Bayesian Learning

By Hayden Gemeinhardt

1. Bayes' Theorem

Introduced by Revered Thomas Bayes, Bayes' Theorem calculates the probability of an event given some prior knowledge. Mathematically, this is found by the following equation:

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

Where:

P(A) is the probability of A and P(B) is the probability of B

P(A|B) is the probability of A given B and P(B|A) is the probability of B given A

2. Proof [1]

The theorem can be derived by the following definitions:

Definition of conditional probability:

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

Definition of joint probability:

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

From conditional probability:

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

Combining with joint probability:

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

And then plugging back into the conditional probability, you get the final result:

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

3. Applications of Bayes' Theorem

Let us say someone is in a difficult class, and the likelihood of achieving a good score is 40% based on past students. At first look, they may be worried that there is a 60% chance they will get a bad grade.

However, the probability that those past students studied is 50%, and of those students who got a good score, 90% of them studied.

P(Good score) = 40%

P(Studied) = 50%

P(Studied | Good score) = 90%

P(Good score | studied) = 0.9*0.4/0.5 = 72%

So, if that person studies, there is a 72% chance they will get a good score. That is a whole lot better than 50%!

4. Further Break Down Of Bayes' Theorem [2]

One can break down Bayes' Theorem even further.

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

Often times in the real world, one needs to find P(B) to begin with, thus the above formula combines the two steps into one (finding P(B) and then P(A|B)).

A very popular example of this is used in medical tests, as these tests can provide four possible results—true positive, false positive, true negative, and false negative—instead of only two.

For example, 1% of people will get a particular disease. When testing for this disease, 90% of the tests detect it when it is there (true positive) and 9% detect it when it is not there (false positive).

One cannot then say there is a 90% chance a patient has the disease when tested positive because it fails to account for the 9% false positives. In other words, we have to calculate P(B) first before we calculate P(A|B).

P(A) = Patient has disease

P(B) = Test returned positive

P(test returned positive) = P(true positive)+P(false positive) = 1%*90% + 99%*9% = 9.81%

P(disease is there|test returned positive) = 90%*1% / 9.81% = 9.17%

Thus, there is only a 9.17% chance the patient actually has the disease given a test result returned positive.

A major common misconception in the medical field is based in the misunderstanding of the fundamentals presented by Bayes' Theorem. [5]. Gerd Gigerenzer, director of the Harding Center for Risk Literacy in Berlin, did a series of seminars in which he asked doctors this same question.

> 'In one session, almost half the group of 160 gynecologists responded that the woman's chance of having cancer was nine in 10. Only 21% said that the figure was one in 10 - which is the correct answer.' [3]

This can be described as a veridical paradox, or a statement that appears to be wrong but can be shown to be true, and is the reason people like Gigerenzer have been pushing for better teaching of Bayes' Theorem in education.

5. Classification Using Bayes Theorem

I. Naïve Bayes [4]

Classification is a form of predictive modeling that assigns a class or label to an input (such as a computer detecting and labeling different objects in a photo).

Bayes Theorem is a dependent conditional probability model which lies under the assumption that each input variable is dependent on all other variables. However, finding P(B|A) for each input is not feasible unless the training data is large enough to find the probability distribution for every possible combination of values (which is rarely the case).

If we consider each input variable to be **independent** (say an apple being in the picture is not dependent on a plane being in the picture), then one can calculate each P(Bn|A) **separately** and multiply them together to get an overall probability distribution for the combination, then use that to calculate P(A|B).

In other words:

$P(A|B1,B2,\ldots,Bn) = P(B1|A)*P(B2|A)*\ldots*P(Bn|A)*P(A) / P(B)$

And since P(B) is constant for each calculation, we can drop it to get:

$P(A|B1,B2,\ldots,Bn) = P(B1|A)*P(B2|A)*\ldots*P(Bn|A)*P(A)$

This simplification of Bayes' Theorem is called Naïve Bayes

Unfortunately, Naïve Bayes relies on the assumption that each variable is independent, which is not always the case in a real world scenario.

I. Bayes Optimal Classifier [5]

Another approach called the Bayes Optimal Classifier allows the user to specify which attributes are independent in the case that some amount of the variables are dependent. No other classification method can outperform Bayes Optimal Classifier given the same prior knowledge and hypothesis space, but the method is very computationally expensive (which is why Naïve Bayes is often used in place of it). It can also be used as a regression model.

The hypothesis of Bayes Optimal can be described in the equation:

$$h_{BO} = \max \left[\sum_{h_i \in H} P(v_j | h_i) P(h_i | D) \right]$$

where V is the set of possible classifications, h is a hypothesis in the hypothesis space H, D is the set of training data. This is finding the best (max) hypothesis from the probability of a classification given a hypothesis times the probability of a hypothesis given the training data.

6. Example In Robotics [5]

Say a robot is trying to determine the best path to take. The set V consists of left, right, or forward, and there are 5 hypothesis h1 through h5. There is also a training set D.

H	$P(h_i|D)$	$P(Forward|h_i)$	$P(Left|h_i)$	$P(Right|h_i)$
h_1	0.1	1	0	0
h_2	0.2	0	1	0
h_3	0.1	1	0	0
h_4	0.3	0	0	1
h_5	0.2	0	1	0
h_6	0.1	1	0	0

$$\sum_{h_i \in H} P(Forward|h_i)P(h_i|D)] = 0.1 + 0.1 + 0.1 = 0.3$$

$$\sum_{h_i \in H} P(Left|h_i)P(h_i|D)] = 0.2 + 0.2 = 0.4$$

$$\sum_{h_i \in H} P(Right|h_i)P(h_i|D) = 0.3$$

$$h_{BO} = \max\,[\sum_{h_i \in H} P(v_j|h_i)P(h_i|D)] = \max\,[0.3,0.4,0.3] = 0.4 = \text{Left}$$

This is simply giving each hypothesis, given the training data, a weight and finding which classification or direction has the most weight based on the suggestions from each hypothesis.

However, if one were to take the direction which had the most hypothesis supporting it, that direction would be forward. And if one took the hypothesis with the highest weight, that direction would be turning right.

The extra calculation provides a sturdier overall hypothesis, though requires a lot more computation.

7. Conclusion

In this paper, we looked at Bayes' Theorem, its derivation, and its use in statistics for practical, real world scenarios. We also found how further applying the theorem can create Naïve Bayes and Optimal Bayes Classification systems utilized in everyday machine learning, each with its advantages and disadvantages.

Sources

Admin. (2020, November 05). Bayes theorem (Statement, Proof, derivation, and examples). Retrieved April 05, 2021, from https://byjus.com/maths/bayes-theorem/

[1] Stuart, A.; Ord, K. (1994), Kendall's Advanced Theory of Statistics: Volume I—Distribution Theory, Edward Arnold, §8.7

[2] An intuitive (and short) explanation of bayes' theorem. (n.d.). Retrieved April 05, 2021, from https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

[3] Kremer, W. (2014, July 06). Do doctors understand test results? Retrieved April 05, 2021, from https://www.bbc.com/news/magazine-28166019

[4] Brownlee, J. (2019, December 03). A gentle introduction to bayes theorem for machine learning. Retrieved April 05, 2021, from https://machinelearningmastery.com/bayes-theorem-for-machine-learning/

[5] https://web.cs.ucdavis.edu/~vemuri/classes/ecs271/Bayes.pdf