Naive Bayes

Day	Outlook	Temperature	Humidity	Wind	$> 1,000$?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Strong	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

After graduating from MUIC you decided to become a circus master. You ran a show every single day. You notice that the weather seems to have something to do with the number of audience. So, you want to

You want to be able to predict whether that **given** today's weather would the attendance be more than a thousand. If so you can then, You remember something from Discrete Math called probabilities.

First Attempt: Table Lookup

Suppose to day is Rainy Mild High Humidity and Weak Day. It's quite easy to estimate the amount of attendance since you have this kind of day before. The Rainy-Mild-High-Weak day happened before on day 4. So to our best knowledge, assuming that the history will repeat itself, we would say that the attendnace will be more than 1,000 on this day.

If we collect more and more data, you can then calculate the probability from historical data and do a table look up for the probability. For example, if historically out of 100 Rainy Mild High Humidity and Weak day, 70 of those day you have more than 1,000 audiences. Then you would say that the probability is 70%. This is the most intuitive and the best one if you have more than "enough" data.

Problem Statement

Let us calculate the number of data needed to even cover all the possibilities. We have 3 outlook, 3 temperature, 2 humidity level and 2 wind level. To cover 100 data point each, you will need $3 \times 3 \times 2 \times 2 \times 100 = 3{,}600$days. That's a lot of data to cover all the possibilities.

Let us consider the day that we have not seen exactly before: a **Sunny-Hot-Normal-Weak** day. Brushing off the problem saying that we need 10 more years of data to calculate such a thing would be a bit too premature.

Even though we have not seen that day exactly before, we have seen some that looks similar before. We have seen sunny days. We have seen hot days. We have seen normal humidity days. We have seen weak wind days. It is just that we have not seen them all together.

It is our hope that by combining the individual information from Sunny days, Hot days, Normal humidity day and weak wind day together. We would be able to calculate the probability that the audicence will be greater than 1,000. In other words, we want to "learn" from the past data and predict the unknown.

The attributes of the data like Sunny Hot Normal Weak are called **features** and the our target result whether the audience is greater than 1,000 or not is called **classes**. That is we always want to predict the classes from given features.

It's clear that we need something better than a look up table. Let us use something from Discrete Math class: the conditional probabilities. What we want to calculate is the probability that the there will be more than 1,000 audiences given that today is a Sunny, Hot Normal, Weak day which is denoted by

$$P(> 1000|\text{Sunny}, \text{Hot}, \text{Normal}, \text{Weak}) \tag{1}$$

Bayes Rule

Recall from Statistics and Discrete Mathematics class that

$$P(A|B) = \frac{P(A \cap B)}{\Pr(B)} \tag{2}$$

This also means that

$$P(B|A) = \frac{P(A \cap B)}{\Pr(A)} \tag{3}$$

Combining the two equations above gives us a very useful relation.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{4}$$

This relation is called *Bayes's Theorem*. You will see why this is important later in the next section.

Chain Rule

Recall the problem that we are trying to solve has more than 1 variable:

$$P(> 1000|\text{Sunny} \cap \text{Hot} \cap \text{Normal} \cap \text{Weak}) \tag{5}$$

To make the notaion more succint we will call this

$$P(T|S \cap H \cap N \cap W) \tag{6}$$

From the definition we know that

$$P(T|S \cap H \cap N \cap W) = \frac{P(S \cap H \cap N \cap W \cap T)}{P(S \cap H \cap N \cap W)} \tag{7}$$

The term $P(S \cap H \cap N \cap W \cap T)$ is impossible to estimate from data but it can be simplify greatly using Equation 2.

$$P(S \cap H \cap N \cap W \cap T) = P(S|H, N, W, T) \times P(H, N, W, T) \tag{8}$$

We can continue doing this

$$\begin{aligned} P(S \cap H \cap N \cap W \cap T) &= P(S|H, N, W, T) \times P(H, N, W, T) && (9)\\ &= P(S|H, N, W, T) \times P(H|N, W, T) \times P(N, W, T) && (10)\\ &= P(S|H, N, W, T) \times P(H|N, W, T) \times P(N|W, T) \times P(W, T) && (11)\\ &= P(S|H, N, W, T) \times P(H|N, W, T) \times P(N|W, T) \times P(W|T) \times P(T) && (12) \end{aligned}$$

Combining every thing We got

$$P(T|S \cap H \cap N \cap W) = P(S|H, N, W, T) \times P(H|N, W, T) \times P(N|W, T) \times P(W|T) \frac{\times P(T)}{P(S \cap H \cap N \cap W)} \tag{13}$$

This doesn't really do us much good. Since the terms like $P(S|H, N, W, T)$ is actually quite hard to calculate we will need to resort to lookup table which requires a lot of data.

