Social Media Analytics

Shivani Kumar
shivaniku@iiitd.ac.in

Introduction

Social media analytics is the practice of gathering data from social media websites and analyzing that data using social media analytics tools to make business decisions or for research.

Fake news

- Fake News: fictitious articles deliberately fabricated to deceive readers.
- Hyperpartisan News: extremely biased in favor of a political party.
- Task can be divided into two parts:
  - Fake News Collection: Collecting news contents and social context automatically which provides Datasets for the study of fake news.
  - Fake News Detection: Extracting useful features and build different machine learning models to detect fake news.

Fake reviews

- Opinion spamming
- Not only individuals but companies hire groups to post fake reviews
- Detection: Data collection -> Data classification

Emotions

- Key part of human like artificial intelligence.
- Can be used to mine opinions in social media.
- Can be used in healthcare for psychological analysis.
- A predefined set of emotions.
- Emotion classification or Emotion classification in conversations.

Mental Health

- Can develop computational models to predict the emergence of depression and Post-Traumatic Stress Disorder in Twitter users.
- Collect labelled dataset of Twitter data and details of depression history.
- Extract predictive features measuring affect, linguistic style, and context from participant tweets.
- Build model.

Sentiments

- Sentiment Analysis- you are supplied with a phrase, or a list of phrases and your classifier is supposed to tell if the sentiment behind that is positive, negative or neutral.
- Sometimes, the third attribute is not taken to keep it a binary classification problem.
- Advantages:
    - It helps to predict customer behavior for a particular product.
    - It can help to test the adaptability of a product.
    - It can easily automate the process of determining how well did a movie run by analyzing the sentiments behind the movie's reviews from a number of platforms.
    - And many more!

Sentiment Analysis

Hands-on

Naive Bayes

Consider...

- Three classes

Car

Art

Space

Consider...

- Three classes and three documents

Car	Art	Space
The Mona Lisa was painted by Da Vinci.	Saturn is lighter than water!	A Ferrari is a fast but not faster than a Bugatti.

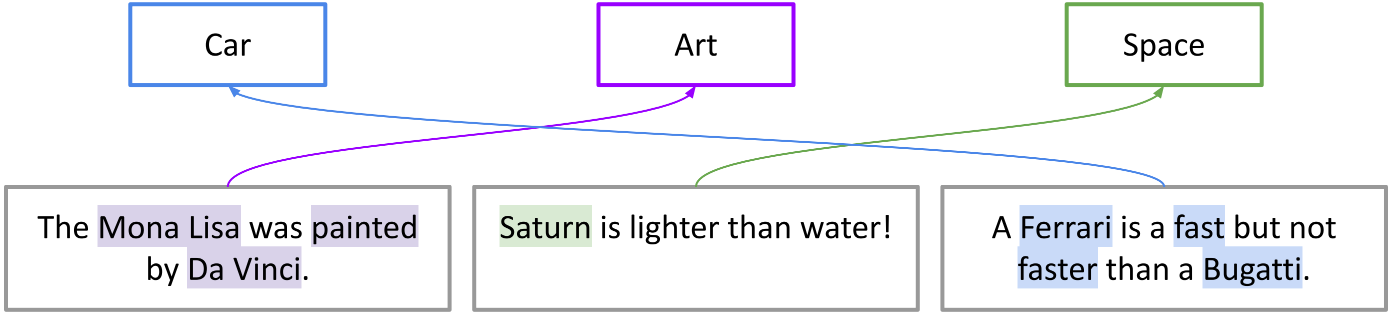
Consider...

