

Bayes' Theorem and Naive Bayes' Classification

Conditional Probability

The **conditional probability** of an event B in relationship to an event A is the probability that event B occurs given that event A has already occurred. The notation for conditional probability is $P(B|A)$, read as *the probability of B given A*.

When two events, A and B are dependent, the probability of both occurring is given by the multiplication rule:

$$P(A \text{ and } B) = P(A).P(B|A).$$

The “and” is usually dropped and replaced with a comma; that is the rule is written:-

$$P(A,B) = P(A).P(B|A).$$

Rearranging this we get:-

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

Example: In the local chippy, the probability that someone buys mushy peas and battered fish together is 0.05. The probability of buying battered fish is 0.2. What is the probability that someone buys mushy peas given they buy battered fish?

Solution: $P(\text{peas}|\text{fish}) = \frac{P(\text{fish,peas})}{P(\text{fish})} = \frac{0.05}{0.2} = 0.25$

Problems such as these where the probabilities are already provided for substitution into a formula are easy to solve. How do we determine these probabilities if we are given only the raw data? We will consider this next in a games context using discrete probabilities.

Imagine we have statistics describing the past performance of a fighter, let's call her Lara. In particular we have data for the last fourteen fights she had (win or lose) against different character types (ogre, human, orc) - see table 1 below.¹

¹ The raw data used for this example is taken from the famous “do I play tennis today?” Naïve Bayes' classifier, however it's been placed in a different context.

Fight Number	Character Type	Win?
1	orc	no
2	orc	no
3	orc	yes
4	orc	yes
5	human	yes
6	human	no
7	orc	yes
8	ogre	no
9	ogre	yes
10	human	yes
11	ogre	yes
12	human	yes
13	orc	yes
14	human	no

Table 1a

Table 1a is the raw data listing the 14 fights, whether Lara won or lost and the character she fought against. Table 1b summarizes this data and table 1c represents the joint probability distribution based on table 1b.

Character	Win	lose	
orc	4	2	6
ogre	2	1	3
human	3	2	5
	9	5	14

Table 1b

Joint Probability Table

Character	Win	lose	
orc	4/14	2/14	6/14
ogre	2/14	1/14	3/14
human	3/14	2/14	5/14
	9/14	5/14	1

Table 1c

Marginal probabilities:-

The word marginal comes from the position of these values at the margins of table 1c – notice they both sum to one, i.e. $9/14 + 5/14 = 1$ and $6/14 + 3/14 + 5/14 = 1$. The marginal probability gives the probability of an event, for example the probability of fighting an orc, $P(\text{Character}=\text{orc})$ is $6/14$ and the probability of winning, $P(\text{Win}=\text{true})$ is $9/14$. Usually we simply write $P(\text{orc})$ instead of $P(\text{Character}=\text{orc})$. Note the use of upper case letters for the variable and lower case for the value it takes for any particular problem.

Probability of winning, $P(\text{win}) = 9/14$,
 Probability of losing, $P(\text{losing}) = 5/14$,
 Probability of fighting an orc, $P(\text{orc}) = 6/14$,
 Probability of fighting an ogre, $P(\text{ogre}) = 3/14$,
 Probability of fighting a human $P(\text{human}) = 5/14$.

Joint probabilities:-

The probability that Lara won when she fought an orc is the simple joint (AND) probability obtained by multiplying the probability that she fought an orc ($6/14$) with the probability that she won ($4/6$), i.e. $6/14 \times 4/6 = 4/14$. Not surprisingly, this figure is found at the intersection of the row “orc” and column “win” in the joint probability table of figure 1c. Here’s all the joint probabilities for table 1c:

Probability of winning and opponent was an orc, $P(\text{win,orc}) = 4/14$,
 Probability of winning and opponent was an ogre, $P(\text{win,ogre}) = 2/14$,
 Probability of winning and opponent was an human, $P(\text{win,human}) = 3/14$,
 Probability of losing and opponent was an orc, $P(\text{lose,orc}) = 2/14$,
 Probability of losing and opponent was an ogre, $P(\text{lose,ogre}) = 1/14$,
 Probability of losing and opponent was a human, $P(\text{lose,human}) = 2/14$.

Conditional Probabilities:-

Note the word “given” here.

We can think of a conditional probability, for example $P(\text{win}|\text{orc})$ as a subset of our sample space. Here, within the set of all characters we have 6 orcs, 3 ogres and 5 humans so our conditional probability of winning given the opponent is an orc is based on the subset of 6 orcs and there were 4 wins out of 6 fights against orcs, hence $P(\text{win}|\text{orc}) = 4/6$.