But the terms like $P(W|T)$ is actually pretty easy to calculate since it just ask you to count that for all the days that we have more than 1,000 audiences, how many days are weak wind day. In particular, from the table above. There are 9 days with more than 1,000 audiences. Out of those 9 days, there are 5 days with weak wind. Thus,

$$P(W|T) = \frac{5}{9} \tag{14}$$

Our hope is to make those hard-to-calculate term easier to calculate. This is where we use the assumption that all the feature are *conditional independent* of each other. In particular, let us consider closely ther term like

$$P(S|H, N, W, T)$$

The humidity, wind speed and temperature doesn't really have much to do with how Sunny the day is. The only thing that is not independence is probably the audience level. This is called the *naive* assumption (thus the name of this note: Naive Bayes Classifier.). This means

$$P(S|H, N, W, T) = P(S|T) \tag{15}$$

This assumption is not true in general. More often than not we will have to deal with correlated features. Yet this greatly simplify Equation 13 to

$$P(T|S \cap H \cap N \cap W) = P(S|T) \times P(H|T) \times P(N|T) \times P(W|T) \frac{\times P(T)}{P(S \cap H \cap N \cap W)} \tag{16}$$

Each of the $P(S|T)$,$P(H|T)$,$P(N|T)$ and $P(W|T)$ terms are easy to calculate from data. All we need to is to count how many of day that has more than 1,000 audiences, is a sunny day. The numbers are given below

$$\begin{aligned} P(S|T) &= 2/9\\ P(H|T) &= 2/9\\ P(N|T) &= 6/9\\ P(W|T) &= 5/9 \end{aligned}$$

Prior

Now that we have 4 terms done. There is still $P(T)$ the probability that we will get more than 1,000 audience. This is called the *Prior* There are two ways to do this each will most likely give you different answers.

One way to do it is to say, given that we don't know anything let us use *equiprobable* assumption that there is a 50% chance that T will happend and another 50% chance that T will not happen. If you really have no idea what you are expected to see then this is a good start.

$$P(T) = 1/2$$

Another way is to assume that the data we collect will represent what we are going to see in the future. Thus, $P(T)$, the probability that there will be more than 1,000 audience can be calculated from the data by just counting the number of days with more than 1,000 audience. In the table given at the beginning, there are 9 out of 14 days that we have more than 1,000 people. Therefore we would say that

$$P(T) = 9/14$$

Which one to use is really a philosophical question of whether how much you can trust the distribution in your data: whether it is a good representation of the frequency you are going to see. For example, if you want to make conclusion about entire Thai population but we only sample college student. We definitely can't hope that college students is a good representation what we are going to see in the entire population. But, if we do a good sampling of Thai population demographics then we can justify that our collection represents the population.

For this example, let assume that the data we have is a good representation of what we are going to see.

Evidence

$P(S \cap H \cap N \cap W)$ is called the evidence term. With independence assumption, it is actually

$$P(S \cap H \cap N \cap W) = P(S)P(H)P(N)P(W)$$

Of course, you can estimate this quantity by counting the number of sunny day from all the sample and so on. But in reality, you don't really need to. This again relies on the fact that the data we collected will be repeated. In this case is it kind of bad assumption. We will get around this in the next section.

Naive Bayes Classifier

The question we want to answer is whether we will have more than 1,000 audience or not. To decide that, we just need to compare two probabilities.

$$P(T|S, H, N, W) \text{ VS } P(\neg T|S, H, N, W)$$

whichever one is greater would be our prediction.

We learn from the discussion above that

$$P(T|S, H, N, W) = P(S|T)P(H|T)P(N|T)P(W|T)\frac{P(T)}{P(S, H, N, W)} \tag{17}$$

and similarly

$$P(\neg T|S,H,N,W) = P(S|\neg T)P(H|\neg T)P(N|\neg T)P(W|\neg T)\frac{P(\neg T)}{P(S,H,N,W)} \tag{18}$$

Furthermore, since we know that the audience is either $> 1{,}000$ or $\leq 1{,}000$. We have

$$\begin{aligned} 1 =& P(T|S,H,N,W) + P(\neg T|S,H,N,W) && (19)\\ 1 =& P(S|T)P(H|T)P(N|T)P(W|T)\frac{P(T)}{P(S,H,N,W)} && (20)\\ &+ P(S|\neg T)P(H|\neg T)P(N|\neg T)P(W|\neg T)\frac{P(\neg T)}{P(S,H,N,W)} && (21) \end{aligned}$$

This also means

$$\begin{aligned} P(S,H,N,W) =& P(S|T)P(H|T)P(N|T)P(W|T)P(T) && (22)\\ &+ P(S|\neg T)P(H|\neg T)P(N|\neg T)P(W|\neg T)P(\neg T) && (23) \end{aligned}$$

Example

Since there are so many pieces needed to calculate

$$P(T|S,H,N,W) = P(S|T)P(H|T)P(N|T)P(W|T)\frac{P(T)}{P(S,H,N,W)}$$

Let us go through the process.

First Let us calculate

$$P(S|T)P(H|T)P(N|T)P(W|T) = \frac{2}{9}\times\frac{2}{9}\times\frac{6}{9}\times\frac{5}{9} = 0.0183$$

Then,

$$P(S|\neg T)P(H|\neg T)P(N|\neg T)P(W|\neg T) = \frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{2}{5} = 0.0192$$

Next, $P(T) = 9/14$ since out of 14 days, there are 9 days which we have more than 1,000 audiences. Thus $P(\neg T) = 5/14$ since the other 5 days we have less than 1,000 audiences.