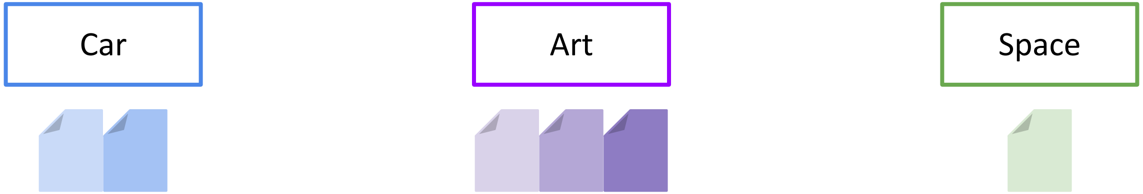
- Three classes and three documents

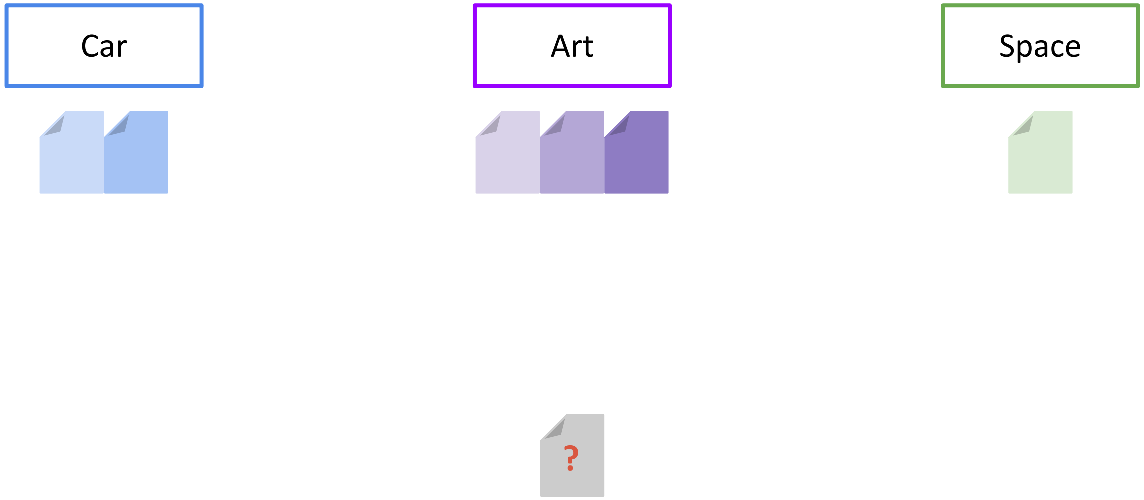
Car	Art	Space
The Mona Lisa was painted by Da Vinci.	Saturn is lighter than water!	A Ferrari is a fast but not faster than a Bugatti.

Consider...

- Three classes and three documents




Car
Art
Space


Car
Art
Space
?

What do you think?

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Which of the following do you find more likely?

1. Steve is a librarian.
2. Steve is a farmer.

Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

What do you think?

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Which of the following do you find more likely?

1. Steve is a librarian.
2. Steve is a farmer.

Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

What do you think?

Number of farmers in the world?

Number of librarians in the world?

Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

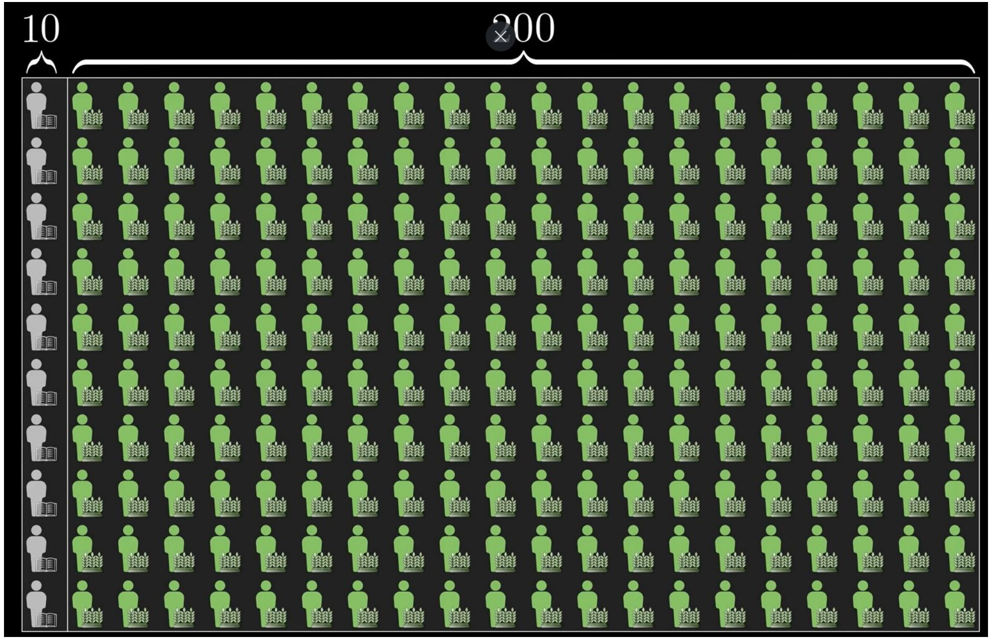
What do you think?

