- 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

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

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

$$A \perp B$$

$$\| \quad \| \quad \| \quad \|$$

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

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

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

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

$$= \frac{2}{10} + \frac{3}{10} = \frac{5}{10} = \frac{1}{2}$$

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

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

Conditional Prob

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

Likelihood · Prior

$$= \frac{P(B|A)\,P(A)}{P(B)} \in \text{Marshall}$$

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

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

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

$$0.2 \quad 1/5 = 0.2$$

$$3/10 = 0.3$$

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

Posterior $= 0.133$

$A \mid B$

Text-as-data

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

- N=100 emails

Spam Data

Email	Spam	Text
1	1	"You win.."
	0	
!	0	
:		
	1	
	0	
100	1	

"You win 1 million dollars..."

1. Pre-processing (words tolli

2. Document term matrix

Pre-processing

① Tokenization - split text into terms

n-grams

(unigram) 1 - gram = word

2 - gram = phrase

$n \geq 2$ = phrase

"I went to New York"

n=1 I went to New York

n=2 { _____ _____ _____ _____ .

"You win $1 million dollars."

you win million dollars you win

win million million dollars

→ stemming ⇒ millions → million
millions → million

→ standardize text
- lowercase
- remove #s, punctuation

going
go
gone
→ go

→ stop-word removal → "the", "a", "for"

(1) Preprocessing

- → tokenization
- → stemming
- → standardize text
- → stop word removal

(2) Turning text into a document term matrix

DTM

$N = 100$, each 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 minim dollars →

	spam	you	win	minim	dollars	you wh
①	1	1	1	1	1	1
②	0	0	0	0	0	0

② ↘ less hang out tomorrow

Naive Bayes for spam detection

Target/Class { 1 Spam
0 Not spam

Features { Terms / words

$S \equiv$ Spam Class, Words $\equiv W$

$$P(S=1 \mid W) = \frac{P(W \mid S=1)\,P(S=1)}{P(W) \in}$$

(A) (B)

$W \in \mathbb{R}^{N \times 1}$

$W_1 = \{\text{you}, \text{win}, \text{million}, \text{dollars}\}$

$P(w_1, w_2, w_3 \mid S=1) = P(w_1 \mid S=1)\,P(w_2 \mid$ _

$$P(w \mid S=1) = \prod_{i=1}^{w_n} P(w_n \mid S=1)$$

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

$$= P(\text{you} \mid S=1) \cdot P(\text{win} \mid S=1) \cdot P(\text{million} \mid S=1) \cdot P(\text{dollars} \mid S=1)$$

$$= \frac{0.01}{1} \cdot \frac{0.02}{2} \cdot \frac{0.10}{5} \cdot \frac{0.20}{4}$$

$$\frac{\#\text{ of times "you" appears in spam docs}}{\#\text{ of words in spam docs}}$$

$P(w_j|s=1) = 0.000004$

$j = 1$ Likelihood

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

$= \%$ of spam docs

$= 0.50$

Marginal Likelihood

$P(w_1) = P(w_1|s=1)P(s=1) + P(w_1|s=0)P(s=0)$

0.000004

0.000003

$$P(S=1 \mid w_1) = \frac{P(w_1 \mid S=1)\, P(S=1)}{P(w_1 \mid S=1)\, P(S=1) + P(w_1 \mid S=0)\, P(S=0)}$$

$$= \frac{0.00004 \times 0.00003}{0.00004 \times 0.00003 + 0.00005 \times 0.50}$$

$$= 0.57$$

$P(S=1 \mid w_1) = 0.57$

$P(S=1 \mid w_1) > 0.5 \Rightarrow 1$

$< 0.5 \Rightarrow 0$

$0.57 > 0.5 \Rightarrow$ Spam