Probability of winning, given opponent was an orc, $P(\text{win}|\text{orc}) = 4/6$,
Probability of winning, given opponent was an ogre, $P(\text{win}|\text{ogre}) = 2/3$,
Probability of winning, given opponent was a human, $P(\text{win}|\text{human}) = 3/5$,
Probability of losing, given opponent was an orc, $P(\text{lose}|\text{orc}) = 2/6$,
Probability of losing, given opponent was an ogre, $P(\text{lose}|\text{ogre}) = 1/3$,
Probability of losing, given opponent was a human, $P(\text{lose}|\text{human}) = 2/5$.

Probability that opponent was an orc, given Lara won $P(\text{orc}|\text{win}) = 4/9$,
Probability that opponent was an ogre, given Lara won $P(\text{ogre}|\text{win}) = 2/9$,
Probability that opponent was a human, given Lara won $P(\text{human}|\text{win}) = 3/9$,
Probability that opponent was an orc, given Lara lost $P(\text{orc}|\text{lose}) = 2/5$,
Probability that opponent was an ogre, given Lara lost $P(\text{ogre}|\text{lose}) = 1/5$,
Probability that opponent was a human, given Lara lost $P(\text{human}|\text{lose}) = 2/5$.

The joint, marginal and conditional probabilities listed above are derived from the table by directly reading the relevant cell, row and column. We can also use the multiplication rule to calculate the joint probabilities from the marginal and conditional probabilities just as we did in our earlier experiment, for example, recall $P(A, B) = P(A) \cdot P(B|A)$. So,

$P(\text{Character}=\text{orc}, \text{Win}=\text{true}) = P(\text{Character}=\text{orc}) \cdot P(\text{Win}=\text{true} | \text{Character}=\text{orc})$
 $= (6/14) \cdot (4/6)$
 $= 4/14$,
and equivalently,

$P(\text{Win}=\text{true}, \text{Character}=\text{orc}) = P(\text{Win}=\text{true}) \cdot P(\text{Character}=\text{orc} | \text{Win}=\text{true})$
 $= 9/14 \cdot 4/9$
 $= 4/14$,

which agrees with the entry in the joint probability table 1c above (where the orc row meets the win column). We might use this approach if we have incomplete data, for example if we want to find the joint probability but have only conditional probabilities and marginal probabilities.

The marginal probabilities can also be calculated using the law of total probability. This says that the probability of A is the sum of the conditional probabilities of A given B_i , each weighted by $P(B_i)$. This is easily written as:

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

For example if $A = \text{orc}$, $B_1 = \text{win}$ and $B_2 = \text{lose}$, then

$$\begin{aligned} P(\text{orc}) &= P(\text{orc}|\text{win})P(\text{win}) + P(\text{orc}|\text{lose})P(\text{lose}) \\ &= 4/9 \cdot 9/14 + 2/5 \cdot 5/14 = \frac{4+2}{14} = 6/14 \end{aligned}$$

We can also use the conditional probabilities we derived from the table to help us decide whether to back Lara in her next fight. For example, if she is to fight an orc next then we can look up the probability of her winning given her opponent is an orc, $P(\text{win}|\text{orc})=4/6$ and the probability of her losing given her opponent is an orc $P(\text{lose}|\text{orc})=2/6$. Then we choose the one with the highest probability and act accordingly, i.e. put 10p on Lara to win. This is a simple example of classification – choosing a particular class (win or lose in this example) given a set of parameters (character = orc).

Bayes' Rule

Later we'll want to solve classification problems with many parameters however calculating all the relevant conditional probabilities becomes extremely time consuming even with the help of powerful computers. To make matters worse we often do not have all the necessary information to do these calculations in the first place, so to overcome this problem we will use the eponymous Naïve Bayes method which is based on Bayes' Rule developed by the Reverend Thomas Bayes, an eighteenth century mathematician. This rule of probability allows us to express one conditional probability in terms of another and is derived from the multiplication rule that we used earlier.

Using the multiplication rule we can write:-

$$P(A \text{ and } B) = P(A).P(B|A) \text{ and} \\ P(B \text{ and } A) = P(B).P(A|B)$$

Since $P(A \text{ and } B)$ equals $P(B \text{ and } A)$ we can equate these to get:

$$P(B).P(A|B) = P(A).P(B|A).$$

We can re-arrange this to get:

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

A way of describing this rule is to say:

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

Prior refers to our "prior" belief of the value of $P(A)$ before we had additional information or evidence, $P(B)$ and the likelihood of that evidence occurring given A, $P(B|A)$.

Posterior $P(A|B)$ refers to the updated probability after taking into account this new evidence $P(B)$ and its likelihood given A, i.e. $P(B|A)$.

