# LOGIC BEHIND TYPES OF NB

## ① Multinoulli / Categorical NB

→ used when all input features are categorical or (can say) follow multinoulli distribution

Eg Data →

| Gender | Pclass | Survived |
|---|---|---|
| M | P1 | 1 |
| F | P2 | 0 |
| F | P3 | 0 |

• even though Gender is Bernoulli but as we are applying Categorical NB, we are assumming it to be categorical / multinoulli distribution

• Acc to Naive Bayes:

new query = { M, P3 } → Predict → 0, 1

$P(1 | M, P3)$

$= P(1) \times P(M|1) \times P(P3|1)$

$\Rightarrow P(1) = \frac{392}{891}$

$\Rightarrow P(M|1) = \frac{2}{3}$

$\Rightarrow P(P3|1) = \frac{1}{5}$

$\frac{392}{891} \times \frac{2}{3} \times \frac{1}{5} = 0.6$

$P(0 | M, P3)$

$= P(0) \times P(M|0) \times P(P3|0)$

$\Rightarrow P(0) = \frac{496}{891}$

$\Rightarrow P(M|0) = \frac{4}{15}$

$\Rightarrow P(P3|0) = \frac{5}{8}$

$\Rightarrow \frac{496}{891} \times \frac{4}{15} \times \frac{5}{8} = 0.9$

- Since $P(0 \mid m, P_3)$ is highest, 0 will be prediction i.e person will not survive
- that's how categorical / multinoulli works
- and LAS will be applied as same
→ ie $\frac{\;+\alpha}{\;+n\alpha}$, Generally $\alpha = 1$, $n$ is no of categories in column.

## ② Bernoulli Naive Bayes

- works well on data where each feature follows Bernoulli distribution ie binary features (2 categories)

- 

| $f_1$ | $f_2$ |
|---|---|
| 1 | 0 |
| 0 | 0 |
| 0 | 1 |
| 1 | 0 |

- where can we find this type of data?
- → Using BOW (binary)
→ In this, it only tells whether the word is there or not, instead of count
- Ex →

# Binary BOW

| Ex | Chinese | Beijing | Shanghai | Macao | Tokyo | Japan | |
|---|---|---|---|---|---|---|---|
| d1 | 1 | 1 | 0 | 0 | 0 | 0 | Y |
| d2 | 1 | 0 | 1 | 0 | 0 | 0 | Y |
| d3 | 1 | 0 | 0 | 1 | 0 | 0 | Y |
| d4 | 1 | 0 | 0 | 0 | 1 | 1 | N |
| d5 | 1 | 0 | 0 | 0 | 1 | 1 | ? |

- All features follows Bernoulli distribution
- hence we will apply Bernoulli NB
- New query → d5 → prediction Y or No?

d5

$P(Y \mid Chinese=1, bei=0, shan=0, mac=0, tok=1, Jap=1)$

$= P(Y) \times P(Chinese=1 \mid Y) \times P(bei=0 \mid Y) \times P(Shan=0 \mid Y) \times P(mac=0 \mid Y) \times P(Tok=1 \mid Y) \times P(Jap=1 \mid Y)$

$\Rightarrow P(Chinese=1 \mid Y) = \frac{3}{3} = 1$

$\Rightarrow P(bei=0 \mid Y) = \frac{2}{3}$

$\Rightarrow P(Shanghai=0 \mid Y) = \frac{2}{3}$

$\Rightarrow P(Mac=0 \mid Y) = \frac{2}{3}$

$\Rightarrow P(Tok \mid Y) = P(Jap \mid Y) = \frac{0}{3} = 0$

$\Rightarrow P(Y) = \frac{3}{4}$

$P(N \mid chinese=1, bei=0, chan=0, mac=0, tok=1, Jap=1)$

$P(N) = \frac{1}{4}$

$P(Chinese=1 \mid N) = \frac{1}{1} = 1$

$P(bei=0 \mid N) = 1$

$P(sh=0 \mid N) = 1$

$P(Tok=1 \mid N) = 1$

$P(Jap=1 \mid N) = 1$

⇒ Behind you are using Bernoulli dist

- $P(X=k) = p^k + (1-p)(1-k)$
- where $p$ is probability of success i.e. (1)

Or • k value = {0, 1}

Ex $P(Chinese=1 \mid Y) \Rightarrow$

- $P(X=1) = p^1 \times (1-p)^{(1-k)}$
  $= p^1 \times 1$

$P(Chinese=1 \mid Y) = \frac{3}{3}$

$1 \times \frac{2}{3} \times \frac{2}{3} \times \cancel{1}$

- That's how Bernoulli is used.

