MACHINE LEARNING (ML-11)

Dr. NEERAJ GUPTA, Department of CEA, GLA University, Mathura

AGENDA

- Naïve Bayes Classifier

WHAT IS NAIVE BAYES ALGORITHM?

- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes model is easy to build and particularly useful for very large data sets.
- Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Prob.

PREREQUISITES FOR BAYES' THEOREM

What is an Experiment?

"An experiment is a planned operation carried out under controlled conditions."

Tossing a coin, rolling a die, and drawing a card out of a well-shuffled pack of cards are all examples of experiments.

SAMPLE SPACE

The result of an experiment is called an outcome. The set of all possible outcomes of an event is called the sample space.

For example, if our experiment is throwing dice and recording its outcome, the sample space will be:

$$S1 = \{1, 2, 3, 4, 5, 6\}$$

What will be the sample when we're tossing a coin?

$$S2 = \{H, T\}$$

EVENT

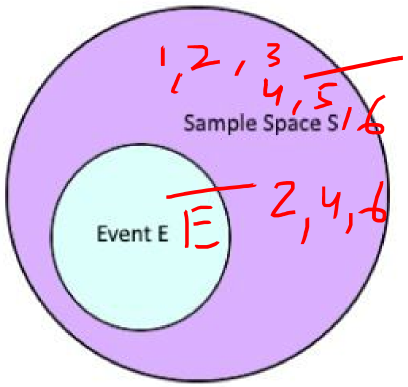
An event is a set of outcomes (i.e. a subset of the sample space) of an experiment.

Let's get back to the experiment of rolling a dice and define events E and F as:

E = An even number is obtained = {2, 4, 6}

F = A number greater than 3 is obtained = {4, 5, 6}

The probability of these events:



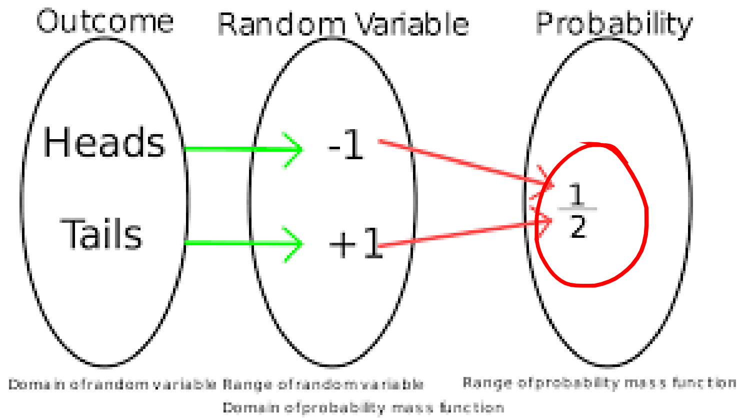
P(E) = Number of favorable outcomes / Total number of possible outcomes = 3 / 6 = 0.5

P(F) = 3 / 6 = 0.5

RANDOM VARIABLE

A Random Variable is exactly what it sounds like – a variable taking on random values with each value having some probability (which can be zero).

It is a real-valued function defined on the sample space of an experiment:



RANDOM VARIABLE

Let's take a simple example (refer to the above image as we go along). Define a random variable X on the sample space of the experiment of tossing a coin. It takes a value +1 if "Heads" is obtained and -1 if "Tails" is obtained. Then, X takes on values +1 and -1 with equal probability of 1/2.

Consider that Y is the observed temperature (in Celsius) of a given place on a given day. So, we can say that Y is a continuous random variable defined on the same space, S = [0, 100] (Celsius Scale is defined from zero degree Celsius to 100 degrees Celsius).

EXHAUSTIVE EVENTS

A set of events is said to be exhaustive if at least one of the events must occur at any time. Thus, two events A and B are said to be exhaustive if A U B = S, the sample space.

For example, let's say that A is the event that a card drawn out of a pack is red and B is the event that the card drawn is black. Here, A and B are exhaustive because the sample space S = {red, black}. Pretty straightforward stuff, right?

INDEPENDENT EVENTS

If the occurrence of one event does not have any effect on the occurrence of another, then the two events are said to be independent. Mathematically, two events A and B are said to be independent if:

independent

$$\boxed{P(A \cap B) = P(AB) = P(A)*P(B)}$$

dice

cards

For example, if A is obtaining a 5 on throwing a die and B is drawing a king of hearts from a well-shuffled pack of cards, then A and B are independent just by their definition. It's usually not as easy to identify independent events, hence we use the formula I mentioned above.