Number of farmers in the world?
Number of librarians in the world?

1 librarian for every 20 farmers!

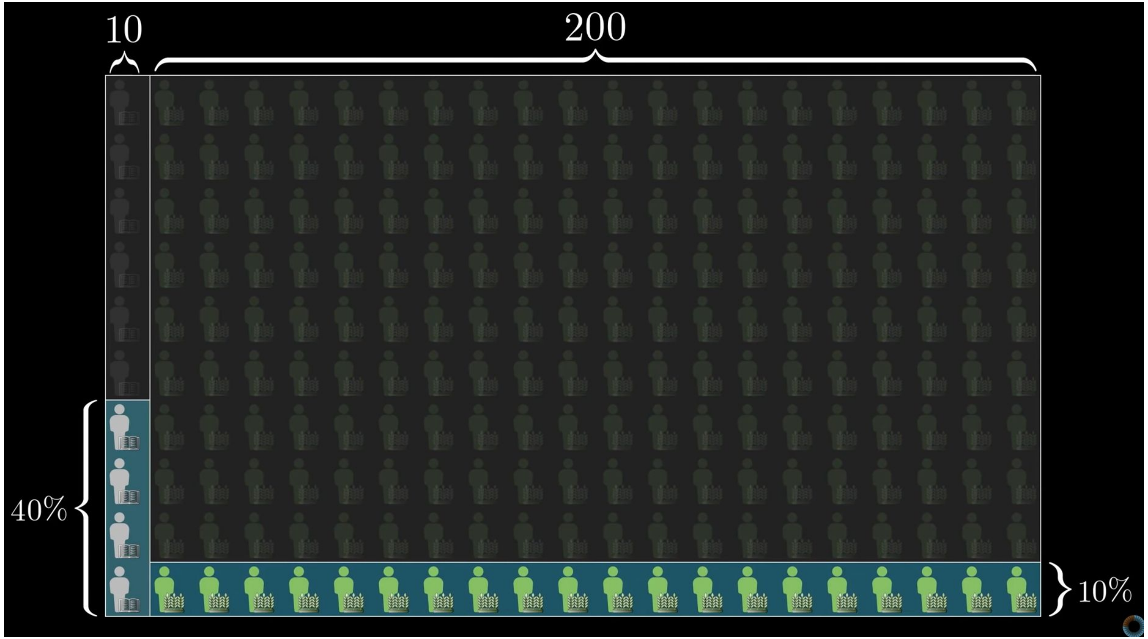
Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

What do you think?



