

- Prob review
- Text as data
- Naive Bayes

# Probability

A - Rain

B - Dog barks

Frequentist / Bayesian

A - Rain on a day

$$P(A) = \frac{20}{100} = \frac{1}{5}$$

B - Dog barks in  
this hour

$$P(B) = \frac{300}{1000} = \frac{3}{10}$$

$$P(A \text{ and } B) = P(A \cap B)$$

$$A \perp B$$

$$P(A)P(B)$$

$$= \left(\frac{1}{5}\right) \left(\frac{3}{10}\right) = \frac{3}{50}$$

$$P(A \text{ or } B) = P(A \cup B)$$

$$= P(A) + P(B)$$

$$= \frac{2}{10} + \frac{3}{10} = \frac{5}{10}$$

*(Note: The fraction 5/10 is circled in the original image, with a '1' written above the 5 and a '2' written below the 10, suggesting a simplification to 1/2.)*

$$P(A^c) = 1 - P(A) = \frac{4}{5}$$

$$P(B^c) = 1 - P(B) = \frac{7}{10}$$



# Conditional Prob

$P(\underline{A} | B)$  Likelihood Prior

$$= \frac{P(B | A) P(A)}{P(B)}$$

$P(A), P(B | A) + P(A^c | B) \leq 1$



$$P(A|B) \propto P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\frac{P(B)}{1} = 0.3$$

$$P(A|B) = \frac{(0.2)(0.2)}{0.3}$$

↓  
Posterior =

$$0.133$$

$$A \perp B$$

# Text - as - data

"You win 1 million dollars. Just give us your bank account information."

-  $N = 100$  emails

# Spam Data

<u>Email</u>	<u>Spam</u>	<u>Text</u>
1	1	~'you win..'
.	0	
!	0	
:	1	
.	0	
100	1	

~ You win ~~X~~ million  
dollars ... //

① Pre-processing <sup>words</sup> / total

② Document term  
matrix

# Pre-processing

① Tokenization - split text  
data into terms

$n$ -grams  
(unigram) 1 - gram = word  
2 - gram = phrase  
 $n \geq 2$  = phrase

$n=1$  "I went to New York"

$n=2$  {  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_



"you win \*million dollars."

you win million dollars you win  
win million million dollars

- stemming  $\Rightarrow$  <sup>millions</sup> million  $\rightarrow$  million-
- standardize text
  - o lowercase
  - o remove #s, punctuation

going  
so  
some  $\rightarrow$  go
- stop-word removal  $\rightarrow$  "the", "is", "for"



# ① Preprocessing

- tokenization
- stemming
- standardize text
- stop word removal

② Turning text into a  
document term matrix

DTM

$N=100$ , even row is a labeled  
email

DTM is a matrix with  
Structure {  $N$  rows (observation / document)  
 $W$  columns (each term in all  
 $N$  documents)

Entries ? ① Term frequency (TF)  
② Term-frequency /

inverse doc.

frequency (TF-IDF)

Note: Collection of  
 $N$  docs is called  
a "Corpus"

① You win million dollars →

①	<u>Span</u>	<u>you</u>	<u>win</u>	<u>million</u>	<u>dollars</u>	<u>You</u> <u>win</u>
②	0	0	0	0	0	0

② Let's hang out tomorrow

# Naive Bayes for Spam detection

Target/Class

{ 1 Spam  
0 Not Spam

Features

{ Terms / words

$S \equiv \text{spam class}, \text{Words} \equiv w$

$$P(S=1 | w)$$

(A)

(B)

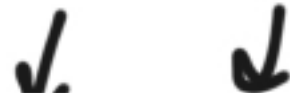


$$\stackrel{(A)}{=} \boxed{P(w | S=1) P(S=1)}$$

$$P(w) \in$$

$$w \in \prod^{N \times T}$$

$$w_1 \equiv \{ \text{you, win, million, dollars} \}$$



$$P(w_1, w_2, w_3 | S=1) = P(w_1 | S=1) P(w_2 | S=1) \dots$$

$$P(w | S=1) = \prod_{i=1}^n p(w_i | S=1)$$

$$P(\text{you, win, million, dollars} | S=1)$$

$$= p(\text{you} | S=1) \cdot p(\text{win} | S=1)$$

$$\textcircled{1} \quad p(\text{million} | S=1) p(\text{dollars} | S=1)$$

$$= (0.01 / 0.02) p(0.10) p(0.20)$$

# of times "you" appears in span docs

# of words in span docs



$$P(w; | s=1) = 0.000004$$

$i=1$  Likelihood

Prior  $P(s=1) = \frac{\# \text{ of spam docs}}{\# \text{ of docs}}$

$$= \frac{\% \text{ of spam}}{\text{docs}}$$

$$= 0.50$$

Marginal  
Likelihood

$$0.000003$$

$$P(w_i) = \overbrace{P(w_i | s=1) P(s=1)}$$

$$+ \overbrace{P(w_i | s=0) P(s=0)}$$



$$P(S=1 | W_1) = \frac{P(W_1 | S=1) P(S=1) + P(W_1 | S=0) P(S=0)}{P(W_1 | S=1) P(S=1) + P(W_1 | S=0) P(S=0)}$$

$$= \frac{0.000004 (0.50)}{0.000004 (0.50) + 0.000003 (0.50)}$$

$$= 0.57$$

$$P(S=1 | w_i) = 0.57$$

$$P(S=1 | w_i) > 0.5 \Rightarrow 1$$

$$< 0.5 \Rightarrow 0$$

$$0.57 > 0.5 \Rightarrow 1 \text{ spam}$$