Q How Bernoulli is different from categorical?

Ans ~~Because in Bernoulli, you find prob of ~~ ~~No~~ ~~also~~

Bernoulli

- Because in ~~prob~~ Bernoulli, you not only find the prob of word present but also prob of word not present

- In Bernoulli ⇒ $P(Y|X)$ indicates word is present
  ⇒ $P(N|X_i)$ " " is not "

Categorical

- In categorical, you only account ~~of~~ for prob of word present

Ex Classification → Dog, Cat, horse

- No of prob will be $P(Dog|X)$, $P(Cat|X)$, $P(horse|~~Dog~~X)$
- It is not accounting for probability of not a dog, prob of not cat, or horse.

LAS

- In this, you can also apply LAS in same way

## 3 Multinomial Naive Bayes

- Applied when all features are discrete

Ex

| | Chinese | Beijing | Shanghai | Macao | Tokyo | Jap | |
|---|---|---|---|---|---|---|---|
| d1 | 2 | 1 | 0 | 0 | 0 | 0 | Y |
| d2 | 2 | 0 | 1 | 0 | 0 | 0 | Y |
| d3 | 1 | 0 | 0 | 1 | 0 | 0 | Y |
| d4 | 1 | 0 | 0 | 0 | 1 | 1 | N |
| Query { d5 | 3 | 0 | 0 | 0 | 1 | 1 | ? |

- Predict for Q5

• Two prob will be calculated

i) P(Y | chinese = 3 | bei = 0 | sha = 0 | Jap = 0 | mac = 0 | tok = 1, Jap = 1)

= ~~Y~~

• Note

• In other types, ex Bernoulli NB, we considered all features to follow Bernoulli distribution

Imp

Q • So, similarly in multinomial, what is the thing that follows multinomial distribution?

Ans You will treat each column as a multinoulli distribution and each row as multinomial distribution

understanding by example

Eg:

• Treat every new sentence as bag which contain words ⇒ $\boxed{0\,0\,0\,0\,0}$

• and whenever new query comes, effectively you are asking tell what is the probability of 3 chinese words, 1 tokyo word, 1 Japan word

• Similar to Q4

• where we are asking P of 3 students get placed 1 student opt out and 6 not placed.

• Similarly we are asking prob of 3 Chinese words, 1 tokyo and 1 Japan out of the 5 words taken out of bag.

Can be rewritten as:

- chinese chinese chinese Tokyo Japan
- P( chinese | Y) X P( chinese | Y) X P( chinese | Y) X P( Tokyo | Y) X P( Ja | Y)

⇒ What is this P ( Chinese | Y ) ?

Ans Out of total Yes words, how many are chinese

$= \frac{5}{8} = P(\text{chinese} \mid Y)$

⇒ What is P ( Tokyo | Y ) ?

$= \frac{0}{8}$, P( Jap | Y) is also same

In short $P(Y \mid x) =$

$$= \left(\frac{5}{8}\right)^3 \left(\frac{0}{8}\right)^1 \left(\frac{0}{8}\right)^1 \times P(Y) = P(Y \mid X)$$

- There are also prob of beijing and shang but they are not present in test data ∴ their raised to power is 0 ⇒ $P(\text{bei} \mid Y)^0 = 1$. Hence no need to mention

- Refer to Q4, you will understand

Q2 In multinomial, we are also multiplying it with permutation but in this we are not. why?

Ans

Why we are not multiplying it with:

$$\frac{5!}{3!\,1!\,1!}$$ → no of total words in query

→ 3 categories ⇒ chinese = 3, tok = 1, Jap = 1

Ans : Because, we are only asking for that test combination ch ie. :-

chinese ($w_1$) chinese ($w_2$) chinese ($w_3$) Tokyo ($w_4$) Japan ($w_4$)

Note :

- Since, rows represent multinomial that's why in laplace additive smoothing ie

$$\frac{+\alpha}{+n\alpha}$$ → n is no of features not no of categories in features

(4) Out of core Naive Bayes

- can apply partial-fit
- useful when dataset is large
- divide large data into chunks
- Because text data is usually large

- Training will continue from at that chunk part
- Other algos also provide this method