Lesson 16

Classification Techniques – Naïve Bayes

Kush Kulshrestha

Bayes Theorem

Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age.

Bayes' theorem is stated mathematically as the following equation:

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

where A and B are events and P(B) <> 0

- P (A|B) is a conditional probability: the likelihood of event A occurring given that B is true.
- P (B|A) is also a conditional probability: the likelihood of event B occurring given that A is true.
- P(A) and P(B) are the probabilities of observing A and B independently of each other. P(B) is also called marginal probability of B and P(A) is called prior.

Bayes Theorem

Why is Prior important?

Image you are standing in a line of people and the person standing in front of you is having long hairs. What can you conclude about the gender of the person?

Most people will suggest that the person with long hair is probably woman.

Now if some more information is given to you (prior) that the line you are standing in is the line towards Gents restroom. Now what would be your estimate about the gender?

You can say with very high certainty that the person is a male.

This is how including prior into your analysis could help you in predicting the best fit category / class.

Intro to Naïve Bayes

Candy Selection Example

Once there lived an old Grandma. Every year on her birthday, her entire family would visit her and stay at her mansion. Sons, daughters, their spouses, her grandchildren. It would be a big bash every year, with a lot of fanfare. But what Grandma loved the most was meeting her grandchildren and getting to play with them. She had ten grandchildren in total, all of them around 10 years of age, and she would lovingly call them "**random variables**".

Every year, Grandma would present a candy to each of the kids. Grandma had a large box full of candies of ten different kinds. She would give a single candy to each one of the kids, since she didn't want to spoil their teeth.

But, as she loved the kids so much, she took great efforts to decide which candy to present to which kid, such that it would maximize their total happiness. (the **maximum likelihood estimate**, as she would call it)

Intro to Naïve Bayes

But that was not an easy task for Grandma. She knew that each type of candy had a certain **probability** of making a kid happy. That probability was different for different candy types, and for different kids. Rakesh liked the red candy more than the green one, while Sheila liked the orange one above all else. Each of the 10 kids had different preferences for each of the 10 candies. Moreover, their preferences largely depended on external factors which were unknown (**hidden variables**) to Grandma. If Sameer had seen a blue building on the way to the mansion, he'd want a blue candy, while Sandeep always wanted the candy that matched the colour of his shirt that day. But the biggest challenge was that their happiness depended on what candies the other kids got! If Rohan got a red candy, then Niyati would want a red candy as well, and anything else would make her go crying into her mother's arms (**conditional dependency**). Sakshi always wanted what the majority of kids got (**positive correlation**), while Tanmay would be happiest if nobody else got the kind of candy that he received (**negative correlation**). Grandma had concluded long ago that her grandkids were completely **mutually dependent**

It was computationally a big task for Grandma to get the candy selection right. There were too many conditions to consider and she could not simplify the calculation. Every year before her birthday, she would spend days figuring out the optimal assignment of candies, by enumerating all configurations of candies for all the kids together.

She was getting old, and the task was getting harder and harder. She used to feel that she would die before figuring out the optimal selection of candies that would make her kids the happiest all at once.

Intro to Naïve Bayes

But an interesting thing happened. As the years passed and the kids grew up, they finally passed from teenage and turned into **independent** adults. Their choices became less and less dependent on each other, and it became easier to figure out what is each one's most preferred candy.

Grandma was quick to realize this, and she joyfully began calling them "**independent random variables**".

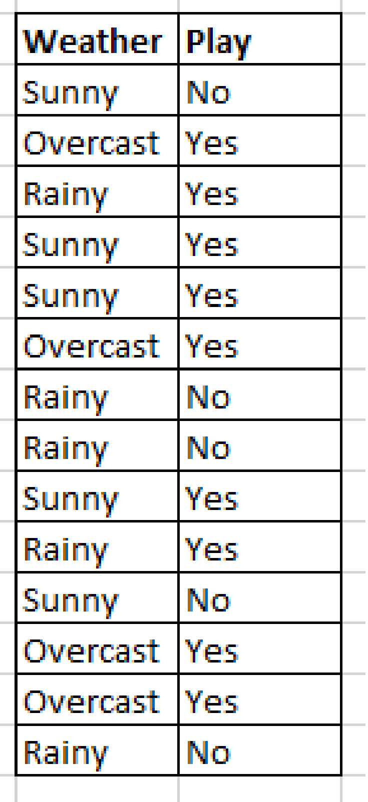
It was much easier for her to figure out the optimal selection of candies - she just had to think of one kid at a time and, for each kid, assign a happiness probability to each of the 10 candy types for that kid. Then she would pick the candy with the highest happiness probability for that kid, without worrying about what she would assign to the other kids. This was a super easy task, and Grandma was finally able to get it right.

