lottery

w_1	w_2	w_3	0
w_2	w_3	w_4	1
w_4	w_1	w_5	1
w_2	w_9	w_3	1
w_1	w_3	w_4	0

0 → spam

1 → Non spam

$[w_2 \; w_9 \; w_4]$

Training

w_1 = lottery

$P(w_1 / y=1)$

$P(\text{lottery} / y=1) = 1/3$

$P(\text{lottery} / y=0) = 2/2 = 1$

$\boxed{P(\text{lottery}) = 3/5}$ X

$P(w_1 / y=1)$

$P(w_2 / y=1)$

$P(w_3 / y=1)$

$P(w_4 / y=1)$

$P(w_9 / y=1)$

$P(w_1 / y=0)$

$P(w_2 / y=0)$

$P(w_3 / y=0)$

$P(w_4 / y=0)$

$P(y=1)$ $P(y=0)$

$x_q = [\; w_2 \quad w_3 \quad w' \;]$ new

$$P(y=1 \mid x_q) = \frac{P(w_2 \cap w_3 \cap w' / y=1) \cdot P(y=1)}{P(w_2 \cap w_3 \cap w')} \rightarrow ①$$

Compare

$$P(y=0 \mid x_q) = \frac{P(w_2 \cap w_3 \cap w' / y=0) \cdot P(y=0)}{P(w_2 \cap w_3 \cap w')} - ②$$

ignore

(1) $P(y=1)\,P(w_2 \cap w_3 \cap w' / y=1) = P(y=1) \cdot P(w_2 / y=1)\, P(w_3 / y=1) \cdot P(w' / y=1)$

A_1

(2) $P(y=0)\,P(w_2 \cap w_3 \cap w' / y=0) = P(y=0)\, P(w_2 / y=0) \cdot P(w_3 / y=0)\, P(w' / y=0)$

A_2

A_1 v/s A_2

Laplace Smoothing

Data Imbalance

Likelihood

1000

$P(y=1)$ $\prod_{i=1}^{d} P(w_i / y=1)$ nonspam

0.7   0.1

$P(y=0)$ $\prod_{i=1}^{d} P(w_i / y=0)$ / spam

0.3   0.2

$P(y=1/\text{text}) = 0.07$

$P(y=0/\text{text}) = 0.06$

nonspam

1   700

0   300

spam.

text: w_1 . w_d

w_1 lottery w_2 Prince

w_3 Money w_4 transfer

$P(w_1 \mid y=0)$ 0.2

$P(w_2 \mid y=0)$ 0.2

$P(w_3 \mid y=0)$ 0.1

$P(w_4 \mid y=0)$ 0.1

$\prod_{i=1}^{4} P(w_i \mid y=0)$

$\Rightarrow 0.0004$

$P(w_1 \cap w_2\, w_3 \cap w_4 \mid y=0) = 0.0004$

$P(w_1 \mid y=1)$ 0.1

$P(w_2 \mid y=1)$ 0.1

$P(w_3 \mid y=1)$ 0.1

$P(w_3 \mid y=1)$ 0.1

$\prod_{i=1}^{4} P(w_i \mid y=1)$

$\Rightarrow 0.0001$

$P(w_1 \cap w_2 \cap w_3 \cap w_4 \mid y=1) = 0.0001$

0 → Spam

1 → Nonspam

$w_1 \cap w_2 \cap w_3 \; w_4 \longrightarrow$

priors

$P(y=0)$

$P(y=0 \mid w_1 \cap w_2 \cap w_3 \cap w_4) = 0.0002 \times 0.3 = 0.00006$

$P(y=1 \mid w_1 \cap w_2 \cap w_3 \cap w_4) = 0.0001 \times 0.7$

$P(y=1)$

0.00007

$0.0002 \times 0.5 = 0.0001$

$0.0001 \times 0.5 = 0.00005$

Rebalance $\longrightarrow$ upsampling

Will Upsampling impact the likelihoods to change?..

