ra nếu Y xảy ra** với **xác suất của Y** – kí hiệu là **P(Y).** Lấy tích đó chia cho **xác suất của X** – kí hiệu p(X).

$$P(Y \mid X) = \frac{P(X \mid Y) * P(Y)}{P(X)}$$

Tóm tắt các kí hiệu:

- X và Y là hai biến cố.
- P(Y | X): Xác suất có điều kiện của Y dựa trên X. Hoặc xác suất Y xảy ra nếu X là đúng.
- P(X | Y): Tương tự kí hiệu ở trên. Đây là xác suất X xảy ra nếu Y là đúng.
- P(Y): là xác suất của Y. Đây là xác suất độc lập
- P(X): là xác suất của X. Đây là xác suất độc lập.

Tình huống minh họa

Để hiểu được định lý Bayes thì chúng ta bàn một chút về tình huống thực tế trong Y Khoa – Vấn đề chuẩn đoán ung thư phổi.

Để biết một người có ung thư phổi hay không thì bác sĩ cần lấy mẫu để đi sinh thiết và kết luận là có ung thư hay không?

Tìm hiểu thêm một chút về khái niệm sinh thiết:

Sinh thiết là một quá trình lấy mẫu mô hoặc tế bào khỏi cơ thể để kiểm tra dưới kính hiển vi. Sinh thiết phổi là quá trình lấy mẫu từ mô phổi (bằng kim sinh thiết đặc biệt hoặc trong quá trình phẫu thuật) để kiểm tra xem phổi bị bệnh lý gì, và có sự tồn tại của ung thư hay không.

[25] https://en.wikipedia.org/wiki/Thomas_Bayes

Tạm thời coi phương pháp sinh thiết phổi là tiêu chuẩn vàng để kết luận là người đó có ung thư phổi hay không? Như vậy khi nói người bị ung thu phổi có nghĩa là kết quả sinh thiết cho kết luận là ung thư.

Ngoài ra để tầm soát ung thư phổi thì có thể dùng phương pháp chụp ảnh cắt lớp (gọi tắt là chụp ảnh phổi, trong bài viết này gọi là xét nghiệm). Tôi phịa (có tham khảo) ra vài thông tin bên dưới để chúng ta hiểu định lý Bayes nhé.

Chúng ta có vài thông tin sau:

1) Các nghiên cứu cho thấy tỉ lệ dân số Việt Nam mắc bệnh ung thư phổi là 4%.

2) Cứ 100 người ung thư phổi mà đi chụp ảnh phổi thì có 80 người cho kết quả là dương tính (kí hiệu +ve).

3) Cứ 100 người KHÔNG ung thư phổi mà đi chụp ảnh phổi thì có 90 người cho kết quả âm tính (kí hiệu -ve)

Bàn luận:

- Theo ý 2) ở trên thì có 20 trong 100 người bị ung thư phổi không được phát hiện bằng cách xem ảnh phổi. 20 người gọi là âm tính giả. Đây là một tình huống sai, gọi là **âm tính giả**; sai số 20%.

- Theo ý 3) ở trên thì có 10 người bình thường khi chụp ảnh phổi sẽ được cho kết quả là dương tính (+ve). Đây cũng là một tính huống sai, gọi là dương tính giả; sai số 10%.

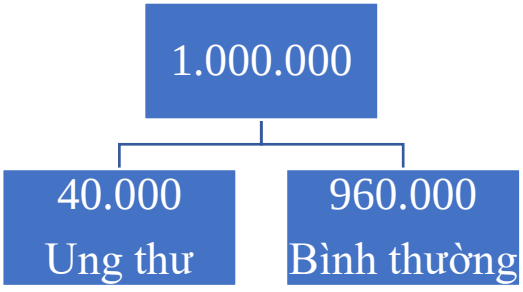
  Tình huống là nếu 1 người nào đó đi chụp phổi và được chẩn đoán là dương tính thì xác xuất bị ung thư phổi là bao nhiêu?

  Kí hiệu P(cancel | +ve)

  Phương pháp suy luận theo tầng số: sử dụng chỉ số P, khoảng tin cậy 95%.