CONDITIONAL PROBABILITY

Consider that we're drawing a card from a given deck.

What is the probability that it is a black card?

That's easy – 1/2, right?

However, what if we know it was a black card – then what would be the probability that it was a king?

This is where the concept of conditional probability comes into play.

Conditional probability is defined as the probability of an event A, given that another event B has already occurred (i.e. A conditional B). This is represented by P(A|B) and we can define it as:

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

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

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

CONDITIONAL PROBABILITY

Let event A represent picking a king, and event B, picking a black card. Then, we find P(A|B) using the above formula:

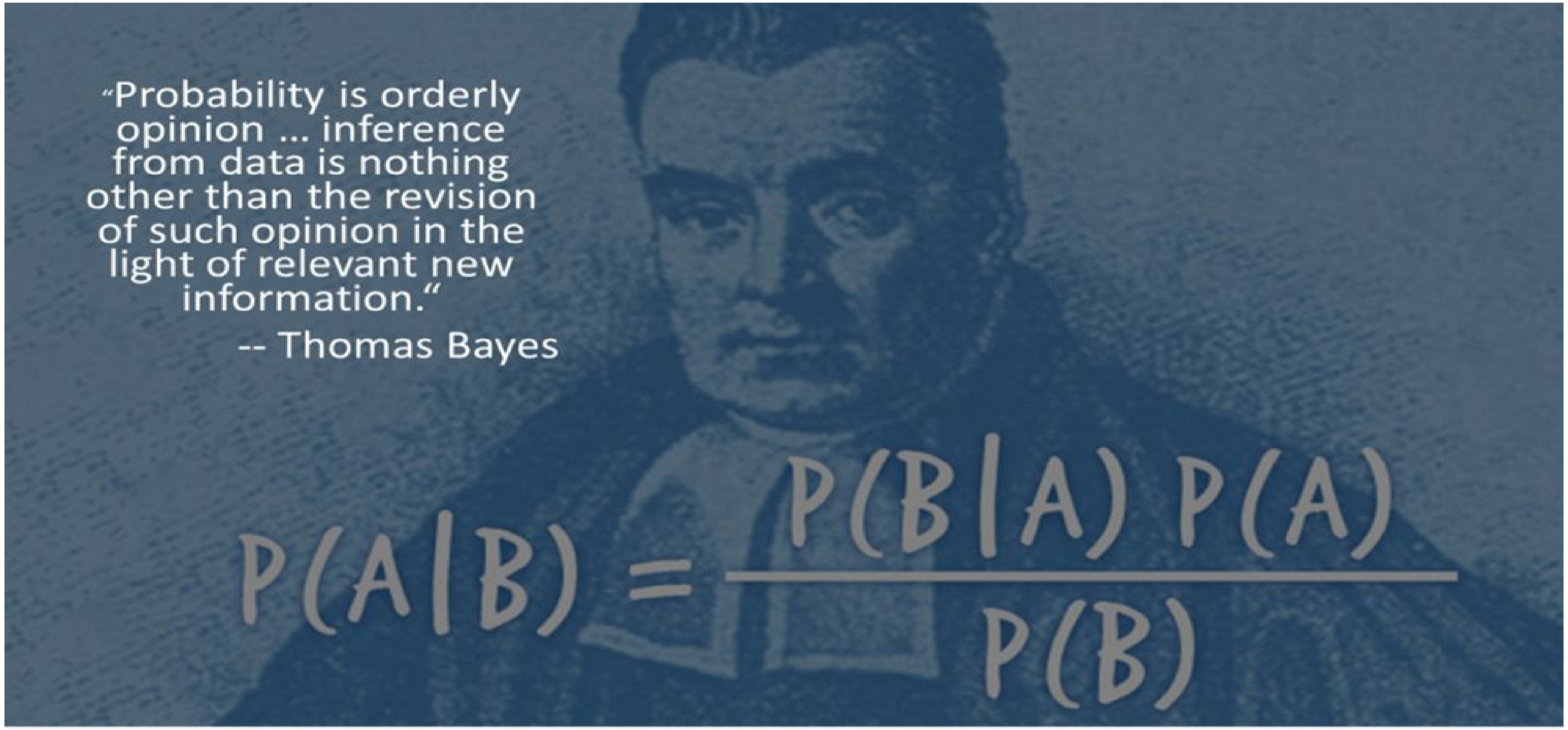
King Black

P(A ∩ B) = P(Obtaining a black card which is a King) = 2/52

P(B) = P(Picking a black card) = 1/2

Thus, P(A|B) = 4/52. Try this out on an example of your choice.