$P(\,|\,y=1)$ likelihood $y=1$

$P(\,|\,y=0)$ likelihood $y=0$

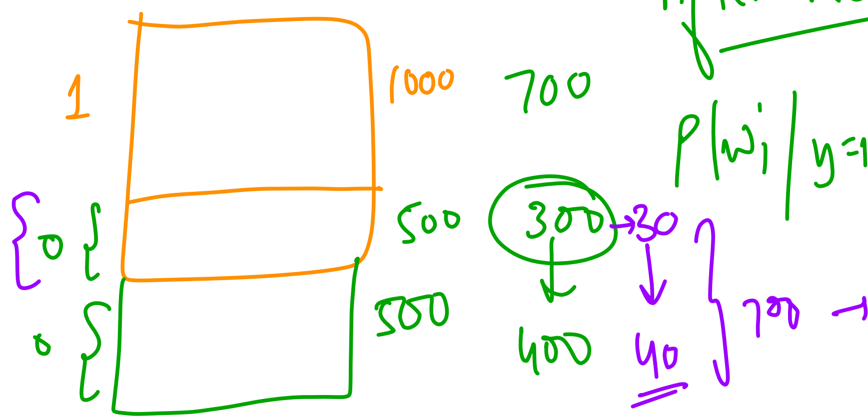
before Rebalancing

$$P(w_j \mid y=0) = \frac{50+\alpha}{500+\alpha C}$$

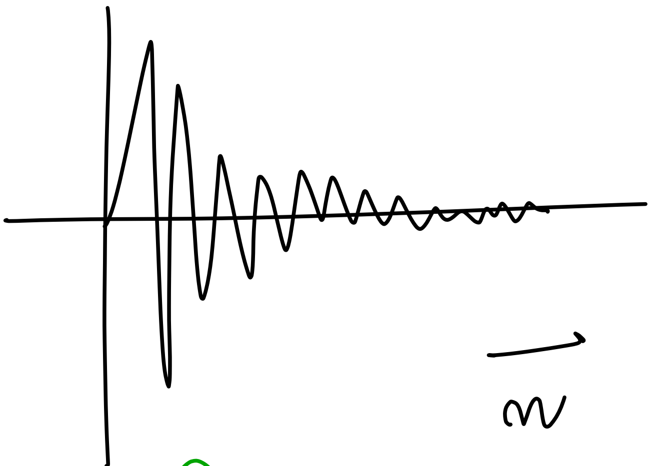
After Rebalancing

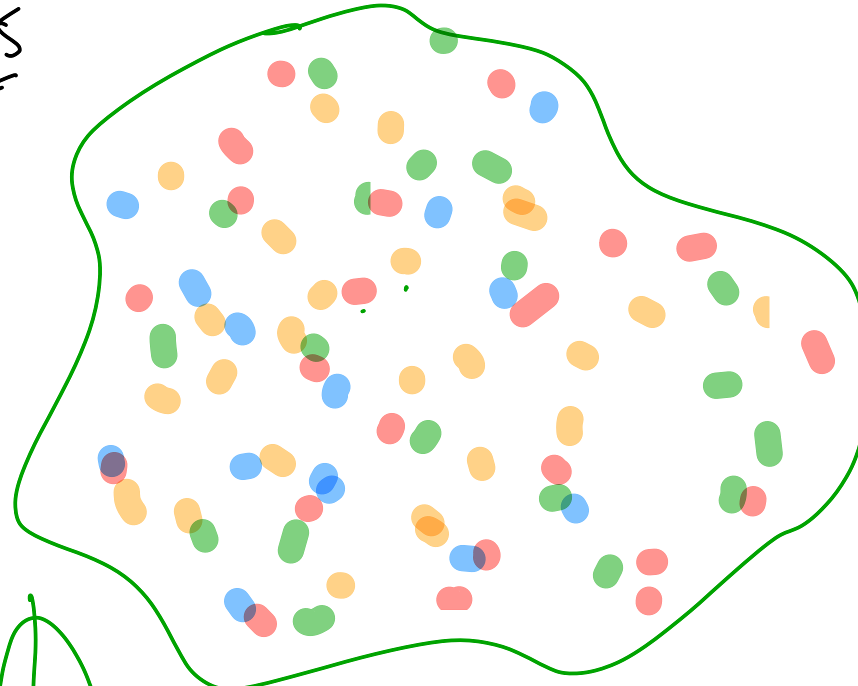
$$P(w_j \mid y=1) = \frac{100+\alpha}{1000+\alpha C}$$

