# Naive Bayes

1. Intro
2. Spam data
3. Mathmetical Intution
4. Naive Bayes Assumption
5. Train / Test time complexities
6. Space complexity

## Google

Mail 1 : " I, Nigerian prince need your help. Send money "

Mail 2 : " Meeting scheduled at 6pm. Kindly revert "

lottery , Million dollar , Jackpot

## Sentiments

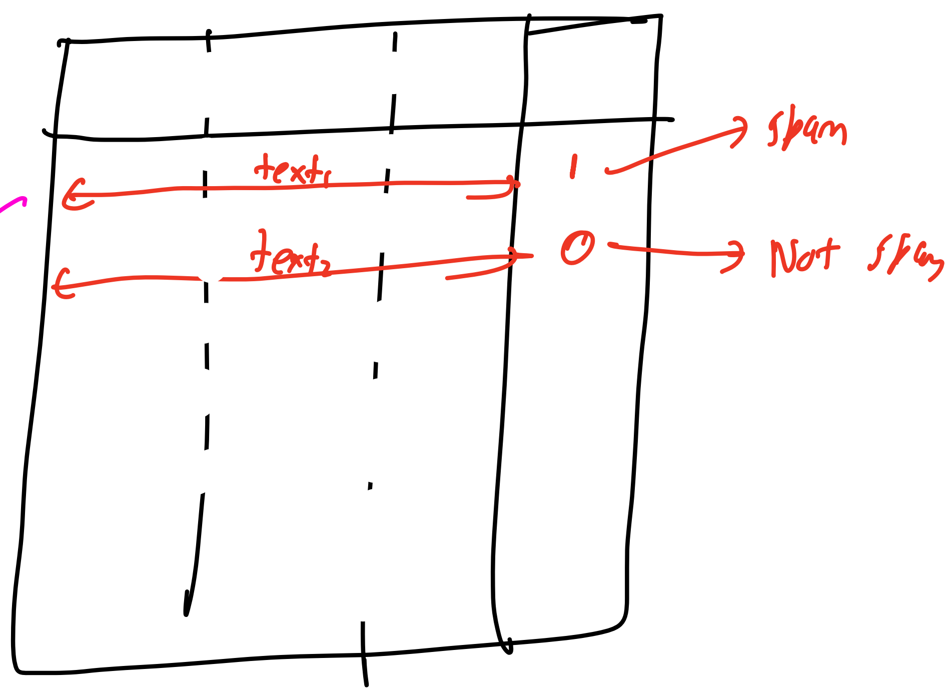
"terrible" , "bad" , "pathetic" → -ve

"great" , "good" → +ve

NB
80%

SOTA
↓
transformers (later)
98.5%

⇒ "I am shivank"
$w_1$ $w_2$ $w_3$




$$text_1 = w_1^i \; w_2^i \; w_3^i \; .... \; w_d^i$$

"This is shivank"

↓

① Tokenisation [ "This", "is", "shivank" ]

good Good

② Lowercase

⇒ I, Shivank Agrawal wants to teach!

⇓

removing all punctuation ⇒ regex

④ text → remove stopwords

"I", "and", "the"

⑤ gramatical errors ←

Lema → Steming ⇒ eating eat

money

$p(\text{money} \mid \underline{\text{spam}})$

$p(\text{money} \mid \underline{\text{ham}})$

play plays playing

I eat food
I eaten food
I eats foods

I eat food

## Mathematical Intuition

email, $\longrightarrow$ spam or not spam

email, $\longrightarrow$ 0/1

$P(y=1 \mid \text{email}_1)$

$P(y=0 \mid \text{email}_1)$

$P(y=0 \mid w_1^i, w_2^i, w_3^i, w_4^i, \ldots, w_d^j)$

$P(y=1 \mid w_1^i, w_2^i, w_3^i, w_4^i, \ldots, w_d^j)$

## Bayes Theorem

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

$P(y=1 \mid w_1^i, w_2^i, w_3^i, w_4^i, \ldots, w_d^j)$

$$P_1(y=1 \mid x) = \frac{P(x \mid y=1) * P(y=1)}{P(x)}$$

$$P_2(y=0 \mid x) = \frac{P(x \mid y=0) * P(y=0)}{P(x)}$$

40/3 ✓

20/3 ✓

$\boxed{y=1}$

$$P(y=1 \mid w_1 \dots w_d) = \frac{P(w_1 \dots w_d \mid y=1) \cdot P(y=1)}{\cancel{K}}$$

$$P(y=0 \mid w_1 \dots w_d) = \frac{P(w_1 \dots w_d \mid y=0) \cdot P(y=0)}{\cancel{K}}$$