This is the simple form of the famous Bayes' Rule of probability which is very useful as it allows us to calculate the probability of A given B in terms of the probability of B given A. This might appear to be not very useful at all but in fact there are many situations where we might want to find the probability of A given B but we only have experimental data describing the probability of B given A. Common examples are taken from the medical literature, for

instance we may know the probability of having the symptoms of meningitis, $P(B)$, the probability of having meningitis, $P(A)$ and the probability of having the symptoms of meningitis given that the person has meningitis, $P(B|A)$ but we're interested in knowing the probability that someone has meningitis given that they have the symptoms $P(A|B)$. Notice also that the probability of having meningitis $P(A)$ is very small when compared to $P(B|A)$ so overall $P(A|B)$ is small indeed. It is easy to overlook this and it can lead to some serious overestimates - see [1].

Bayes' rule can be re-written by using the total law of probability described earlier to replace $P(B)$ in the denominator. Recall that the total law of probability says:-

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

So for any conditional probability that we are interested in say $P(A_j|B)$, we can also write Bayes rule as:

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

There are a few things to note here. First, the numerator also appears in the denominator as one of the terms to be summed. Secondly, the denominator reveals itself to be a normalizing constant in much the same way that the number 14 was used to normalize probabilities in our earlier example. This ensures that the probabilities sum to one as they always should. Third, if we are comparing conditional probabilities then we can forget about the denominator entirely – more on this later.

Classification and Bayes' Rule

Bayes' rule can be used to classify data and as already mentioned it is the basis of the powerful Naive Bayes classification method used in machine learning. Let's use the data of Lara's 14 fights to illustrate the approach. We will solve the same problem we posed earlier, whether or not to back Lara in a fight against an orc. Be aware that this example has only two variables and the problem is trivial – we've already derived the answer from table 1c. We will normally use this classification method with problems involving several variables, not just two. This example merely illustrates the method.

We want to know whether Lara is likely to win or lose given that we know her next opponent's character type. To do this we first define two classes, win and lose. Then we determine the probabilities of belonging to each class given the opponent's character type and choose the class with the highest probability (and if we feel lucky we'll put 10p on our choice).

In this example we would use Bayes' rule twice, once for each class:-

1. The probability of winning given the opponent's character type and
2. The probability of losing given the opponent's character type.

Calculation of the probability of winning given the opponent's character type, $P(A|B)$ requires us to first calculate the following probabilities:

- $P(A)$ is the probability of winning, whatever type of character she fought. We can work this out from the table.
- $P(B|A)$ is the probability that her opponent was a particular character type, given that she had won. We can work this out from the table too.
- $P(B)$ is just the probability that her opponent was a particular character type, win or lose – again we can work this out from the data in the table.

First let's calculate the probability of winning given the character is an orc.

$P(A|B)$ becomes $P(\text{win}|\text{orc})$ – we want to calculate this,

$P(A)$ is $P(\text{win})=9/14$,

$P(B|A)$ is $P(\text{orc}|\text{win})=4/9$

$P(B)$ is $P(\text{orc})=6/14$

Substitute these into Bayes equation:

$$P(\text{win}|\text{orc}) = \frac{\frac{9}{14} * \frac{4}{9}}{\frac{6}{14}}$$

Therefore

$$P(\text{win}|\text{orc}) = \frac{4}{6}$$

Which in this simple example we could have got straight away from table 1c, but please bear with me. Repeating this exercise to find $P(\text{lose}|\text{orc})$ we obtain $P(\text{lose}|\text{orc}) = \frac{2}{6}$. Try it yourself. We classify the fight as a win for Lara since $P(\text{win}|\text{orc})$ is greater than $P(\text{lose}|\text{orc})$.

This approach works if we have just one conditional probability but as we've already noted we usually have several variables that form part of the problem so we need to extend Bayes' Rule.

Bayes' rule with many features and Naïve Bayes'

Imagine Lara's fight data is more comprehensive and we have information regarding not just the type of character she fought but also the fitness, skill and strength of the character (see table 2 below). These additional variables make the problem more complex but we can simplify it by making a few assumptions. The resulting method we use is called Naïve Bayes and the word "naïve" appears because of the assumptions we make.

The additional variables need to be incorporated into the equation and we use the usual probability notation as follows,

$$P(A|B_1, B_2, \dots, B_n) = \frac{P(A)P(B_1, B_2, \dots, B_n|A)}{P(B_1, B_2, \dots, B_n)}$$

Here the left hand side reads "the probability of A given B_1 and B_2 and B_3 ." If we use the extended Lara fights example data below then $B_1, B_2, \dots, B_n$ would be character type, fitness, skill and strength. The variable A would be the result of the contest - win or lose.