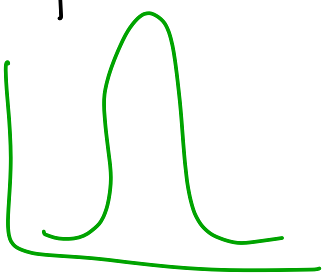
$w_j = 50$

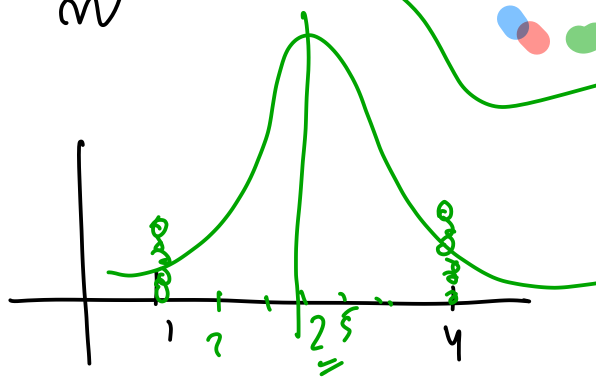


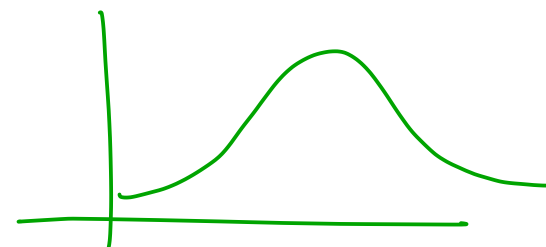
1 , 2 , 3 , 4 $= \underline{\underline{2.5}}$











1000 Unique word $\rightarrow$ x_q $\rightarrow$ Spam
$\rightarrow$ Nonspam

100000 Unique word $\rightarrow$ Training data

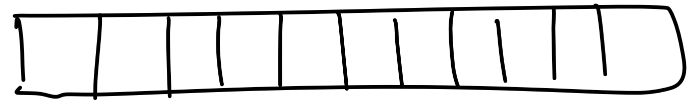
$$P(w_j / y=1) = [0,1]$$

$$y=0 = [0,1]$$

$$\text{prior}_{y=1} \times \prod_{i=1}^{1000} \text{Liklihood}$$

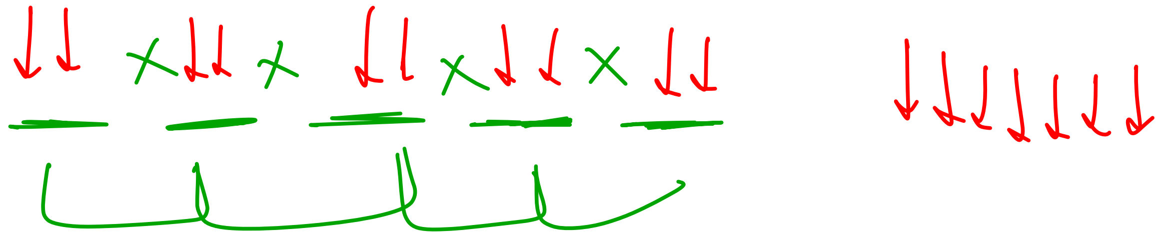
$$\text{prior}_{y=0} \times \prod_{i=1}^{1000} \text{Liklihood}$$

d+1 $\Rightarrow$ Very Very Very small

float64 →  Max
Min

Underflowing → NAIVE BAYES

→ Overflowing



$$\log\left(P(y=1 \mid w_1 \cap w_2 \cap w_3)\right) = \log\left(P(y=1) \cdot P(w_1 \mid y=1)\, P(w_2 \mid y=1)\, P(w_3 \mid y=1)\right)$$