$P(DF) = 0.01$

$P(\text{smoke}) = 0.1$

$P(\text{smoke} \mid DF) = 0.9$

$$P(DF \mid \text{smoke}) \Rightarrow ? \Rightarrow \frac{P(DF) * P(\text{smoke} \mid DF)}{P(\text{smoke})}$$

$$\Rightarrow \frac{0.01 * 0.9}{0.1}$$

$$\Rightarrow 0.09$$

$$P(y=1 \mid w_1 \dots w_d) = \frac{P(w_1 \dots w_d \mid y=1) \cdot P(y=1)}{\cancel{K}}$$

$$P(y=0 \mid w_1 \dots w_d) = \frac{P(w_1 \dots w_d \mid y=0) \cdot P(y=0)}{\cancel{K}}$$

$$P(y=1) \rightarrow \frac{\#\text{ train points with } y_i = 1}{\text{total } \#\text{ of train pts}}$$

$$P(y=0) \rightarrow \frac{\#\text{ of train point with } y=0}{\text{total } \#\text{ of train pts}}$$

prince | spam

viagra | spam

$$P(\underset{\text{prince}}{\underset{\downarrow}{w_1}}, \underset{\text{viagra}}{\underset{\downarrow}{w_2}} \mid \text{spam}) \approx P(w_1 \mid \text{spam}) \cdot P(w_2 \mid \text{spam})$$

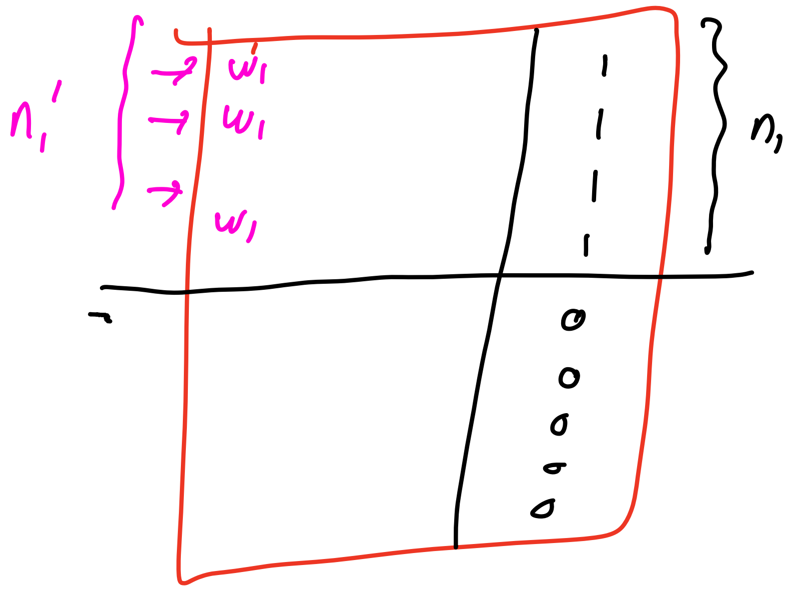
$$P(w_1, w_2 \mid y=1) = P(w_1 \mid y=1) * P(w_2 \mid y=1)$$

Naive Assumption

$$P(w_1 w_2 w_3 \ldots w_d \mid y=1)$$

$$= P(w_1 \mid y=1) \cdot P(w_2 \mid y=1) \ldots P(w_d \mid y=1)$$

$$P(w_1 \mid y=1) = \frac{n_1'}{n_1}$$




$$P(y=1 \mid \overset{w_1\, w_2 \ldots w_d}{\text{text}}) \simeq P(w_1 w_2 \ldots w_d \mid y=1) \frac{* P(y=1)}{K}$$

$$= P(w_1 \mid y=1) * P(w_2 \mid y=1) \ldots P(w_d \mid y=1) \cdot P(y=1)/K$$

$$\simeq \prod_{i=1}^{d} P(w_i \mid y=1) \cdot P(y=1)/K$$

$$P(y=1 \mid \text{text}) \approx \prod_{i=1}^{d} \underbrace{p(w_i \mid y=1)}_{\text{likelihood}} \cdot \overbrace{P(y=1)}^{\text{class priors}} \big/ K$$

compare

$$P(y=0 \mid \text{text}) \approx \prod_{i=1}^{d} p(w_i \mid y=0) \cdot P(y=0) \big/ K$$

$$P(w_1\ w_2 \mid y=1)$$

$$\approx P(w_1 \mid y=1) \cdot P(w_2 \mid y=1)$$