20200211【Data Analyst Nanodegree】P04M01L07

> Part 04 : Practical Statistics
> Learn how to apply inferential statistics and probability to important, real-world scenarios, such as analyzing A/B tests and building supervised learning models.

Module 01: Practical Stats

Lesson 07: Bayes Rule

> Learn one of the most popular rules in all of statistics – Bayes rule.

01. Bayes Rule

We're going to talk about perhaps the holy grail of probabilistic inference. It's called **Bayes Rule(贝叶斯定理,Bayes's Theorem/Bayes's Law)**.

Bayes Rule is based on **Reverend Thomas Bayes**, who used this principle to infer the existence of God, but in doing so, he created a new family of methods that has vastly influenced artificial intelligence and statistics.

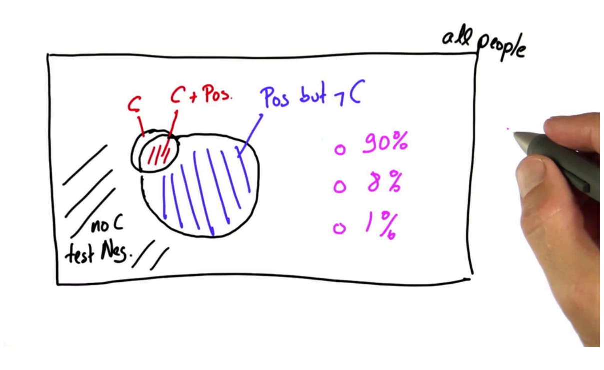
02. Cancer Test

The question being asked is this:

1% of the population has cancer. Given that there is a 90% chance that you will test positive if you have cancer and that there is a 90% chance you will test negative if you don't have cancer, what is the probability that you have cancer if you test positive?

$P(C) = 0.01$

The answer is 8%



03. Prior And Posterior

Bayes Rule

So this is **the essence of Bayes Rule**, which I'll give to you to you in a second. There's some sort of a prior, of which we mean the probability before you run a test, and then you get some evidence from the test itself.

and that all leads you to what's called a **Posterior Probability(后验概率)**. Now this is not really a plus operation. In fact, in reality, it's more like a multiplication, but semantically, what Bayes Rule does is it incorporates some evidence from the test into your **Prior Probability(先验概率)** to arrive at a posterior probability.

$\text{Prior Probability} \cdot \text{Test Evidence} \rightarrow \text{Posterior Probability}$

Prior:

$P(C) = 0.01 = 1\%$

$P(\neg C) = 0.99 = 99\%$

$P(Pos|C) = 0.9 = 90\%$

$P(Neg|C) = 0.1 = 10\%$

$P(Pos|\neg C) = 0.1 = 10\%$

$P(Neg|\neg C) = 0.9 = 90\%$

Posterior

$P(C|Pos) = P(C)P(Pos|C) = 0.01 \cdot 0.9 = 0.09$

$P(\neg C|Pos) = P(\neg C)P(Pos|\neg C) = 0.99 \cdot 0.1 = 0.099$

04. Normalizing 1

The **normalization** proceeds in two steps. We just normalized these guys to keep ratio the same but make sure they add up to 1.

Prior:

$P(C) = 0.01 = 1\%$

$P(\neg C) = 0.99 = 99\%$

$P(Pos|C) = 0.9 = 90\%$

$P(Neg|C) = 0.1 = 10\%$

$P(Pos|\neg C) = 0.1 = 10\%$

$P(Neg|\neg C) = 0.9 = 90\%$

Joint Probability(联合概率)

$P(C, Pos) = P(C)P(Pos|C) = 0.01 \cdot 0.9 = 0.09$

$P(\neg C, Pos) = P(\neg C)P(Pos|\neg C) = 0.99 \cdot 0.1 = 0.099$

Normalizer(归一化):

$P(Pos) = P(C, Pos) + P(\neg C, Pos) = 0.009 + 0.099 = 0.108$

05. Normalizing 2

Prior Probability:

$P(C) = 0.01 = 1\%$

$P(\neg C) = 0.99 = 99\%$

$P(Pos|C) = 0.9 = 90\%$

$P(Neg|C) = 0.1 = 10\%$

$P(Pos|\neg C) = 0.1 = 10\%$

$P(Neg|\neg C) = 0.9 = 90\%$

Joint Probability

$P(C, Pos) = P(C)P(Pos|C) = 0.01 \cdot 0.9 = 0.09$

$P(\neg C, Pos) = P(\neg C)P(Pos|\neg C) = 0.99 \cdot 0.1 = 0.099$

Normalizer

$P(Pos) = P(C, Pos) + P(\neg C, Pos) = 0.009 + 0.099 = 0.108$

Posterior Probability

$P(C|Pos) = 0.009/0.108 = 0.0833$

$P(\neg C|Pos) = ?$

06. Normalizing 3

Posterior Probability

$P(C|Pos) = 0.009/0.108 = 0.0833$

$P(\neg C|Pos) = 0.099/0.108 = 0.9167$

07. Total Probability

Posterior Probability

$P(C|Pos) = 0.009/0.108 = 0.0833$

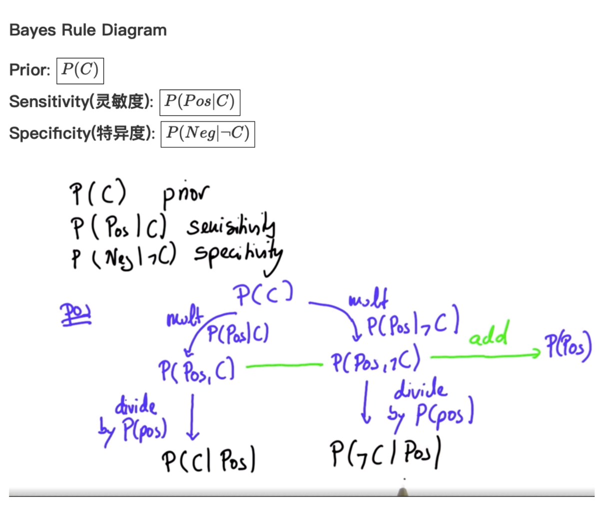
$P(\neg C|Pos) = 0.099/0.108 = 0.9167$

$P(C|Pos) + P(\neg C|Pos) = 1$

08. Bayes Rule Diagram

Bayes Rule Diagram

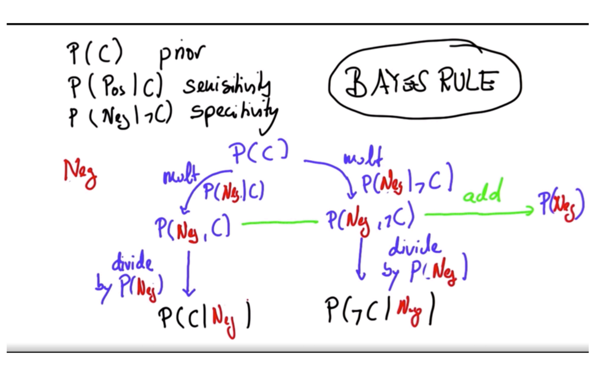
Prior: $\boxed{P(C)}$
Sensitivity(灵敏度): $\boxed{P(Pos|C)}$
Specificity(特异度): $\boxed{P(Neg|\neg C)}$



09. Equivalent Diagram

Equivalent Diagram

Prior: $\boxed{P(C)}$
Sensitivity(灵敏度): $\boxed{P(Pos|C)}$
Specificity(特异度): $\boxed{P(Neg|\neg C)}$



$P(Pos, C) = P(Pos|C)P(C)$

$P(Neg, C) = P(Neg|C)P(C)$

$P(Pos, \neg C) = P(Pos|\neg C)P(\neg C)$

$P(Neg, \neg C) = P(Neg|\neg C)P(\neg C)$

10. Cancer Probabilities

Given

Prior: $\boxed{P(C)}$
Sensitivity(灵敏度): $\boxed{P(Pos|C)}$
Specificity(特异度): $\boxed{P(Neg|\neg C)}$

Quiz

$P(\neg C) = 1 - 0.01 = 0.99$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.9 = 0.1$

11. Probability Given Test

Given

$P(C) = 0.01$

$P(Pos|C) = 0.9$

$P(Neg|\neg C) = 0.9$

$P(\neg C) = 1 - 0.01 = 0.99$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.9 = 0.1$

Quiz

$P(C, Neg) = P(C)P(Neg|C) = 0.01 \cdot 0.1 = 0.001$

$P(\neg C, Neg) = P(\neg C)P(Neg|\neg C) = 0.99 \cdot 0.9 = 0.891$

12. Normalizer

Given

$P(C) = 0.01$

$P(Pos|C) = 0.9$

$P(Neg|\neg C) = 0.9$

$P(\neg C) = 1 - 0.01 = 0.99$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.9 = 0.1$

$P(C, Neg) = P(C)P(Neg|C) = 0.01 \cdot 0.1 = 0.001$

$P(\neg C, Neg) = P(\neg C)P(Neg|\neg C) = 0.99 \cdot 0.9 = 0.891$

Quiz

$P(Neg) = P(C, Neg) + P(\neg C, Neg) = P(Neg|C)P(C) + P(Neg|\neg C)P(\neg C) = 0.1 \cdot 0.01 + 0.9 \cdot 0.99 = 0.892$