Ví dụ cụ thể

Xét một cộng đồng gồm 1 triệu người Việt Nam với thông tin số 1) ở trên thì sẽ có 4% bị ung thư phổi, còn lại tạm cho là bình thường. Xem sơ đồ sau:



Nếu trong 40.000 người bị ung thư này mà đi chụp phổi thì sẽ có 80% sẽ cho kết quả dương tính (+ve). Số 80% hay 0.80 gọi là độ nhạy (**Sensitivity**):

Sensitivity = 0.80

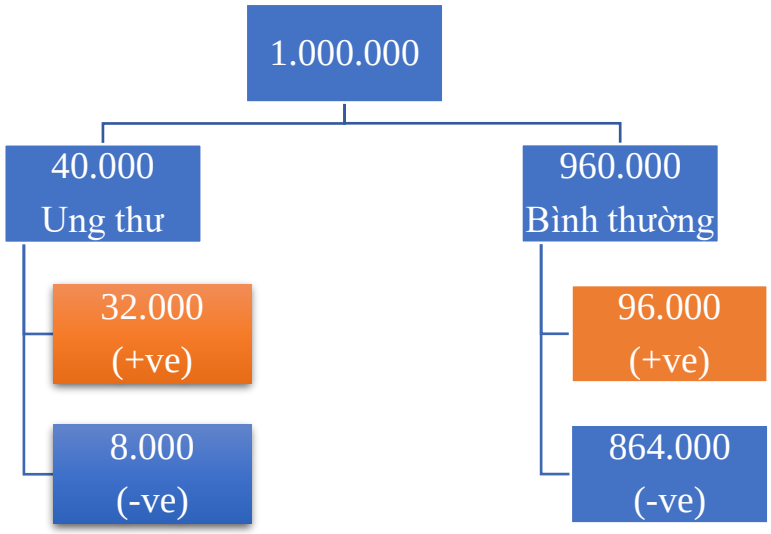
Tức là trong 40.000 người **ung thư** thì có 80% x 40.000 = 32.000 sẽ có kết quả xét nghiệm dương tính (+ve); và 20% x 40.000 = 8.000 người cho **kết quả âm tính**.

Nếu trong 960.000 người bình thường nếu đi chụp phổi thì sẽ có 10% sẽ cho kết quả dương tính (+ve). Trường hợp này gọi là **dương tính giả**. Số 10% hay 0.10 gọi là **độ đặc hiệu** (Specificity).

Specificity = 0.10

Tức là trong 960.000 người được cho là **bình thường** nếu đi xét nghiệp thì sẽ có 10% x 960.000 = 96.000 người sẽ cho **kết quả dương tính**; và có 90% x 960.000 = 864.000 người sẽ cho **kết quả âm tính**.

Cập nhật sơ đồ như sau:



Như vậy tổng số người dương tính (+ve) là 32.000 + 96.000 = **128.000 người dương tính**. Trong đó thì chỉ có **32.000 người là ung thư**.

Nếu một xét nghiệm cho kết quả là dương tính thì xác suất người đó ung thư được viết dạng công thức là:

P (cancer | +ve) = 32.000 / 128.000 = **0.25 = 25%**

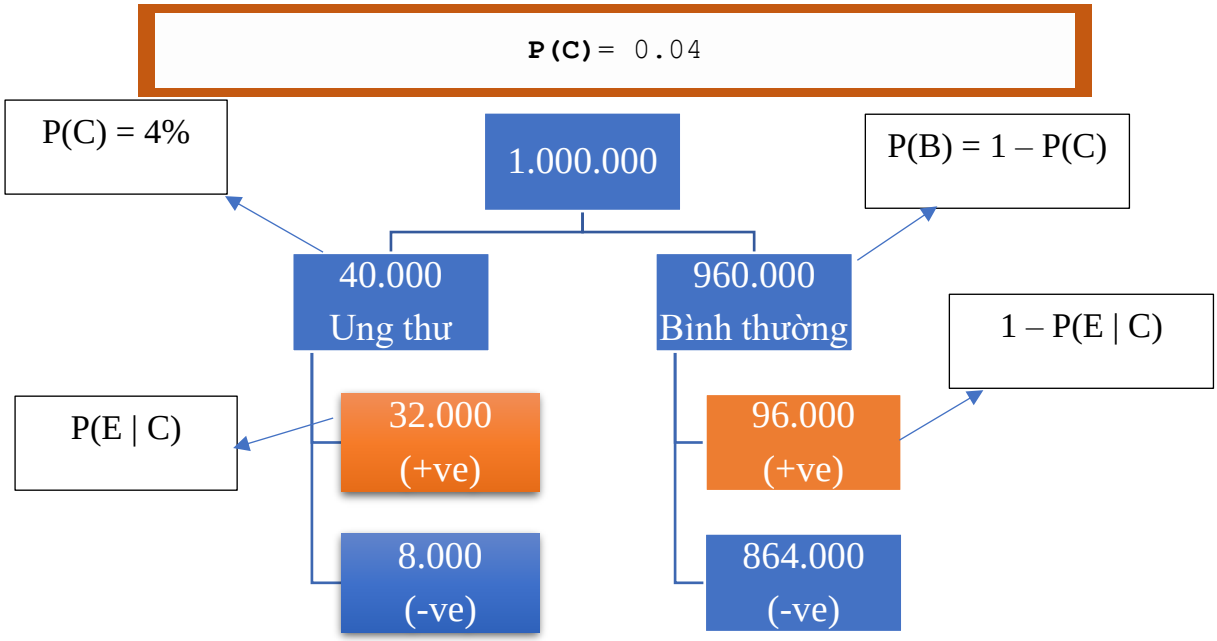
Nhu vậy có thể phát hiểu:

```
Nếu một người đi xét nghiệm phổi mà cho kết quả dương tính thì
xác suất người này bị ung thư phổi là 25%.
```

Độ nhạy và độ đặc hiệu (Sensitivity & Specificity)

Tóm tắt lại tình huống ở trên dưới dạng các công thức sau đây.

Thông tin xác suất đã có từ các nghiên cứu trước đây gọi là **xác suất tiền định** (prior probability). Gọi C là biết cố ung thư, thì xác suất của C đã biết trong cộng đồng kí hiệu là:



Gọi E là biến cố "Dương tính".

Sensitivity

P(E | C) là xác suất dương tính nếu đã biết chắc là ung thư (C là đúng).

Độ nhạy được tính là: P(E | C)

```
P(E | C) = 0.08
```

Specificity

Gọi B là biến cố "bình thường" tức là không ung thư.

```
P(B) = 1 – P(C)
```

P(E | B) là xác suất **dương tính** nếu biết người đó được xem là bình thường.

```
P(E | B) = 0.10
```

Độ đặc hiệu chính là **xác suất KHÔNG dương tính giả**, được tính bằng:

```
Specificity = 1 -  P(E | B)
```

Tóm tắt ý tưởng Bayes

Phương pháp Bayes **sử dụng lại kiến thức đã có** và **dữ liệu thực tế** để **suy luận ra thông tin mới**. Cụ thể là:

- Dựa vào kiến thức trước đây: tỉ lệ mắc bệnh ung thư phổi trong Việt Nam là 4%. Khi có một bệnh nhân đến gặp bác sĩ nhờ khám phổi thì bác sĩ chỉ biết có mỗi thông tin là 4% ở trên. Thông tin này gọi là **prior knowledge**.

- Dữ liệu thực tế: Bác sĩ định là đi chụp phim phổi. Dữ liệu này gọi là **data**. Và bác sẽ cũng có thêm thông tin độ nhạy, độ đặc hiệu ([26]): hai thông tin này gọi là **likelihood**.

- Trên cơ sở thông tin **prior knowledge, data, likelihook** thì bác sĩ có thể suy luận ra thông tin mới (gọi là posterior knowledge)– xác suất hậu định để chẩn đoán cho bệnh nhân: xác suất ung thư là bao nhiêu.

Tóm làm ý tưởng của phương pháp Bayes là:

```
Thông tin sẵn có + Dữ liệu mới => Thông tin mới
```

