Naive Bayes Algorithm

Sentence Classification

What is Bayes Algorithm

Simple algorithm to classify text

Low training time and resources

Requires a set of labeled training data

Will be used to classify new sentences

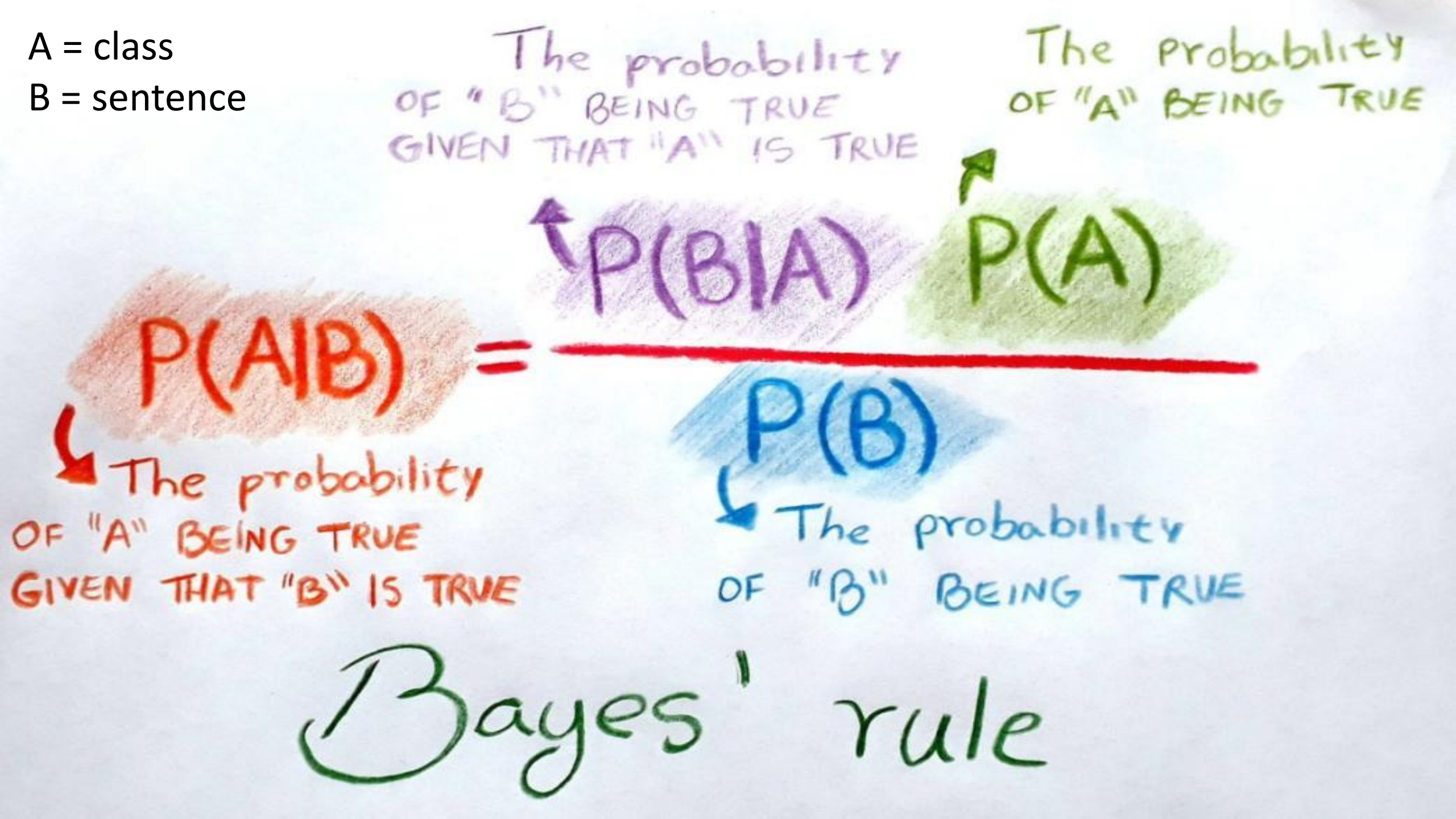
Our data

No.	Training-Text	Label
1	"A great game”	Sports
2	“The election was over”	Not Sports
2	“Very clean match”	Sports
4	“A clean but forgettable game”	Sports
5	“It was a close election”	Not Sports

- Training data consists of two classes
  - Sport or not sport

A = class

B = sentence



How does it work?