13. Normalizing Probability

Given

$P(C) = 0.01$

$P(Pos|C) = 0.9$

$P(Neg|\neg C) = 0.9$

$P(\neg C) = 1 - 0.01 = 0.99$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.9 = 0.1$

$P(C, Neg) = P(C)P(Neg|C) = 0.01 \cdot 0.1 = 0.001$

$P(\neg C, Neg) = P(\neg C)P(Neg|\neg C) = 0.99 \cdot 0.9 = 0.891$

$P(Neg) = P(C, Neg) + P(\neg C, Neg) = P(Neg|C)P(C) + P(Neg|\neg C)P(\neg C) = 0.1 \cdot 0.01 + 0.9 \cdot 0.99 = 0.892$

Quiz

Please specify your answer to at least **4 decimal places**, rounded properly.

$P(C|Neg) = P(C, Neg)/P(Neg) = 0.001/0.892 = 0.0011$

$P(\neg C|Neg) = P(\neg C, Neg)/P(Neg) = 0.891/0.892 = 0.9989$

Now, what's remarkable about this outcome is really what it means.

Before the test, we had a 1% chance of having cancer, now, we have about a 0.9% chance of having cancer. So, a cancer probability **went down** by about a factor of 9. So, the test really helped us gaining confidence that we are cancer-free.

Conversely, before we had a 99% chance of being cancer free, now it's 99.89%. So, all the numbers are **working exactly how we expect them to work**.

14. Disease Test 1

Given

$P(C) = 0.1$

$P(Pos|C) = 0.9$

$P(Neg|\neg C) = 0.5$

Quiz

$P(\neg C) = 1 - 0.1 = 0.9$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.5 = 0.5$

$P(C, Neg) = P(C)P(Neg|C) = ?$

$P(\neg C, Neg) = P(\neg C)P(Neg|\neg C) = ?$

$P(Neg) = P(C, Neg) + P(\neg C, Neg) = P(Neg|C)P(C) + P(Neg|\neg C)P(\neg C) = ?$

$P(C|Neg) = P(C, Neg)/P(Neg) = ?$

$P(\neg C|Neg) = P(\neg C, Neg)/P(Neg) = ?$

15. Disease Test 2

--snip--

16. Disease Test 3

--snip--

17. Disease Test 4

--snip--

18. Disease Test 5

Quiz

$P(C, Neg) = P(C)P(Neg|C) = 0.1 \cdot 0.1 = 0.01$

$P(\neg C, Neg) = P(\neg C)P(Neg|\neg C) = 0.9 \cdot 0.5 = 0.45$

$P(Neg) = P(C, Neg) + P(\neg C, Neg) = P(Neg|C)P(C) + P(Neg|\neg C)P(\neg C) = 0.1 \cdot 0.1 + 0.9 \cdot 0.5 = 0.46$

$P(C|Neg) = P(C, Neg)/P(Neg) = 0.01/0.46 = 0.0217$

$P(\neg C|Neg) = P(\neg C, Neg)/P(Neg) = 0.45/0.46 = 0.9783$

19. Disease Test 6

Given

$P(C) = 0.1$

$P(Pos|C) = 0.9$

$P(Neg|\neg C) = 0.5$

$P(\neg C) = 1 - 0.1 = 0.9$

$P(Neg|C) = 1 - 0.9 = 0.1$

$P(Pos|\neg C) = 1 - 0.5 = 0.5$

Quiz

$P(C, Pos) = P(C)P(Pos|C) = 0.1 \cdot 0.9 = 0.09$

$P(\neg C, Pos) = P(\neg C)P(Pos|\neg C) = 0.9 \cdot 0.5 = 0.45$

$P(Pos) = P(C, Pos) + P(\neg C, Pos) = P(Pos|C)P(C) + P(Pos|\neg C)P(\neg C) = 0.54$

$P(C|Pos) = P(C, Pos)/P(Pos) = 0.09/0.54 = 0.1667$

$P(\neg C|Pos) = P(\neg C, Pos)/P(Pos) = 0.45/0.54 = 0.833$

20. Bayes Rule Summary

In **Bayes rule**, we have a **hidden variable** we care about--whether they have cancer or not. But **we can't measure it directly and instead we have a test**.

We have a prior of how frequent this variable is true and the test is generally characterized by how often it says positive when the variable is true and how often it is negative and the variable is false.

Bayes' rule looks like this:

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

Bayes rule takes a prior, **multiplies** in the measurement, which in this case we assume to be the positive measurement to give us a new variable and does the same for all actual measurement, given the opposite assumption about our hidden variable of cancer and that multiplication gives us this guy over here.

We **add** those two things up and then it gives us a new variable and then we **divide** these guys to arrive the **best estimate of the hidden variable c** given our test result.