Learnings:

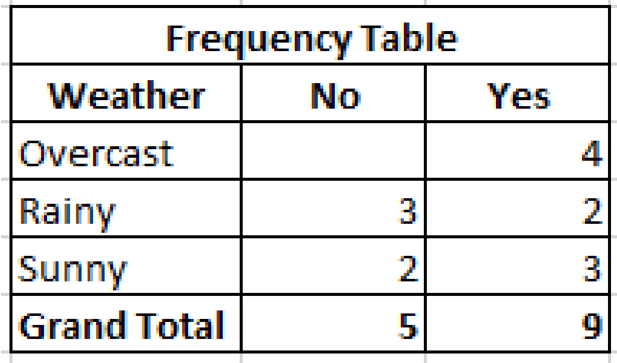
- In statistical modelling, having mutually dependent random variables makes it really hard to find out the optimal assignment of values for each variable that maximizes the cumulative probability of the set.
- However, if the variables are independent, it is easy to pick out the individual assignments that maximize the probability of each variable, and then combine the individual assignments to get a configuration for the entire set.
- In Naive Bayes, you make the assumption that the variables are independent (even if they are actually not). That’s why it is called Naïve.
- This simplifies your calculation, and in many cases, it actually gives estimates that are comparable to those which you would have obtained from a more (computationally) expensive model that takes into account the conditional dependencies between variables.

How Naïve Bayes works?

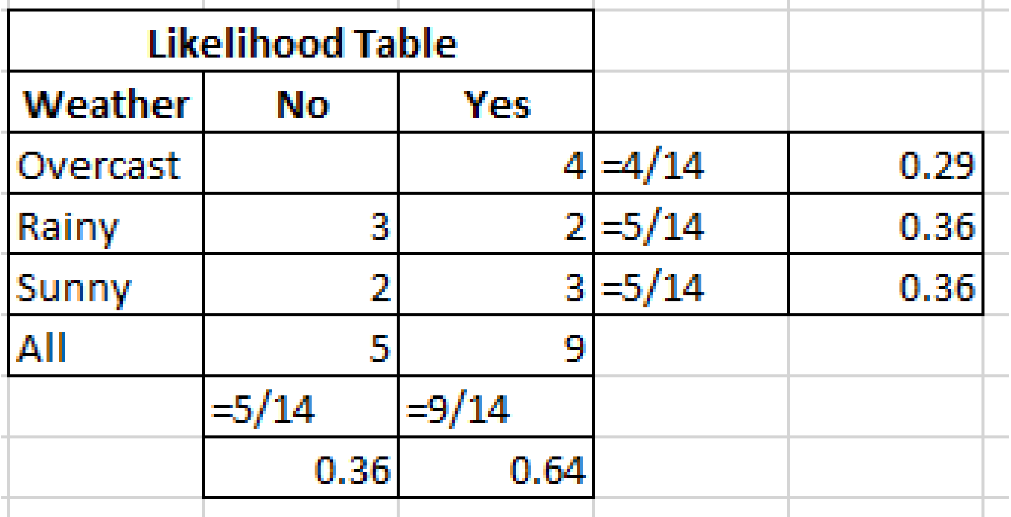
Below I have a training data set of weather and corresponding target variable ‘Play’ (suggesting possibilities of playing).



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No



Frequency Table		
Weather	**No**	**Yes**
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	**5**	**9**



Likelihood Table				
Weather	**No**	**Yes**		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

How Naïve Bayes works?

Players will play if weather is sunny. Is this statement is correct?

We can solve it using above (Bayes Theorem) discussed method of posterior probability.

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

We know from the data: P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

Pros and Cons of Naïve Bayes

Pros:

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction.
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It performs reasonably good even when the assumption of independence is somewhat not true.

Cons:

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.