$$\log\left(P(y=1 \mid w_1 \cap w_2 \cap w_3)\right) = \log\left[P(y=1) \prod_{i=1}^{3} P(w_i \mid y=1)\right]$$

$$\log(ab) = \log a + \log b$$

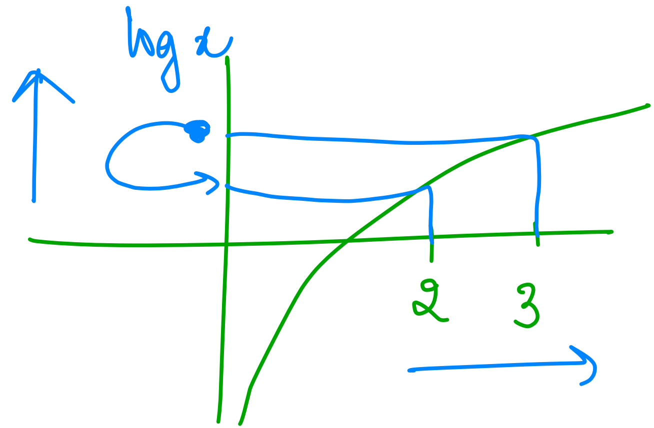
$$\log(P(y=1 \mid w_1 \cap w_2 \cap w_3)) = \log(P(y=1)) + \log(P(w_1 \mid y=1)) + \log(P(w_2 \mid y=1)) + \log(P(w_3 \mid y=1))$$

1000 Unique

$$\log(P(y=1 \mid \text{text}))$$

Compare

$$\log(P(y=0 \mid \text{text}))$$



Feature Importance & Interpretability

w_1 w_2 w_3

words

Class priors }.

Top likelihood.

{Likelihood} w & Class

Outliers →

1. New word → outlier Test

2. Rare word → Training time.

Flabbergasted

$P(w_{rare} | y=1)$ = Very Very

$y=0$ = 0.0001

10000 ⇒ 1

How to solve?

Laplace Smoothing

$$\frac{0+\alpha}{10000+\alpha C} \qquad \frac{1+\alpha}{1000+\alpha C}$$

How do we identify?

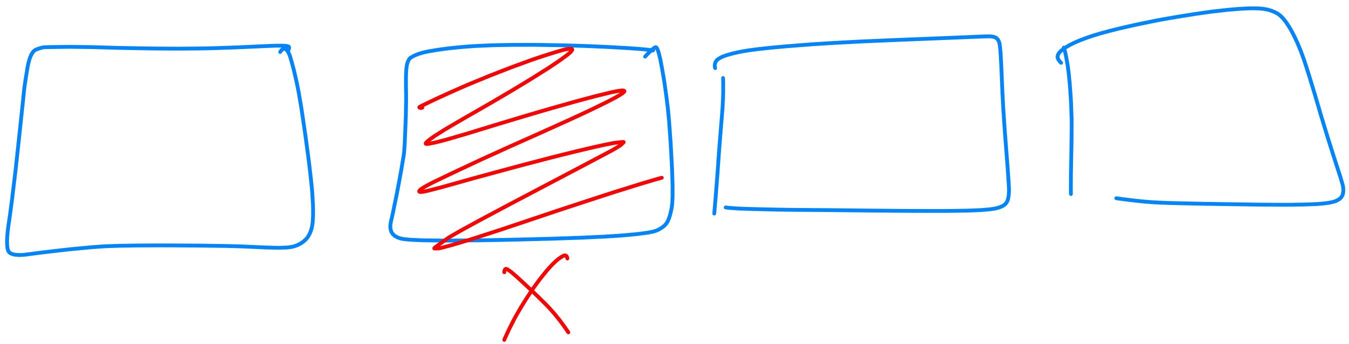
$$P(w_{rare} / y = 0/1) = \downarrow\downarrow\downarrow \quad 0.0001$$

Threshold $\rightarrow$ less than τ (10)

$\times$ Not consider

$$\alpha \rightarrow \frac{0+1}{10000 + 1 \times 2} = \frac{1}{10000 + 2} \quad \frac{1}{10002} =$$

$$\frac{0 + 500}{1000 + 500 \times 2} = \frac{500}{11000} = 0.05$$



Multinomial Naive Bayes

Unique words → 0 → Present or Not present
→ 1 → Present or Not present.