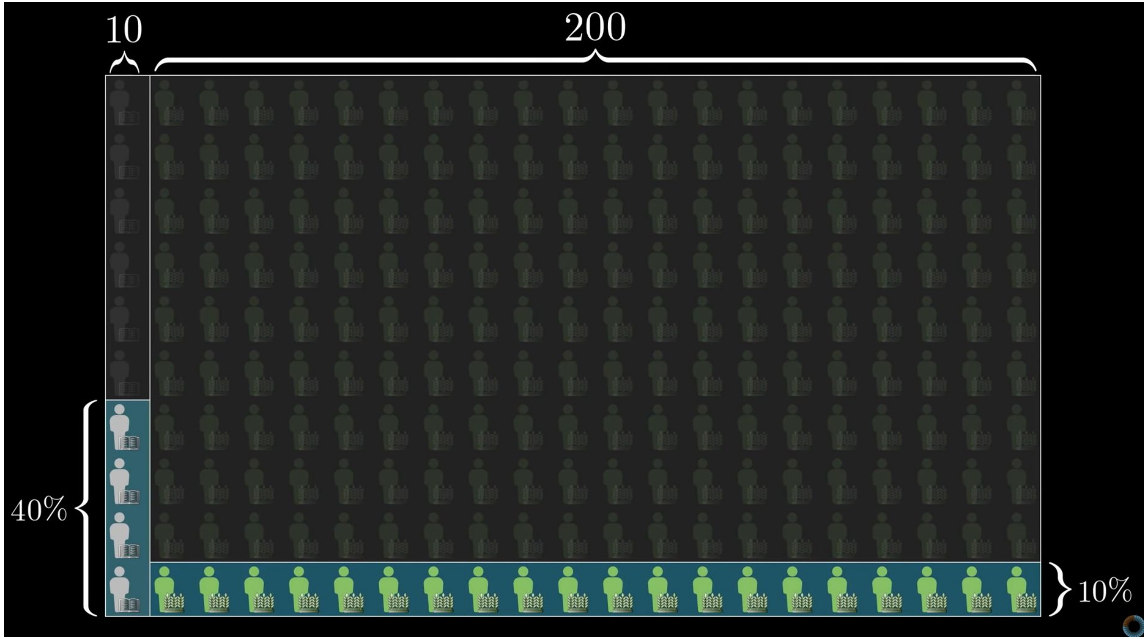
Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

What do you think?



Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

What do you think?



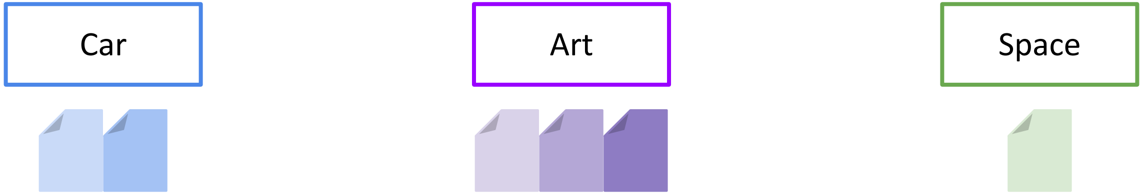
P(Librarian given description)
= 4/(4+20) = 16.7%

P(Farmer given description)
= 20/(4+20) = 83.3%

Source: https://www.youtube.com/watch?v=HZGCoVF3YvM

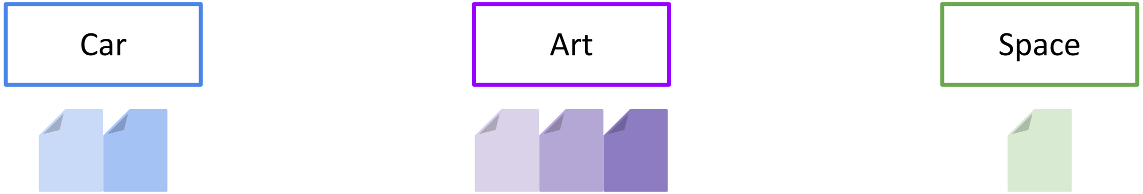
Prior Probability

- The probability of the class



Prior Probability

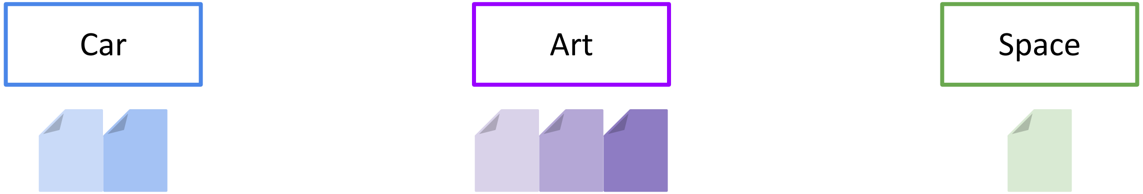
- The probability of the class



$$P(c_1) = \frac{\text{\# docs in } c_1}{\text{\# total docs}}$$

Prior Probability

- The probability of the class



$$P(c_1) = \frac{\#\text{ docs in } c_1}{\#\text{ total docs}}$$

$P(c_1) = 2/6$ $P(c_2) = 3/6$ $P(c_3) = 1/6$

Likelihood

- The chance of something happening.
  - The chance that given a class, the document belongs to it.