To use this as a classifier we calculate $P(A|B_1, B_2, \dots, B_n)$ for differing outcomes, e.g. calculate probabilities when $A=\text{win}$ and when $A=\text{lose}$ and assign the result to the class with the highest probability.

This means we would need to have calculated these conditional probabilities for all possible combinations of variables. A moment's thought reveals that this is a lot of calculations as there are many possible combinations and this requires lots of observed data. We can simplify this equation by making the assumption (this is where the naïve bit comes in) that B_1, B_2 , etc. are all conditionally independent, i.e. the probability of one occurring is independent of the other whatever the given condition. In our example we assume that the probability of fighting an orc with high strength given we win is the same as the joint probability of fighting an orc given we win and fighting an opponent with high strength given we win. This isn't always true (there probably is a causal relationship between high strength and being an ogre) but nevertheless Naïve Bayes still works very well when this assumption is violated. So how do we modify Bayes rule to include the idea of conditional independence?

The multiplication rule says $P(A, B) = P(A).P(B|A)$.
Extending this to several variables we have

$$P(A, B_1, B_2, \dots, B_n) = P(A)P(B_1, B_2, \dots, B_n|A)$$

Note the right hand side is the numerator of Bayes rule.

Looking at the left hand side of the multiplication rule and applying the chain rule of probability (see Appendix A) we get

$$P(A, B_1, B_2, \dots, B_n) = P(B_n|A, B_1, \dots, B_{n-1})P(B_{n-1}|A, B_1, \dots, B_{n-2}) \dots P(B_2|A, B_1)P(B_1|A)P(A)$$

We now apply the assumption of conditional independence between $B_1, B_2, \dots, B_n$. For example, the first term in the rhs of the equation above is:

$$P(B_n|A, B_1, \dots, B_{n-1})$$

and since B_n does not depend upon $B_1, B_2, \dots$ or B_{n-1} then we can write

$$P(B_n|A, B_1, \dots, B_{n-1}) = P(B_n|A)$$

Repeating this logic for all terms we get

$$P(A, B_1, B_2, \dots, B_n) = P(B_1|A)P(B_2|A) \dots P(B_n|A) P(A)$$

Substituting this result back into Bayes rule finally gives us the Naïve Bayes equation:-

$$P(A|B_1, B_2, \dots, B_n) = \frac{P(A)P(B_1|A)P(B_2|A) \dots P(B_n|A)}{P(B_1, B_2, \dots, B_n)}$$

Note we could also replace the denominator term with the equivalent summation as we did earlier. The formula now looks like:

$$P(A|B_1, B_2, \dots, B_n) = \frac{P(A)P(B_1|A)P(B_2|A) \dots P(B_n|A)}{\sum_{i=1}^m P(B_1, B_2, \dots, B_n|A_i)P(A_i)}$$

Remember that ultimately we are not interested in calculating the denominator when comparing two or more classes as the denominator is the same in all cases.

Another important point needs to be made here. If any one of the numerator terms is zero then the result is zero and we can't use this method. Why would any result be zero? Probably because we have not got any example data in our set of samples, so we must ensure we have enough samples. If not we simply set any zero values to be a finite but small constant.

Example.

Armed with the extended Naive Bayes classifier let's revisit the problem. Lara's fight data is more comprehensive and we have information regarding the type of character she fought, the fitness, skill and strength of the character (see table 2 below). We want to know whether Lara will win or not against an orc who's fitness is high, speed is slow, skill is high and strength is strong.

Step1 – Training or learning based on previous knowledge.

As more clashes with opponents are recorded in the table, so the classifier can become more accurate and unusual (outlying) results become less important in determining the overall probability. This is a powerful feature of the Bayesian method. Lara's record is shown in table two and we use this to calculate the various probabilities required by the Naive Bayes classifier we have just developed.

Fight	Character Type	Opponent's Fitness	Opponent's Skill	Opponent's Strength	Win?
1	orc	high	high	weak	no
2	orc	high	high	strong	no
3	orc	<i>medium</i>	<i>high</i>	<i>weak</i>	<i>yes</i>
4	orc	<i>low</i>	<i>high</i>	<i>weak</i>	<i>yes</i>
5	human	<i>low</i>	<i>normal</i>	<i>weak</i>	<i>yes</i>
6	human	low	normal	strong	no
7	orc	<i>medium</i>	<i>normal</i>	<i>strong</i>	<i>yes</i>
8	ogre	high	high	weak	no
9	ogre	<i>high</i>	<i>normal</i>	<i>weak</i>	<i>yes</i>
10	human	<i>low</i>	<i>normal</i>	<i>weak</i>	<i>yes</i>
11	ogre	<i>high</i>	<i>normal</i>	<i>strong</i>	<i>yes</i>
12	human	<i>medium</i>	<i>high</i>	<i>strong</i>	<i>yes</i>
13	orc	<i>medium</i>	<i>normal</i>	<i>weak</i>	<i>yes</i>
14	human	low	high	strong	no

