W207.2 Notes 05/22/18

— Review of prob.

— Text-as-data

— Naive Bayes + application

Probability

A - It will rain today
B - My dog will bark today

→ frequentist

→ Bayesian

A - RnM $P(A) = \frac{800}{1000} = \frac{4}{5}$

$P(B) = \frac{980}{1000} = \frac{98}{100} \approx 98\%$

Bayesian - when we calculate probabilities we need to take into account beliefs about a probability distribution

$A \perp B$ A - Rain B - dog bark

$P(A) = 4/5$, $P(B) = 9/10$

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

$= \binom{4}{5}\left(\frac{9}{10}\right) = \frac{36}{50}$

$P(A \text{ or } B) = P(A \cup B) \approx 72\%$

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

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

$$P(B^c) = 1 - P(B) = 1 - \frac{9}{10} = \frac{1}{10}$$

Bayes Rule

conditional probability

likelihood

$$\underbrace{P(A|B)}_{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{likelihood}} \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{marginal likelihood}}}$$

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

$$P(B) = \underbrace{P(B|A)}\underbrace{P(A)} + \underbrace{P(B|A^c)}\underbrace{P(A^c)}$$

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

$P(A)$ = 1 − 4/5 = 1/5

$$P(\text{Rain} \mid \text{My dog barked today}) = \frac{(1/2)(1/5)}{(9/10)} = \frac{(1/10)}{9/10} = \frac{1}{9} \approx 4\%$$

Assn. for Bayes Thm $A \mid B$

$$P(A) = 9/10$$

$$P(A \mid B) = \frac{(1/2)(9/10)}{(9/10)} = 50\%$$

.

Text-as-Data

$N = 100$ email (documents)

(Corpus)

Spam { 1 email is spam
0 " is not spam

Target

class label

(Features)

Text - words/terms

= "You have win one million dollars give bank info"

Spam

1

0

0

1

.

.

.

1

Text

"you have win ..."

Span = f(Text)

$\widehat{Spam} = f(Text)$

Text to Data

① Preprocessing

→ tokenization

→ cleaning

- o stemming
- o punctuation removal
- o stop word removal

NLP

② Building a DTM (Document term matrix)

..

Pre-processing

① Tokenization

"You have won one million."

a) partition text into terms

n-grams unigrams (words)

2-grams (phrase)

n≥2 - phrase

$W_1 = \{$"you", "have", "won", "one", "million", "you have" ... $\}$

$w_1 \in \mathbb{R}^9$ ⊗ Clean texts

— Stemming - groups → group

groups
grouped
won
wins

- ing
- ed
- s

→ win

→ remove punctuation, remove its lowercase

→ Stop word removal — removal of common and frequent terms

$w_1 = \{$ "you", "have", "with" ... $\}$

(2) DTM

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

$N \equiv$ obs / documents

$W \equiv$ terms in all N documents

$N = 100$ docs / email

$W = 10{,}000$ terms $X \in \mathbb{R}^{100 \times 10000}$

Term 1 . . . . — . Term 10,000

1
2
⋮
.
.
.

2 1 0 0 0 0 0 0 . .

load. 2 DTM element/entries

① Term frequency

$w_1 = \{$ "you", "have", "with" ... $\}$

Doc	Span	you	have	with	...		
1	1	1	1	1		0	0
2	0	0	0	0			

② TF-IDF – Term-frequency

Inverse Document frequency

you have wm

1 (1/80) 1 (1/50) 1 (1/5)