Likelihood

d: The Mona Lisa was
painted by Da Vinci.

Conditioned on d
belonging to the
class c

P(d|c) = P(The Mona Lisa was painted by Da Vinci | c). Normally,

P(The Mona Lisa was painted by Da Vinci) = P(The | Mona Lisa was painted by Da Vinci) P(Mona | Lisa was painted by Da Vinci) P(Lisa | was painted by Da Vinci) … P(Vinci)

Likelihood

d: The Mona Lisa was painted by Da Vinci.

Conditioned on d belonging to the class c

P(d|c) = P(The Mona Lisa was painted by Da Vinci | c). Normally,

P(The Mona Lisa was painted by Da Vinci) = P(The | Mona Lisa was painted by Da Vinci) P(Mona | Lisa was painted by Da Vinci) P(Lisa | was painted by Da Vinci) … P(Vinci)

The Naive Bayes assumption: All terms are independent of each other.

P(The Mona Lisa was painted by Da Vinci) = P(The) P(Mona) P(Lisa) … P(Vinci)

Likelihood

d: The Mona Lisa was painted by Da Vinci.

Conditioned on d belonging to the class c

P(d|c) = P(The Mona Lisa was painted by Da Vinci | c). Normally,

P(The Mona Lisa was painted by Da Vinci) = P(The | Mona Lisa was painted by Da Vinci) P(Mona | Lisa was painted by Da Vinci) P(Lisa | was painted by Da Vinci) … P(Vinci)

The Naive Bayes assumption: All terms are independent of each other.

P(The Mona Lisa was painted by Da Vinci) = P(The) P(Mona) P(Lisa) … P(Vinci)

P(d|c) = P(The | c) P(Mona | c) P(Lisa | c) P(was | c) P (painted | c) P(by | c) P(Da | c) P(Vinci | c)

$$P(\text{The} \mid c) = \frac{\#\ \text{'The' occurs in class } c}{\#\ \text{words in class } c}$$

Probability of the document

d: The Mona Lisa was painted by Da Vinci.

NOT conditioned on d belonging to the class c

P(d) = P(The Mona Lisa was painted by Da Vinci)

= P(The) P(Mona) P(Lisa) P(was) P(painted) P(by) P(Da) P(Vinci)

Probability of the document

d: The Mona Lisa was
painted by Da Vinci.

P(d) = P(The Mona Lisa was painted by Da Vinci)

= P(The) P(Mona) P(Lisa) P(was) P(painted) P(by) P(Da) P(Vinci)

NOT conditioned
on d belonging to
the class c

$$P(\text{The}) = \frac{\#\ \text{‘The’ occurs in the corpus}}{\#\ \text{total words}}$$

Probability of the document

d: The Mona Lisa was painted by Da Vinci.

NOT conditioned on d belonging to the class c

P(d) = P(The Mona Lisa was painted by Da Vinci)

= P(The) P(Mona) P(Lisa) P(was) P(painted) P(by) P(Da) P(Vinci)

$$P(\text{The}) = \frac{\text{\# 'The' occurs in the corpus}}{\text{\# total words}}$$

Same for all classes.

Naive Bayes

Bayes Theorem:

$$P(c \mid d) = \frac{\overset{\text{Likelihood}}{P(d \mid c)}\,\overset{\text{Prior}}{P(c)}}{\underset{\text{Normalization Constant}}{P(d)}}$$

To find the most probable class:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

Dropping the denominator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

(Since documents are made up of words)

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$ <-- The Naive Bayes assumption

Finally-

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{x \in X} P(x \mid c)$$

Working example

Car	Bugatti looks good in black
	Ferrari is fast
Art	The Mona Lisa is in Louvre
	Sistine ceiling was painted by Michelangelo
	Da vinci painted
Space	Jupiter is the biggest planet

d: The Mona Lisa was painted by Da Vinci.