App

↳ ~~freq~~ → words

0
$w_1 \to 1$
$w_2 - 1$
$w_3 - 5$

$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & w_4 & w_5 \\ \hline 0\ T_1 & 1 & 0 & 0 & \cancel{0} & 1 \\ 0\ T_2 & 1 & 1 & 0 & 0 & 1 \\ 1\ T_3 & 0 & 0 & 0 & 1 & 1 \end{array}$$

absent $\leftarrow$ (0)   present $\leftarrow$ (1)

$$P(w_3 \mid y=0) = 0 \qquad P(w_2 \mid y=0) = 1/2$$

$$P(w_3 = 1 \mid y = 0) = 0 \qquad P(w_2 = 1 \mid y = 0) = 1/2$$

		w_1	w_2	w_3	w_4	w_5
0	T_1	3	2	4	1	2
0	T_2	1	2	3	4	1
1	T_3	2	2	2	3	3

freq

$$P(w_3 = 3 / y = 0) = \frac{1}{2}$$

$P(w_j \mid y=1)$ → Bernoulli Naive Bayes

→ Multinomial Naive Bayes.

Storing more information.

10000 unique

$x_q = 10$

[1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1 1 0 1]

Sparse vector

9990 ⇒ 0's

$$\begin{bmatrix} 1 & 1 & 1 & 1 & \boxed{1} & 1 & \boxed{1} & 1 & \boxed{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & & 1 & & 1 & & 1 & & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\underline{\underline{10}}$ $\underbrace{\{ w_i : 1 \; , \; w_{10} : 1 \; , \; w_{100} : 1 \}}$

$\underline{\underline{10}}$ $[\, 2 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 3 \; \ldots \ldots \; 5 \,]$