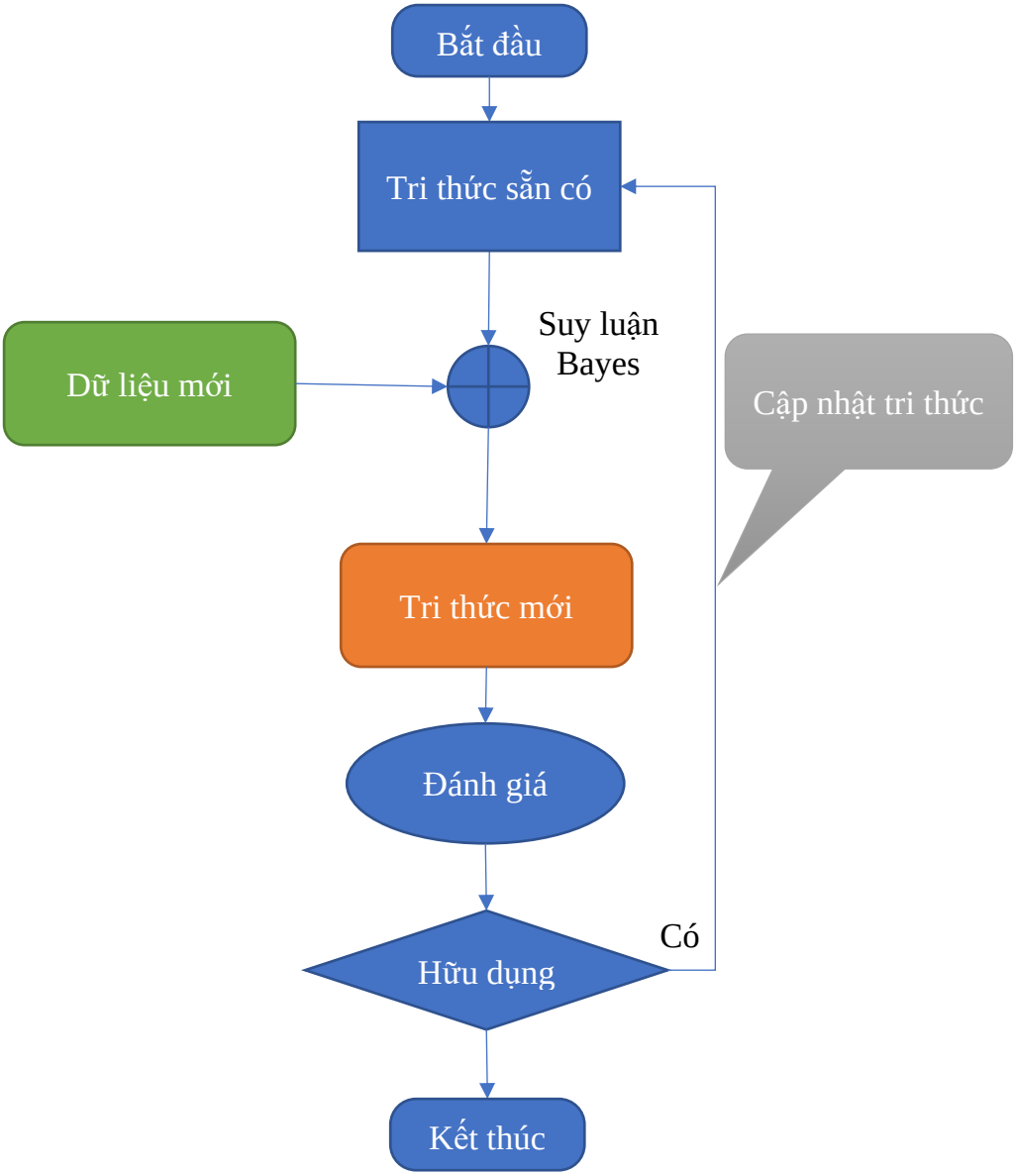
Hoặc văn hoa hơn thì:

```
Tri thức sẵn có + Dữ liệu mới => Trí thức mới
```

Công thức này rất phù hợp cho các thuật toán Machine learnining. Tức là cứ nạp thi tri thức sẵn có và cài đặt thuật toán Bayes cho cái máy. Mỗi lần có thông tin mới được nạp vào máy thì máy sẽ cho ra trí thức mới. Sau đó tri thức mới sẽ được con người kiểm tra nếu thấy hữu dụng thì nạp lại vào máy coi là tri thức sẵn có. Cứ như vậy cái máy sẽ ngày càng thông minh hơn.

Sơ đồ hóa chu trình cho cái máy nó học như sau:

[26] Thường thì thông tin này được cập nhật thông qua nghiên cứu thực tế, trong đó có liên quan đến cái máy và phương pháp chụp phim phổi



Học liệu tham khảo:

[1] Nguyễn Văn Tuấn - Bài giảng 53: Phương pháp Bayes: nhập môn - https://youtu.be/kbog4CvKaM8