## Artificial Intelligence

Unit 06 Naïve Bayes

By: Syeda Saleha Raza



AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.

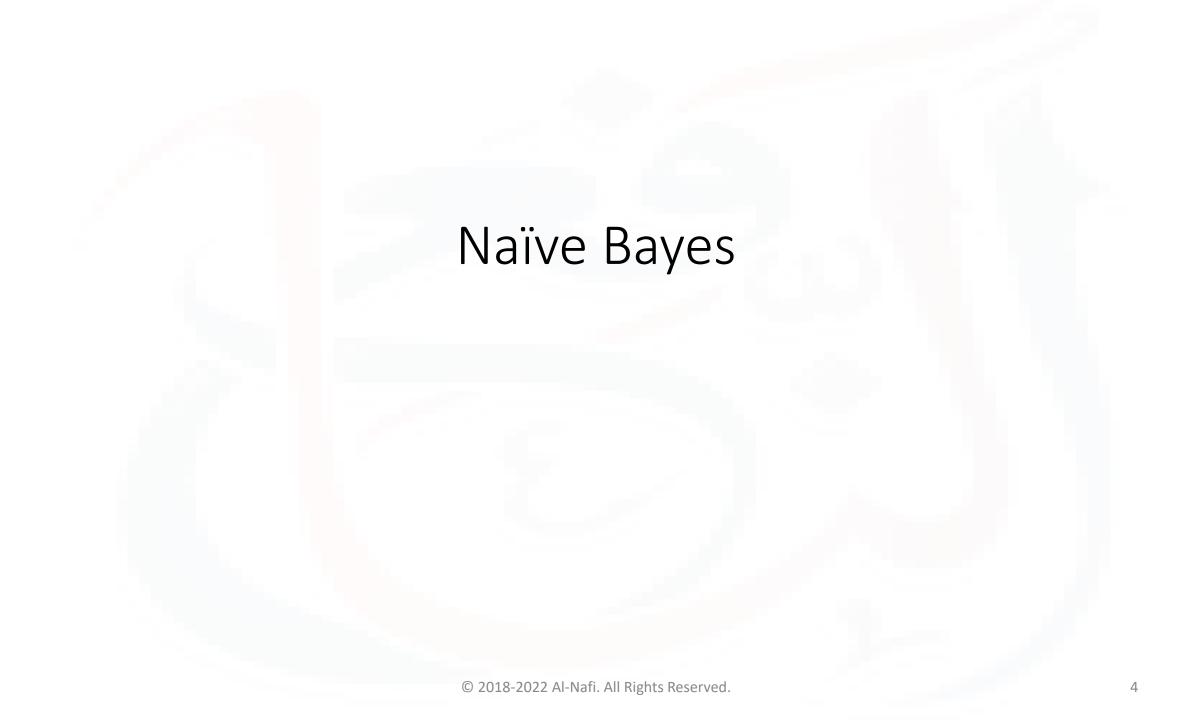
# اللَّهُمَّ إِنِي أَسُالُكَ عِلْمًا تَّافِعًا، وَرِزْقًا طَيِّبًا، وَعَمَلًا مُّتَقَبَّلًا،

(O Allah, I ask You for beneficial knowledge, goodly provision and acceptable deeds)

اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طبیب رزق کا، اور اس عمل کا
(Sunan Ibn Majah: 925)

#### Outline

- Naïve Bayes Model
  - Probabilities Recap
  - Learning Naïve Bayes
  - Strength and Weaknesses
  - Code Demo

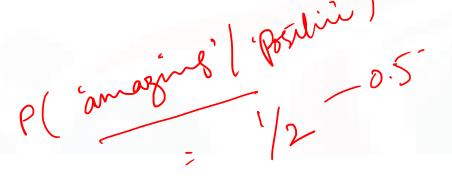


# Bayes' Rule Applied to Documents and Classes

For a document d and a class c

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

#### Parameter Estimation



$$\hat{P}(\underline{w_i} | \underline{c_j}) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

 $\hat{P}(\underline{w_i} | c_j) = \frac{count(w_i, c_j)}{\sum count(w, c_j)}$  fraction of times word  $w_i$  appears among all words in documents of topic  $c_j$ 

Review	Class
"The movie was amazing!"	Positive
"I didn't like the film."	Negative
"Great acting and an engaging plot."	Positive
"Terrible movie, I would not recommend it to anyone."	Negative

## Applying Multinomial Naive Bayes Classifiers to Text Classification

$$c_{NB} = \underset{c_{j} \in C}{\operatorname{argmax}} P(c_{j}) \prod_{i \in positions} P(x_{i} \mid c_{j})$$

