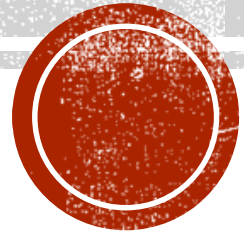
TEXT CLASSIFICATION

BY: Khaled Mohammed

Teaching Assistant at Faculty of Engineering

Khaled.edu.engineer@gmail.com



NAIVE BAYES

- Naïve Bayes is a **probability-based classifier.**
  - Its goal is to find the class **C** (for example: *positive vs negative* sentiment) that is most likely given the words in a document:

$$\hat{C} = \arg\max_{c} P(c \mid \text{words})$$

- Using Bayes' theorem:

$$P(c \mid w_1, w_2, \ldots, w_n) = \frac{P(c)P(w_1, w_2, \ldots, w_n \mid c)}{P(w_1, w_2, \ldots, w_n)}$$

- Since the denominator is the same for all classes, we compare only:

$$P(c) \cdot P(w_1, w_2, \ldots, w_n \mid c)$$

THE NAÏVE ASSUMPTION

- We assume **every word is independent of the others** given the class, So:

$$P(w_1, w_2, \ldots, w_n \mid c) = \prod_{i=1}^{n} P(w_i \mid c)$$

- This assumption is **not true in real language**, but it works extremely well because:
  - It simplifies the math
  - Works well in high-dimensional sparse text
  - Word frequencies capture a lot of signal (ex: "good", "love", "great" → positive)

HOW NAÏVE BAYES WORKS FOR TEXT

- Given a document with words:
  - I love this movie
- We want to compute:

$$P(\text{positive} \mid \text{document}) \propto P(\text{positive}) \times P(I|\text{pos}) \times P(love|\text{pos}) \times P(this|\text{pos}) \times P(movie|\text{pos})$$

- And the same for negative.
- The class with the **higher** probability is selected.

WHICH NAÏVE BAYES IS USED FOR TEXT?

- We use **Multinomial Naïve Bayes** because it counts words:

$$P(w|c) = \frac{\text{count}(w, c) + 1}{\text{total words in class } c + |V|}$$

- Why Laplace smoothing (+1)?
  - To avoid zeros
  - If a word never appears in a class → its probability becomes 0 → whole product becomes 0
  - Laplace smoothing solves it

DETAILED EXAMPLE

- We have 4 training documents, each belonging to class **c** or **j**:

Doc	Words	Class
1	Chinese Beijing Chinese	c
2	Chinese Chinese Shanghai	c
3	Chinese Macao	c
4	Tokyo Japan Chinese	j

- Test document (Doc 5):
  - Chinese Chinese Chinese Tokyo Japan
- We want to decide: **Is it class c or class j?**

STEP 1: PRIORS

- Because:
  - 3 documents are class **c**
  - 1 document is class **j**

$$\hat{P}(c) = \frac{3}{4}, \qquad \hat{P}(j) = \frac{1}{4}$$

STEP 2: COUNT WORDS WITH LAPLACE SMOOTHING

- **Vocabulary:**
  - Chinese, Beijing, Shanghai, Macao, Tokyo, Japan
  - $|V| = 6$
- **Count total words per class:**
  - **For class c**
    - Docs 1, 2, 3:
    - Doc1: Chinese Beijing Chinese → 3 words
    - Doc2: Chinese Chinese Shanghai → 3 words
    - Doc3: Chinese Macao → 2 words
      Total words = 8
  - **For class j**
    - Doc 4:
    - Tokyo Japan Chinese → 3 words
      Total words = 3

STEP 3: COMPUTE CONDITIONAL PROBABILITIES

$$P(w|c) = \frac{\text{count}(w,c)+1}{\text{total words in c} + |V|}$$

- For class **c**:
  - count(Chinese, c) = 5
  - count(Tokyo, c) = 0
  - count(Japan, c) = 0
- So:

$$P(\text{Chinese}|c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$$

$$P(\text{Tokyo}|c) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$P(\text{Japan}|c) = \frac{0+1}{8+6} = \frac{1}{14}$$

CONT.

- For class **j**:
- Counts from Doc 4:
  - Tokyo: 1
  - Japan: 1
  - Chinese: 1
- Total = 3

$$P(\text{Chinese}|j) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Tokyo}|j) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Japan}|j) = \frac{1+1}{3+6} = \frac{2}{9}$$

STEP 4: COMPUTE PROBABILITY OF CLASS FOR DOC 5

- Document 5:
  - Chinese Chinese Chinese Tokyo Japan
- Number of each word:
  - Chinese ×3
  - Tokyo ×1
  - Japan ×1
- **For class c**

$$P(c|d_5) \propto P(c) \cdot P(\text{Chinese}|c)^3 \cdot P(\text{Tokyo}|c) \cdot P(\text{Japan}|c)$$

- Substitute:

$$= \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14}$$

- Numerically ≈ **0.0003**

CONT.

- **For class j**

$$P(j|d_5) \propto P(j) \cdot P(\text{Chinese}|j)^3 \cdot P(\text{Tokyo}|j) \cdot P(\text{Japan}|j)$$

$$= \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9}$$

- Numerically ≈ **0.0001**

FINAL DECISION

$$P(c|d_5) > P(j|d_5)$$

So the document is classified as **class c**.