$\{ w_1 = 2 \; , \; w_{10} = 1 \; , \; w_{15} = 3 \; , \; w_{100} = 5 \}$

Cascading → Banks hospital

(1) Fraudlent : 1%

(0) Non fraudulent : 99%

$\mathcal{D} \rightarrow \boxed{m1}$

$P(T=1)$

$P(T=0)$

$\mathcal{D} - \mathcal{D}_1$

$P(T=1) \ll 0.1$

$\boxed{m2}$

$P(T=1)$

$P(T=0)$

$P(T=1) \ll 0.01$

$\mathcal{D} - \mathcal{D}_1 - \mathcal{D}_2$

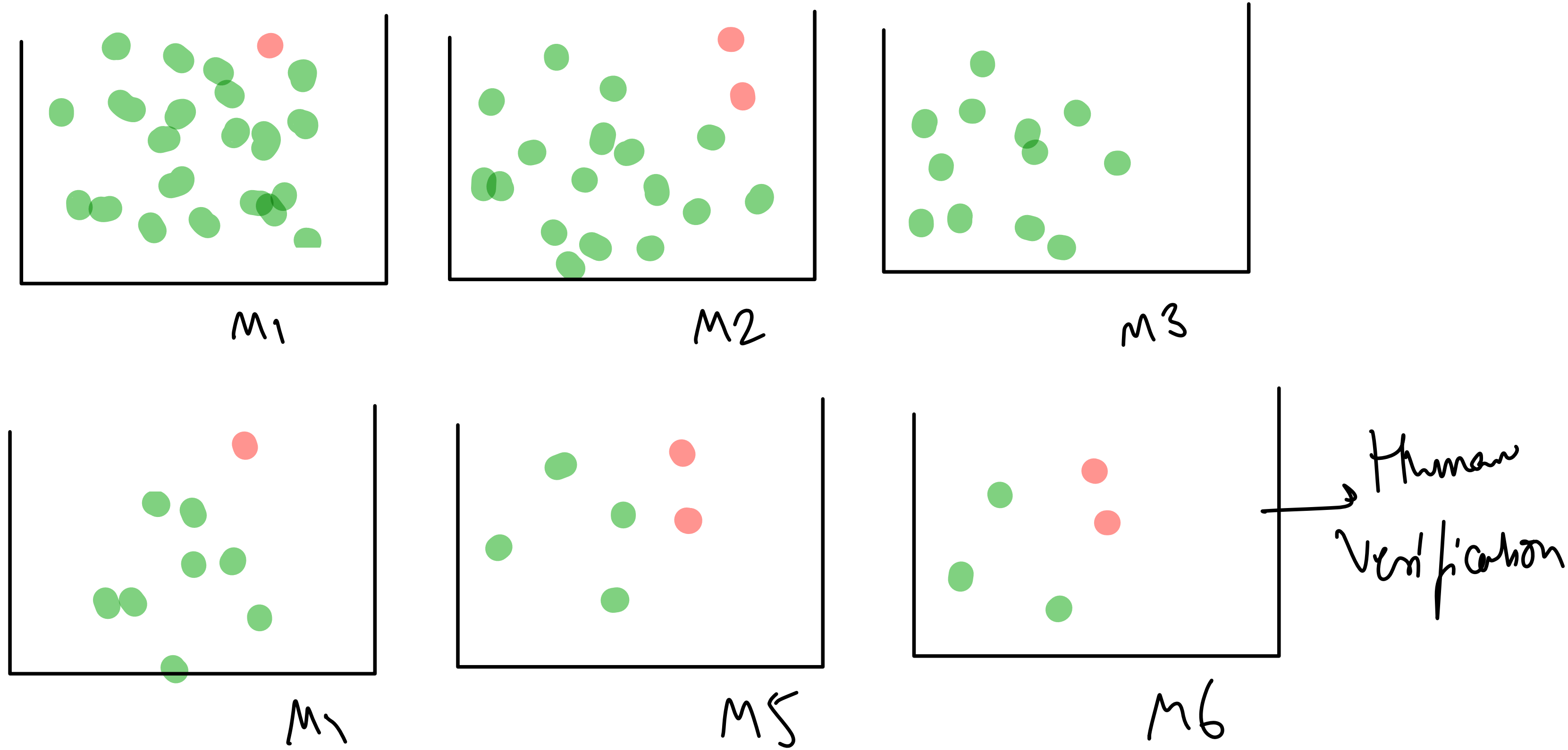
$\boxed{m_3}$

$P(T=1) < 0.001$

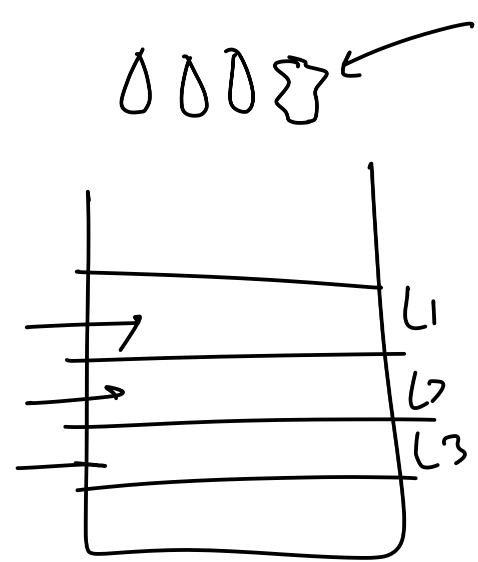
$\mathcal{D} - \mathcal{D}_1 - \mathcal{D}_2 - \mathcal{D}_3$

$m_4 \rightarrow$

Human Verification


M1
M2
M3
M4
M5
M6
Human
Verification

Non Scheduler



$T_2 = 1\%$

$T_4 = 9\%$

$T_3 = 10\%$

$T_1 = 22\%$

T_6 25%

T_7 75%

T_8 90%

$T_5 = 50\%$

$T_{10} = 70\%$

$T_9 = 80\%$

$\hat{y} \rightarrow$ error

$\hat{y}_i$

spam

120 → 8 | 0 → 4 → — — — -

1

0

1

6

1

0

4

1

0

1

0

MSE

logloss