- Comparing the probabilities:
  - P (Sport| Hermann played a TT match) = $\frac{P(A\ very\ close\ game|Sports)*P(Sports)}{P(A\ very\ close\ game)}$
  - P (Not Sports | Hermann played a TT match) = $\frac{P(A\ very\ close\ game|not\ Sports)*P(Not\ Sports)}{P(A\ very\ close\ game)}$

Probability of a sentence

- Likelyhood a setence is a sport sentence:
  - P(Sport) = $\frac{Number\ of\ sentences\ in\ class\ "Sports"}{Total\ number\ of\ sentences\ in\ the\ training\ set}$
    - Similarily calculate P(Not Sports)

- „Naive“ Bayes because we think each word is indipendent from the other ones
  - Possibility of a sentence „A very close game“ is calculated like this:
    - $P(A\ very\ close\ game) = P(A) * P(very) * P(close) * P(game)$

Probability of a sentence in a class

- Applying the probabilites of the words to Bayes formula:

$$P(A\ very\ close\ game\ |Sports) =$$

$$P(A|Sports) * P(very|Sports) * P(close|Sports) * P(game|Sports) * P(Sports)$$

Calculating the probabilities

- Now all we have to do is calculate all the different probabilites by counting everything in our training data

No.	Training-Text	Label
1	"A great game”	Sports
2	“The election was over”	Not Sports
3	“Very clean match”	Sports
4	“A clean but forgettable game”	Sports
5	“It was a close election”	Not Sports

- P(Sports) = 3/5 | P(Not Sports) = 2/5
- Probability of a word in class Sports:
  - $P(game|Sports) = \frac{amount\ of\ "game"\ in\ Sports\ sentences}{total\ number\ of\ words\ in\ Sports\ sentences} = \frac{2}{11}$
  - Repeat for other words and other classes

Smoothed probability

- Sentence ”a very close game”
  - P(close|Sports) = 0 !
  - We have “**close**” 0 times in our Sports-data
    - ➔ We would multiply with **0** ➔ everything becomes 0
    - ➔We need Laplace Smoothing

- Laplace smoothing: We add 1 to every count so it’s never zero. To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
  - ➔ P(game|Sports)= $\frac{(2+1)}{(11+14)} = \frac{3}{25}$
    - 2 + 1 times the word “game” in Sports sentences
    - 11 words in Sports sentences
    - 14 possible words

Summary of calculation

- P(Sports) = $\frac{3}{5}$
- P(Not Sports) = $\frac{2}{5}$
- Anzahl (Words|Sports) = 11
- Anzahl (Words|Not Sports) = 9
- Anzahl (possible words) = 14

- $P(A\ very\ close\ game\ |Sports) = P(A|Sports) * P(very|Sports) * P(close|Sports) * P(game|Sports) * P(Sports)$

➔ P(game|Sports)= $\frac{(2+1)}{(11+14)} = \frac{3}{25}$

- 2 + 1 times the word “game” in Sports sentences
- 11 words in Sports sentences
- 14 possible words

- P(a very close game|Sports) =

= P(a|Sports) * P(very|Sports) * P(close|Sports) * P(game|Sports) * P(Sports)

= $\frac{3}{25} * \frac{2}{25} * \frac{1}{25} * \frac{3}{25} * \frac{3}{5} \approx 2{,}7648\ e^{-5}$

➔ Analog für Not Sports:

P(a|Not Sports)*P(very|Not Sports)*P(close|Not Sports)*P(game|Not Sports)*P(Not Sports)

= $\frac{2}{23} * \frac{1}{23} * \frac{2}{23} * \frac{1}{23} * \frac{2}{5} \approx 0{,}57176\ e^{-5}$

➔ Larger number means higher probability

➔ Its more likely that ist a Sports sentence ☺

Homework

- “Hermann plays a TT match”
- P(Sports) = $\frac{4}{6}$
- P(Not Sports) = $\frac{2}{6}$
- Anzahl (Words|Sports) = 15
- Anzahl (Words|Not Sports) = 9
- Anzahl (possible words) = 14

➔ P(game|Sports)= $\frac{(2+1)}{(11+14)} = \frac{3}{25}$

- 2 + 1 times the word “game” in Sports sentences
- 11 words in Sports sentences
- 14 possible words

No.	Training-Text	Label
1	"A great game”	Sports
2	“The election was over”	Not Sports
3	“Very clean match”	Sports
4	“A clean but forgettable game”	Sports
5	“It was a close election”	Not Sports
6	“A very close game”	Sports
---	Target-Text	---
New	“Hermann plays a TT match”	?

Probabilities

- Sports:

  P(Hermann plays a TT match | Sports) =

  $$P(Hermann|Sports) * P(plays|Sport) * P(a|Sports) * P(TT|Sports) * P(match|Sports) * P(Sports)$$

  $$= \frac{1}{29} * \frac{1}{29} * \frac{4}{29} * \frac{1}{29} * \frac{2}{29} * \frac{4}{6} = 2{,}6002e\text{-}7$$

- Not Sports:

  P(Hermann plays a TT match | Not Sports) =

  $$P(Hermann|Not\ Sports) * P(plays|Not\ Sport) * P(a|Not\ Sports) * P(TT|Not\ Sports) * P(match|Not\ Sports)$$

  $$* P(Not\ Sports)$$

  $$= \frac{1}{23} * \frac{1}{23} * \frac{2}{23} * \frac{1}{23} * \frac{1}{23} * \frac{2}{6} = 1{,}0358e\text{-}7$$

  ➔ It will be classified as a Sports-sentence

Vielen Dank für Ihre Aufmerksamkeit

Sources

- https://medium.com/@sweetai/a-brief-introduction-to-naive-bayes-classifier-31cf772195cd
- https://medium.com/analytics-vidhya/naive-bayes-classifier-for-text-classification-556fabaf252b#:~:text=The%20Naive%20Bayes%20classifier%20is,time%20and%20less%20training%20data