Working example: vocab

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

d: The Mona Lisa was painted by Da Vinci.

Car	Bugatti looks good in black
	Ferrari is fast
Art	The Mona Lisa is in Louvre
	Sistine ceiling was painted by Michelangelo
	Da vinci painted
Space	Jupiter is the biggest planet

Working example: prior probabilities

- P(car) = # docs in class car / # total docs = 2/6 = 1/3
- P(art) = # docs in class art / # total docs = 3/6 = 1/2
- P(space) = # docs in class space / # total docs = 1/6

Working example: likelihoods

- P(d | car)

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # 'The' occurs in class car / # words in class car

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # 'The' occurs in class car / # words in class car
  - = (0/8) (0/8) … = 0

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # 'The' occurs in class car / # words in class car
  - = (0/8) (0/8) … = 0
- P(d | art)

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # ‘The’ occurs in class car / # words in class car
  - = (0/8) (0/8) … = 0
- P(d | art)
  - = P(The|art) P(Mona|art) P(Lisa|art) P(was|art) P(painted|art) P(by|art) P(Da|art) P(Vinci|art)

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # ‘The’ occurs in class car / # words in class car
  - = (0/8) (0/8) … = 0
- P(d | art)
  - = P(The|art) P(Mona|art) P(Lisa|art) P(was|art) P(painted|art) P(by|art) P(Da|art) P(Vinci|art)
  - = (1/15) (1/15) (1/15) (1/15) (2/15) (1/15) (1/15) (1/15) = 7.8×10^{-10}

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: likelihoods

- P(d | car)
  - = P(The Mona Lisa was painted by Da Vinci | car)
  - = P(The|car) P(Mona|car) P(Lisa|car) P(was|car) P(painted|car) P(by|car) P(Da|car) P(Vinci|car)
    - P(The|car) = # ‘The’ occurs in class car / # words in class car
  - = (0/8) (0/8) … = 0
- P(d | art)
  - = P(The|art) P(Mona|art) P(Lisa|art) P(was|art) P(painted|art) P(by|art) P(Da|art) P(Vinci|art)
  - = (1/15) (1/15) (1/15) (1/15) (2/15) (1/15) (1/15) (1/15) = 7.8 x 10^{-10}
- P(d | space)
  - = 0

Word	Car	Art	Space
The	0	1	1
Mona	0	1	0
Lisa	0	1	0
was	0	1	0
painted	0	2	0
by	0	1	0
Da	0	1	0
Vinci	0	1	0
…	…	…	…
Total	**8**	**15**	**5**

Working example: Naive Bayes

- P(car|d) = P(d|car) x P(car)
  - = 0
- P(art|d) = P(d|art) x P(art)
  - = 7.8 x 10^{-10} x 1/2
  - = 3.9 x 10^{-10}
- P(space|d) = P(d|space) x P(space)
  - = 0

P(art|d) > P(car|d) = P(space|d). Thus, the document d belongs the class ‘art’

Working example

Car	Bugatti looks good in black
	Ferrari is fast
Art	The Mona Lisa is in Louvre
	Sistine ceiling was painted by Michelangelo
	Da vinci painted
Space	Jupiter is the biggest planet

d: The Mona Lisa was painted by Da Vinci.

Naive Bayes: Hands-on

- Sentiment classification on movie reviews
- Using Python and NLTK

Next up...

- Sentiment Analysis of tweets
  - Authorize twitter API client.
  - Make a GET request to Twitter API to fetch tweets for a particular query.
  - Parse the tweets. Classify each tweet as positive, negative or neutral.

Other tools for Sentiment Analysis...

TextBlob

- A Python library for processing textual data.
- Provides a simple API
  - Part-of-speech tagging
  - Noun phrase extraction
  - Sentiment analysis
  - Classification
  - Translation
  - …

VADER

- VADER: Valence Aware Dictionary and sEntiment Reasoner
- Is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- Not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

Thank You