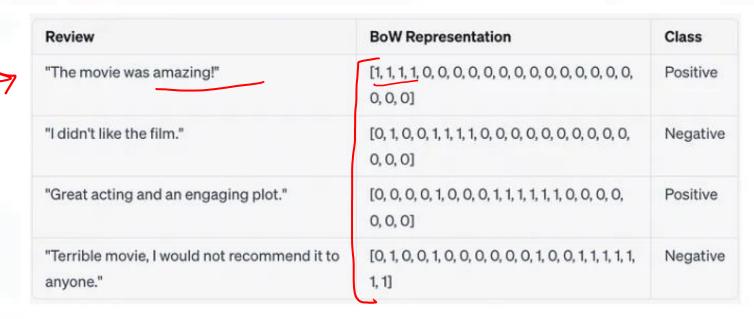
Review	Class
"The movie was amazing!"	Positive
"I didn't like the film."	Negative
"Great acting and an engaging plot."	Positive
"Terrible movie, I would not recommend it to anyone."	Negative

Review	Class
"The movie was amazing!"	Positive
"I didn't like the film."	Negative
"Great acting and an engaging plot."	Positive
"Terrible movie, I would not recommend it to anyone."	Negative

["the," "movie," "was," "amazing," "I," "didn't," "like," "film," "great," "acting," "and," "an," "engaging," "plot," "terrible," "would," "not," "recommend," "it," "to," "anyone"].

Review	Class
"The movie was amazing!"	Positive
"I didn't like the film."	Negative
"Great acting and an engaging plot."	Positive
"Terrible movie, I would not recommend it to anyone."	Negative

["the," "movie," "was," "amazing," "I," "didn't," "like," "film," "great," "acting," "and," "an," "engaging," "plot," "terrible," "would," "not," "recommend," "it," "to," "anyone"].



- Now, let's assume we have a new movie review:
  - "The acting was great!"
- We can convert this review into its BoW representation:
  - [0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].

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

Class Prior Probability		Prior Probability Likelihood	
Positive	0.5	6.004	0.002
Negative	0.5	0.000000000002	0.00000000001

### Classification using Naïve Bayes

$\hat{P}(c) =$	_	$N_c$
<i>I</i> (c)	_	$\overline{N}$

$$\hat{P}(w \mid c) = \frac{count(w,c) + 1}{count(c) + |V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	С
	2	Chinese Chinese Shanghai	С
	3	Chinese Macao	С
	4	Tokyo Japan Chinese	jr
Test	5	Chinese Chinese Tokyo Japan	?

#### **Priors:**

$$P(c) = \frac{3}{4} \frac{1}{4}$$

$$P(j) = \frac{3}{4} \frac{1}{4}$$

#### **Conditional Probabilities:**

$$P(Chinese | c) = (5+1) / (8+6) = 6/14 = 3/7$$
  
 $P(Tokyo | c) = (0+1) / (8+6) = 1/14$   
 $P(Japan | c) = (0+1) / (8+6) = 1/14$   
 $P(Chinese | j) = (1+1) / (3+6) = 2/9$   
 $P(Tokyo | j) = (1+1) / (3+6) = 2/9$ 

$$P(Japan|j) = (1+1)/(3+6) = 2/9$$

#### **Choosing a class:**

$$P(c|d5) \propto (3/4)^3 (3/7)^3 * 1/14 * 1/14$$
  
 $\approx 0.0003$ 

$$P(j|d5) \propto \frac{1/4 * (2/9)^3 * 2/9 * 2/9}{\approx 0.0001}$$

6/2/01/

### Dealing with small values

- Multiplying lots of small probabilities (all are under 1) can
- lead to numerical underflow ...

$$\log \prod_{i} x_i = \sum_{i} \log x_i$$

#### Smoothing

- What if we have seen no training documents with the word fantastic and classified in the topic positive (thumbs-up)
- To deal with low counts, it can be helpful to smooth probabilities
- Smoothing term! is a hyperparameter, which must be tuned on a development set
- Laplace (add-1) smoothing: widely used

#### Strength

- Very Fast, low storage requirements
- Robust to Irrelevant Features Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features Decision
   Trees suffer from fragmentation in such cases especially if little data
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification (But we will see other classifiers that give better accuracy)

## Naïve Bayes - Code Walkthrough

#### Resources

- https://www.coursera.org/learn/machinelearning/lecture/du981/backpropagation-intuition
- https://mattmazur.com/2015/03/17/a-step-by-step-backpropagationexample/
- https://scikit-neuralnetwork.readthedocs.io/en/latest/index.html
- 7 NB.pdf (stanford.edu)



To ask questions, Please use communities link (for respective course) within portal

https://portal.alnafi.com/enrollments