So here we have 4 attributes. What we need to do is to create “look-up” tables for each of these attributes, and write in the probabilities of winning and losing based on this attribute. In these tables we have to note that there are 5 cases of losing, and 9 cases of winning a contest.

Character Type	Win = yes	Win = no
human	3/9	2/5
orc	4/9	2/5
ogre	2/9	1/5

Fitness	Win = yes	Win = no
high	2/9	3/5
medium	4/9	0/5
low	3/9	2/5

Opponent's Skill	Win = yes	Win = no
high	3/9	4/5
normal	6/9	1/5

Opponent's Strength	Win = yes	Win = no
strong	3/9	3/5
weak	6/9	2/5

We must also note that the $P(\text{Win=yes}) = 9/14$ and $P(\text{Win=no})=5/14$

Step 2: Calculating which is the most probable class, or testing

For this, we need to know the attributes of the opponent as stated earlier. These are

X = (Fitness=high, Character Type = orc, Skill=high, Strength=strong)

Firstly we look at the probabilities of the attributes associated with the fact that she won, so we use the lookup tables to get:

- $P(\text{Fitness} = \text{high} \mid \text{Win} = \text{yes}) = 2/9$
- $P(\text{Character Type} = \text{orc} \mid \text{Win} = \text{yes}) = 4/9$
- $P(\text{Skill} = \text{high} \mid \text{Win} = \text{yes}) = 3/9$
- $P(\text{Strength} = \text{strong} \mid \text{Win} = \text{yes}) = 3/9$
- $P(\text{Win} = \text{yes}) = 9/14$

Now we calculate the probability of winning given the attributes by substituting the probabilities into the numerator of the naive Bayes equation. This is just the product of all those probabilities, i.e. we multiply the whole lot together. So the product of the probabilities for Win=yes is:

$$(2/9) * (4/9) * (3/9) * (3/9) * (9/14) = 0.0071$$

We ignore the denominator (can you remember why?)

Next we list the probabilities of the attributes occurring given we did not win.

If you look up the Win=No probabilities, you should get:

- $P(\text{Fitness} = \text{high} \mid \text{Win} = \text{no}) = 3/5$
- $P(\text{Character Type} = \text{orc} \mid \text{Win} = \text{no}) = 2/5$
- $P(\text{Skill} = \text{high} \mid \text{Win} = \text{no}) = 4/5$
- $P(\text{Strength} = \text{strong} \mid \text{Win} = \text{no}) = 3/5$
- $P(\text{Win} = \text{no}) = 5/14$

Now we calculate the probability of NOT winning given the attributes. We repeat the calculation as above using the Win=no probabilities.

We get an answer of 0.0411

Again we ignore the denominator.

Now we compare the “win given the attributes” probability with the “lose given the attributes” probability. Since the denominator of the equation is the same and we are doing a comparison we can ignore this as the result of the comparison will remain unchanged. Recall from your school days that if $A/B < C/B$ then $A < C$. That is why we don't need to calculate the denominator of the Naive Bayes equation.

So, given the probabilities, what are the chances of winning? Well:

$$P(\text{Yes} \mid X) < P(\text{No} \mid X) \quad \Rightarrow \quad 0.0071 < 0.0411$$

So the answer is Lara probably won't win.

Appendix A: The chain Rule of Probability

Let $\{A_1, A_2, \dots, A_n\}$ be a collection of events. The chain rule of probability gives us a useful way to compute the joint probability of the entire collection

$$P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2) \cdots P(A_n | A_1, A_2, \dots, A_{n-1})$$

Proof: Simply expand the conditional probabilities and note how the denominator of the term $P(A_k | A_1, \dots, A_{k-1})$ cancels the numerator of the previous conditional probability, i.e.,

$$P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2) \cdots = P(A_1) \frac{P(A_1, A_2)}{P(A_1)} \frac{P(A_1, A_2, A_3)}{P(A_1, A_2)} \cdots \frac{P(A_1, \dots, A_n)}{P(A_1, \dots, A_{n-1})}$$

Exercises:-

Using the data from table two, answer the question “should I bet on Lara to win given she fights an orc with low fitness, high skill and is strong.”