$$IDF_i = \frac{1}{\#\text{ of documents word/term } i \text{ appears in}}$$

$$IDF(\text{you}) = \frac{1}{80} \quad IDF(\text{have}) = \frac{1}{50}$$

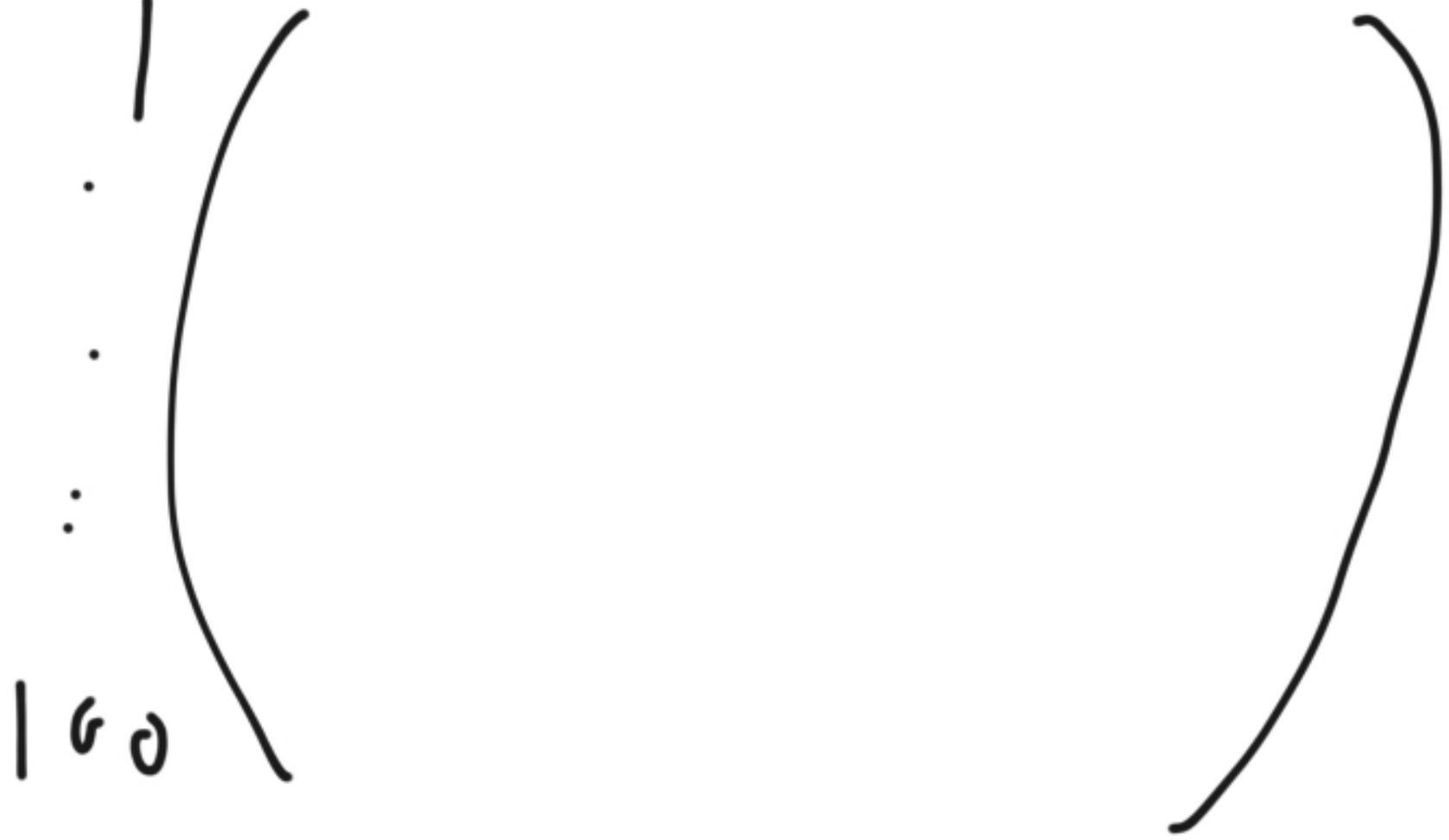
$$IDF(wm) = 1/5$$

$X \equiv DTM$ or Feature matrix

$N \times W$

$N \approx 100$!

You have NM . . .



$|60$

Target = spam $\in \mathbb{R}^{100}$

$X \in \mathbb{R}^{100 \times 10000}$

Naive Bayes

$N = 100$

1. Pre-processing / DTM
2. Divide data into training and test 80% / 10%
3. Train algorithm

④ Naive performance

50 training emails

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

likelihood

$$P(S=1|w_i) = \frac{P(w_i|S=1)\,P(S=1)}{P(w_i)}$$

$w_1 = \{\text{"you"}, \text{"have"}, \text{"win"}\}$

$P(w_1 \mid S=1) = P(\{\text{"you"}, \text{"have"}, \text{"win"}\} \mid S=1)$ $1/30$

$\Rightarrow P(\text{"you"} \mid S=1)\, P(\text{"have"} \mid S=1)\, P(\text{"win"} \mid S=1) = 15/30.$

80 → 50 not spam

→ 30 spam

$P(\text{"you"} \mid S=1) = \frac{5}{30}$

$$P(w_1 \mid s=1) = \left(\frac{6}{30}\right)\left(\frac{1}{30}\right)\left(\frac{15}{30}\right)$$

$$P(s=1) = \frac{30}{50} = \frac{3}{5} \qquad P(s=0) = \frac{20}{50}$$

$$P(w_1) = P(w_1 \mid s=1)P(s=1) +$$

$$P(w_1 \mid s=0)P(s=0)$$

$$P(S=1 \mid w_1) = \frac{\binom{75}{23000}\binom{3}{5}}{\binom{75}{23000}\binom{3}{5} + \binom{20}{23000}\binom{2}{5}}$$

$$=$$