M

E

$$p(M|E) = ?$$

$$p(M|E) = \frac{p(E|M) \cdot p(M)}{p(E)}$$

Bayes' Classifier

$H = S^c$

S: event that email is spam

V: event that email contains "viagra"

$$p(S|V) = \frac{p(V|S) \cdot p(S)}{p(V)}$$

$$= \frac{p(V|S) \cdot p(S)}{p(V|S) \cdot p(S) + p(V|H) \cdot p(H)}$$

Words in email

$x_1, x_2, \ldots, x_n$

$$p(S|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|S) \cdot p(S)}{p(x_1, \ldots, x_n|S)\, p(S) + p(x_1 \ldots x_n|H) \cdot p(H)}$$

Assumption

Words are independent!

$$p(x_1, \ldots, x_n|S) \cdot p(S)$$

$$= p(S) \cdot p(x_1|S) \cdot p(x_2|S) \cdots p(x_n|S)$$

$$\frac{p(S) \cdot p(x_1|S) \cdot \ldots p(x_n|S)}{p(S) \cdot p(x_1|S) \cdots p(x_n|S) + p(x_1|H) \cdots p(x_n|H) \cdot p(H)}$$

Words in email

$x_1, x_2, \ldots, x_n$

$$p(S \mid x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \mid S) \cdot p(S)}{p(x_1, \ldots, x_n \mid S)\, p(S) + p(x_1 \ldots x_n \mid H) \cdot p(H)}$$

Assumption

Words are independent!

$$p(x_1, \ldots, x_n \mid S) \cdot p(S)$$

$$= p(S) \cdot p(x_1 \mid S) \cdot p(x_2 \mid S) \cdots p(x_n \mid S)$$

$$\frac{p(S) \cdot p(x_1 \mid S) \cdots p(x_n \mid S)}{p(S) \cdot p(x_1 \mid S) \cdots p(x_n \mid S) + p(x_1 \mid H) \cdots p(x_n \mid H) \cdot p(H)}$$

Smoothing

Goal: For all words x

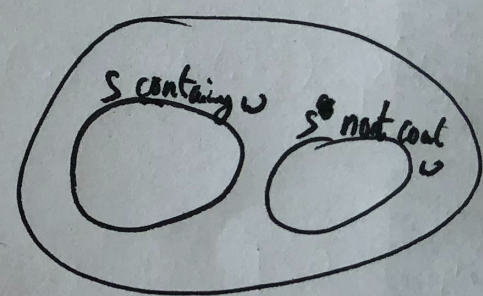
$$p(x \mid S) > 0, \quad p(x \mid H)$$

To estimate

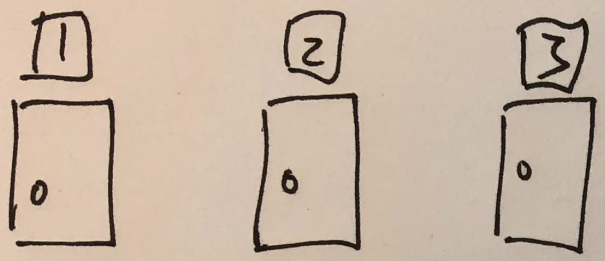
$$p(w \mid S) : \frac{\#\text{ spam emails containing } w}{\#\text{ spam emails}}$$

Smooth

$$\frac{\#\text{spam emails with } w + 1}{\#\text{spam emails} + 2}$$



Monty Hall Gameshow



2 doors → goat

1 door → car

1. Contestant picks a door $\boxed{1}$

$\boxed{3}$ 2. Host picks a random goat-door among two not picked by contestant

3. Contestant stay or switch.

C_1, C_2, C_3 : events that car is behind $\boxed{1}, \boxed{2}, \boxed{3}$

H_2 : host picks door $\boxed{2}$

$H_3 = H_2^c$: host pick door $\boxed{3}$

H_3 happened.

$p(C_1 | H_3)$ vs $p(C_2 | H_3)$

$$p(C_2 | H_3) = \frac{p(H_3 | C_2) \cdot p(C_2)}{p(H_3)}$$

$p(H_3 | C_2) = 1$ $\quad p(C_2) = 1/3$

$p(H_3) = 1/2$ $\quad p(H_2) + p(H_3) = 1$

$$p(H_3) = p(H_3 | C_1) \cdot p(C_1) + p(H_3 | C_1^c) \cdot p(C_1^c).$$

C_1, C_2, C_3 : events that car is behind $\boxed{1}, \boxed{2}, \boxed{3}$

H_2 : host picks door $\boxed{2}$

$H_3 = H_2^c$: host pick door $\boxed{3}$

H_3 happened.

$p(C_1 | H_3)$ vs $p(C_2 | H_3)$

$$p(C_2 | H_3) = \frac{p(H_3 | C_2) \cdot p(C_2)}{p(H_3)}$$

$p(H_3 | C_2) = 1$ $\quad p(C_2) = 1/3$

$p(H_3) = 1/2$ $\quad p(H_2) + p(H_3) = 1$

$p(H_3) = p(H_3 | C_1) \cdot p(C_1) + p(H_3 | C_1^c) \cdot p(C_1^c)$.

$$p(C_2 | H_3) = \frac{1 \cdot (1/3)}{1/2}$$

$$= 2/3.$$

$p(C_1 | H_3) + p(C_2 | H_3) = 1$

$\Rightarrow p(C_1 | H_3) = 1/3.$

Study 1:

Acceptance rates are same.

Study 2:

In every major, acceptance rate for women was lower than for men.

W : event applicant is woman

M : " " " man

A : " " is admitted

Study 1: $p(A|W) = p(A|M)$

2 majors

X : event applicant in Computer science

Y : " app in statistics

Study 2:

$$p(A|W,X) < p(A|M,X)$$

$$p(A|W,Y) < p(A|M,Y)$$