Probability

- study of events
- frequentist - concern with events as frequencies
- Bayesian - prior beliefs/info.

A - Event that I will run today

B - Event that my cat will meow today

$$P(A) = \frac{800}{1000} = \frac{8}{10} = \boxed{\frac{4}{5}}$$

$$P(B) = \frac{900}{1000} = \boxed{\frac{9}{10}}$$

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

$$A \perp B \qquad = \left(\frac{4}{5}\right)\left(\frac{9}{10}\right) = \frac{36}{50}$$

$$= 72\%$$

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

$$- P(A \cap B)$$

$$= \frac{8}{10} + \frac{9}{10} - \frac{36}{50} = \boxed{98\%}$$

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

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

Bayes Theorem

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

Likelihood: $P(B|A) = \frac{100}{200} = \boxed{0.5}$

Prior: $P(A) = \frac{800}{1000} = \boxed{\frac{4}{5} \text{ (empirical prior)}}$

$\frac{9}{10}$ (subjective prior) ✓

$\frac{1}{2}$ (flat prior)

Marginal Likelihood: $\boxed{P(B) = \frac{9}{10}}$

Marginal Likelihood =

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

$$P(A|B) = \frac{\left(\frac{1}{2}\right)\left(\frac{4}{5}\right)}{9/10} = \frac{4/10}{9/10}$$

$$= \frac{40}{90} = \boxed{42\%}$$

empirical

prior

$$P(A|B) = \frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}{9/10} = \frac{1/4}{9/10}$$

$$= \frac{10}{36} \approx \boxed{28\%}$$

flat prior

subjective

$$\Rightarrow \frac{\left(\frac{1}{2}\right)\left(\frac{9}{10}\right)}{9/10} = 50\%$$

Text-as-Data

$\underbrace{N = 100 \text{ emails}}_{\text{corpus}}$ → documents

Spam (target) $\begin{cases} 1 & \text{email is spam} \\ 0 & \text{email '' not spam} \end{cases}$

Text (features) → Text of emails

2 steps to turn text into data

① Pre-processing

NLP

a) Tokenization

b) Cleaning

- → Stemming
- → lowercase, remove #'s, punctuat
- → Stop word removal

② Document-term matrix
(DTM)

Tokenization

Spam	Text
1	"You win million dollars give bank info to claim prize"

n-grams → Unigrams (words)

n-grams → 2-grams (phrases)

n>1 => phrase

$w_1 = \{$ You, win, million, dollars_ ,
You win, win million, million dollars

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

→ stemming - removing prefixes & suffixes and replacing words with thier root

dollars → dollar

groups → group

grouping →

grouped →

→ stop word removal → the → for

→ you

→ a

$w_1 = \{$wm, million, dollar, ...$\}$

Document term matrix

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

$N \equiv$ # of docs in a corpus

$W \equiv$ # of terms in a corpus

spam

$$\begin{pmatrix}1\\0\\1\\0\\\vdots\\1\\\vdots\\0\end{pmatrix}
\quad
\begin{array}{c|c}
 & T_1 \quad T_2 \quad T_3 \quad \cdots \quad T_{10000} \\
\hline
1 & 2 \\
\vdots & \\
100 &
\end{array}$$

$N \times 1$

① Text Frequency (TF)

② TF-IDF

$N \times W$

$W = 10000$

$\boxed{\text{TF?}}$ → # of terms in a doc

w_1 = { win, million, dollar . . }

win million dollar today yay . . .

$$1 \left(\tfrac{1}{50} \downarrow \ \tfrac{1}{2} \checkmark \ \tfrac{1}{4} \checkmark \quad 0 \quad 0 \right.$$

TF-IDF   IDF(win) $= \dfrac{1}{50}$

IDF(Dollar) $= \dfrac{1}{4}$ IDF(million) $= \dfrac{1}{2}$

Naive Bayes

N=100 → 20 test

→ 80 Train

$$P(S=1 \mid w_i) = \frac{\overbrace{P(w_i \mid S=1)}^{\text{Likelihood}}\, P(S=1)}{P(w_i)}$$

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

$$= P(\underline{\text{win} \mid S=1})\, P(\underline{\text{million}} \mid S=1)\, P(\text{dollar} \mid \text{spam})$$

80 → 30 spam

→ 50 no spam

1000 words "win" appears 300

$$P(win \mid spam) = \frac{300}{1000}$$

$$P(S=1) = \frac{30}{80} = \frac{3}{8}$$