Therefore, the evidence term/normalization factor is

$$P(S,H,N,W) = 0.183\times\frac{9}{14} + 0.192\times\frac{5}{14} = 0.186$$

Thus,

$$P(T|S,H,N,W) = \frac{0.183\times\frac{9}{14}}{0.186} = 0.632$$

and(you can do just one minus the above answer too but we will be a pedagogical here)

$$P(\neg T|S,H,N,W) = \frac{0.192\times\frac{5}{14}}{0.186} = 0.368$$

So, with Naive Bayes Classifier we would say that the probability that we will have more than 1,000 audience is 0.632.

Remember that we made a very critical *assumption* here that all the features are *independent*. This is not true in general. However, even when this assumption is not satisfied the Naives Bayes classifier usually do a pretty good job.

As you can see, the calculation is quite tedious. It sounds more like a job for computer to do.

Multinomial Bayes

The whole idea of classifier is to estimate the probability. This depends on how we model the problem. In the last section we model the problem as having a bunch of conditionally independent features each contribute independent to the result. Let us consider the problem of classifying spam email.

The simplest way[1] to do this is to count the frequnecy of word that appear in the document. This would be our features. For example, a spam email reads.

Cheap Windows legit cheap act now.

this can be represent using *feature vector*

$$\vec{x} = \{\text{cheap: 2, windows: 1, act: 1, now: 1}\}$$

Thus, our job of classifying spam email becomes comparing

$$P(\text{Spam}|\vec{x}) \text{ VS } P(\text{Ham}|\vec{x})$$

which means given all the frequencies we found in the given email, which probability is greater: spam or ham?

The problem is now a bit different from before since we started to have frequencies coming into play. You could instead of using frequency you could use a boolean whether a word appear or not and we will be back with Naive Bayes scenario. Let us be fancy here and actually exploit frequency instead of just boolean.

Remember that things like the probability that a word "cheap" will appear in a spam email

$$P(\text{cheap appear}|\text{spam})$$

is very easy to deduce from data. We just count the number of "cheap" from all spam email and divide by the total number of words that appear in all spam email[2].

Recall chain rule

$$\begin{aligned} P(\text{Spam}|\vec{x}) \propto & P(\text{cheap}|\text{spam})^2 \times P(\text{windows}|\text{spam}) \times \\ & P(\text{act}|\text{spam}) \times P(\text{now}|\text{spam}) \times P(\text{spam}) \end{aligned} \tag{24}$$

Note that we use the symbol $\propto$ which means proportional to since we ignore two things. The combinatoric factor and the evidence term[3]. But we don't really need to care about that since the combinatoric factor is the same for both spam and ham case. The square on $P(\text{cheap}|\text{spam})$ term indicate that the word cheap appear twice.

Similarly,

$$\begin{aligned} P(\text{Spam}|\vec{x}) \propto & P(\text{cheap}|\text{spam})^2 \times P(\text{windows}|\text{spam}) \times \\ & P(\text{act}|\text{spam}) \times P(\text{now}|\text{spam}) \times P(\text{spam}) \end{aligned} \tag{25}$$

If we really need the probability we can use the fact that

$$P(\text{Spam}|\vec{x}) + P(\text{Ham}|\vec{x}) = 1$$

and then use the sum of the two number to normalize the probability.

[1] try google for more sophisticated ways.

[2] Normally clean up the data and get rid of article and conjunction words like a, the, an, of, in, on, at, etc. and lower case all the words.

[3] Look up wikipedia if you wonder about them

Unseen Words

Let us see what would happen if we see the word we never seen before. Remember that we estimate

$$P(\text{word}|\text{spam})$$

by counting the number words that appear in the spam. But if we want to decide an email like

Let meet *tommorriw* morning

Note that the tommorrow is misspelled. With the method of counting words in the training data set we will find that

$$P(\text{tomorriw}|\text{spam}) = P(\text{tomorriw}|\text{spam}) = 0$$

which will make both $P(\text{spam}|\vec{x})$ and $P(\text{ham}|\vec{x})$ zero!! and you have no idea whether it is a spam or ham.

That's a really bad situation. Let us think back a little bit, the more sensible way to treat unseen words is to ignore it and just use the rest of the words. That means all we need to do is assign some small non-zero number e.g. 1/totalnumber of words to both $P(\text{unseen}|\text{spam})$ and $P(\text{unseen}|\text{ham})$. Since multiplying by the same non-zero number doesn't change the ranking of the two we have the desirable effect.[4]

In data analysis job, we will often see some details like this very often. There is no way to get around it except to actually understand what we are doing. If you use the algorithm like magic with no real understanding. More often than not you will get unexpected answer. The purpose of this class is making sure you don't use it as magical formula and give you enough basic to understand more magic.

[4]Some people add the number of word found by 1 and normalize by the number of vocabulary. This gives a similar effect. But behave better in edge cases. We won't get there.