WHAT IS BAYES' THEOREM?



WHAT IS BAYES' THEOREM?

"Have you ever seen the popular TV show 'Sherlock' (or any crime thriller show)? Think about it – our beliefs about the culprit change throughout the episode. We process new evidence and refine our hypothesis at each step.

This is Bayes' Theorem in real life!"

BAYES' THEOREM

Now, let's understand this mathematically. Consider that A and B are any two events from a sample space S where P(B) ≠ 0. Using our understanding of conditional probability, we have:

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

Similarly, $P(B|A) = P(A \cap B) / P(A)$

It follows that $\boxed{P(A \cap B)} = P(A|B) * P(B) = P(B|A) * P(A)$

Thus, $P(A|B) = P(B|A)*P(A) / P(B)$

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

$$\boxed{\begin{aligned} P(A \cap B) &= P(A|B) * P(B) \\ P(A \cap B) &= P(B|A) * P(A) \end{aligned}}$$

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

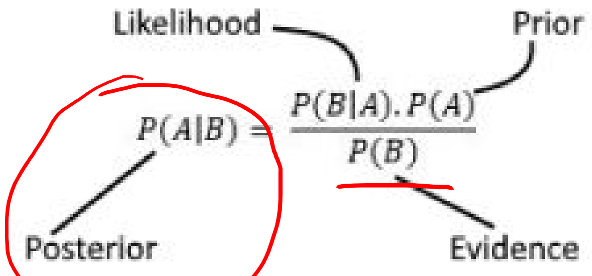
$$\boxed{P(A|B) = \frac{P(B|A) * P(A)}{P(B)}}$$

Here, P(A) and P(B) are probabilities of observing A and B independently of each other. P(B|A) and P(A|B) are conditional probabilities.

BAYES' THEOREM

P(A) is called **Prior probability** and P(B) is called **Evidence.**

P(B|A) is called **Likelihood** and P(A|B) is called **Posterior probability.**



posterior = likelihood * prior / evidence

AN ILLUSTRATION OF BAYES' THEOREM

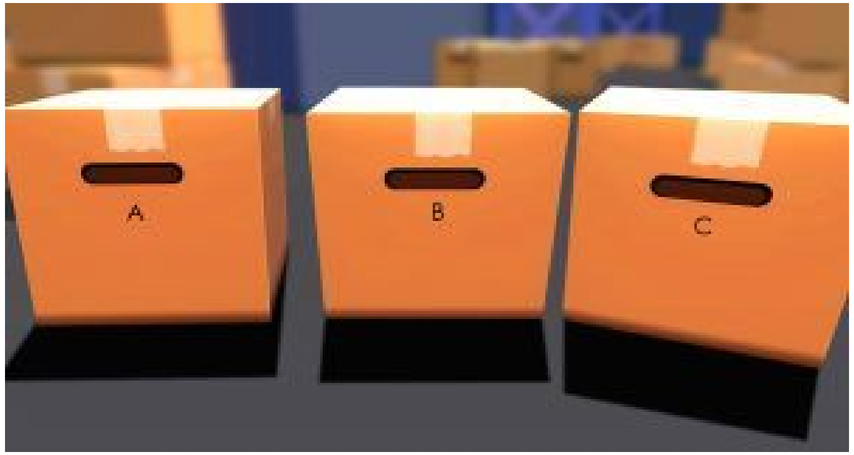
Let's solve a problem using Bayes' Theorem. This will help you understand and visualize where you can apply it.

There are 3 boxes labeled A, B, and C:

Box A contains 2 red and 3 black balls

Box B contains 3 red and 1 black ball

And box C contains 1 red ball and 4 black balls



The three boxes are identical and have an equal probability of getting picked. Consider that a red ball is chosen. Then what is the probability that this red ball was picked out of box A?

CONTD...

We have prior probabilities P(A) = P(B) = P (C) = 1 / 3, since all boxes have equal probability of getting picked.

P(E|A) = Number of red balls in box A / Total number of balls in box A = 2 / 5

Similarly, P(E|B) = 3 / 4 and P(E|C) = 1 / 5

Then evidence P(E) = P(E|A)*P(A) + P(E|B)*P(B) + P(E|C)*P(C)

= (2/5) * (1/3) + (3/4) * (1/3) + (1/5) * (1/3) = 0.45

Therefore, P(A|E) = P(E|A) * P(A) / P(E) = (2/5) * (1/3) / 0.45 = 0.296

APPLICATIONS OF BAYES' THEOREM

The three main applications of Bayes' Theorem:

Naive Bayes' Classifiers

Discriminant Functions and Decision Surfaces

Bayesian Parameter Estimation

NAIVE BAYES' CLASSIFIERS

Main Point

Naive Bayes' Classifiers are a set of probabilistic classifiers based on the Bayes' Theorem. **The underlying assumption of these classifiers is that all the features used for classification are independent of each other.**

That's where the name 'naive' comes in since it is rare that we obtain a set of totally independent features.

The way these classifiers work is exactly how we solved in the illustration, just with a lot more features assumed to be independent of each other.

NAIVE BAYES' CLASSIFIERS

Here, we need to find the probability P(Y|X) where X is an n-dimensional random variable whose component random variables X_1, X_2, …., X_n are independent of each other:

Sunny

$$X = [x_1, x_2, \dots x_n]$$

$$P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)}$$

$$P(X) = P(X_1, X_2, \dots \dots, X_n) = P(X_1).P(X_2) \dots \dots P(X_n)$$

because for independent events A and B, $P(AB) = P(A).P(B)$

Similarly because of conditional independence,

$$P(X|Y) = P(X_1, X_2, \dots \dots, X_n|Y) = P(X_1|Y).P(X_2|Y) \dots \dots P(X_n|Y)$$

Yes or No

Substituting these in (1), *we get*

$$P(Y|X) = \frac{P(X_1|Y).P(X_2|Y) \dots \dots P(X_n|Y).P(Y)}{P(X_1).P(X_2) \dots \dots P(X_n)}$$

$$P(Y|X) \propto P(X_1|Y).P(X_2|Y) \dots \dots P(X_n|Y).P(Y)$$

$X \rightarrow$ feature

$Y^{(1)} \rightarrow$? target

$X = \{x_1, x_2, \dots x_n\}$

$P(x)$

$P(x_1, x_2, \dots x_n)$

$P(x_1) \cdot P(x_2) \cdot P(x_n)$

Finally, the Y for which P(Y|X) is maximum is our predicted class.

WORKING OF NAÏVE BAYES' CLASSIFIER

Suppose we have a dataset of **weather conditions** and corresponding target variable **"Play"**. So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

Convert the given dataset into frequency tables.

Generate Likelihood table by finding the probabilities of given features.

Now, use Bayes theorem to calculate the posterior probability.

Problem: If the weather is sunny, then the Player should play or not?

CONTD...

Solution: To solve this, first consider the below dataset:

feature — target

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Frequency table

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

Likelihood table.

Weather	No	Yes	
Overcast	0	5	5/14 = 0.35
Rainy	2	2	4/14
Sunny	2	3	5/14
All	4/14 = 0.29	10/14 = 0.71	

CONTD...

Solution: To solve this, first consider the below dataset:

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Apply Bayes' theorem:

$$P(Yes|Sunny) = \frac{P(Sunny|Yes) * P(Yes)}{P(Sunny)}$$

$$P(Sunny|Yes) = 3/10 = 0.3$$

$$P(Sunny) = 5/14 = 0.35$$

$$P(Yes) = 10/14 = 0.71$$

$$P(Yes|Sunny) = (0.3 * 0.71)/0.35 = 0.60$$

$$P(No|Sunny) = \frac{P(Sunny|No) * P(No)}{P(Sunny)}$$

$P(Sunny|No) = 0.5$, $P(No) = 0.29$, $P(Sunny) = 0.35$

$$= (0.5 * 0.29)/0.35 = 0.41$$

$$\boxed{\text{Hence on Sunny day Players can Play the game}}$$

ADVANTAGES & DISADVANTAGES

Advantages of Naïve Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for text classification problems.

Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

APPLICATIONS OF NAÏVE BAYES CLASSIFIER:

- It is used for Credit Scoring.
- It is used in medical data classification.
- It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as Spam filtering and Sentiment analysis.

TYPES OF NAÏVE BAYES MODEL

There are three types of Naive Bayes Model, which are given below:

Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

`GaussianNB` implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$\boxed{P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)}$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

TYPES OF NAÏVE BAYES MODEL

Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.

The parameters θ_y is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

TYPES OF NAÏVE BAYES MODEL

Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable.

The decision rule for Bernoulli naive Bayes is based on

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

THANKS


Keep Learning
Keep Growing


Dr. Neeraj Gupta
Assistant Professor, Dept. of CEA
neeraj.gupta@gla.ac.in