* concept → Laplace Additive Smoothing

Eg: We have dataset

review			sentiment
w_1	w_2	w_3	0
w_1	w_3	w_3	1
w_2	w_2	w_1	1

- unique words → w_1 w_2 w_3
- Applying BOW →

	w_1	w_2	w_3	sentiment
review 1	1	1	1	0 (-ve)
review 2	1	0	2	1 (+ve)
review-3	1	2	0	1 (+ve)

- understanding need of Laplace Additive smoothing
- Ex: We got new review → { w_1 w_1 w_2 }
  - Our task is to predict whether the review is +ve or -ve

$P(+ve \mid review 4)$ $P(-ve \mid review 4)$

- Apply BOW on new query

Review 4	w_1	w_2	w_3
	3	0	0

- $P(+ve \mid r4) = P(w_1 = 3 \mid +ve) \times P(w_2 = 0 \mid +ve) \times P(w_3 = 0 \mid +ve) \times P(+ve)$

$= 0 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}$

$= 0$

$P(-ve \mid r4) = P(w_1 = 3 \mid -ve) \times P(w_2 = 0 \mid -ve) \times P(w_3 = 0 \mid -ve) \times P(-ve)$

$= \quad \times 0 \times \frac{1}{3}$

$= 0$

- Both probability become zero, hence can't predict the sentiment

- This is the problem, in some cases the particular feature in new query may for a particular class e.g $P(w_1=3|+w)$ may not exist in input feature for that class, this leads to 0 of probability → whole product will be zero

- Q can we solve it using log prob
Ans: No, because log(0) is undefined

- This problem is solved by Laplace additive smoothing

- It will prevent the particular term to not become zero

- One possible solution can be replacing with very small no ex → 0.00001, but it is generally not good as we are fixing the value

- Second soln → Laplace Additive Smoothing

- In LAS, you add something in numerator as well as denominator in all probabilities

Ex: we have Probability → $\frac{\boxed{1} + \alpha}{\boxed{3} + n\alpha}$

P( )

M	T	W	T	F	S	S

Page No.: YOUVA

Date:

α is generally 1

Numerator

- so even if we have $\Rightarrow \frac{0+\alpha}{1}$, α will prevent it from becoming zero

- Denominator
  - $\frac{1+\alpha}{3+n\alpha}$ $\Rightarrow$ what is this n?
  - n actually vary, It depends on which type of NAIVE Bayes you are using

- Example: $P(+ve \mid x_4)$

$$= P(w_1=3 \mid +ve) \times P(w_2=0 \mid +ve) \times P(w_3=0 \mid +ve) \times P(+ve)$$

$$= 0 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}$$

Applying LAS $(\alpha=1, n=2)$

$$= \frac{0+\alpha}{+n\alpha} \times \frac{1+\alpha}{2+n\alpha} \times \frac{1+\alpha}{2+n\alpha} \times \frac{1+\alpha}{2+n\alpha} \times \frac{1+\alpha}{3+3n\alpha}$$

Benefit: Prevents from total prob $= 0$

- But why did we use complex thing instead of just replacing with small number
- The answer is in Bias Trade Off.
- Except for Gaussian Naive Bayes, all have this α

Bias variance tradeoff

- Why $\frac{+\alpha}{+n\alpha}$?

- because of $_{\wedge}$ Bias variance
  (because to flexibility of)

- If model is high bias, we can control it using α, and vice versa

let's take scenario

① α is small

- let's say $\alpha = 0$ (can't be -ve)
- Data:

f^1, f^2, f^3 ...	y
...	Yes
...	No
...	Yes

1000 rows

- 500 rows are labelled yes & 500 are labelled No

- $P(Yes|X) = P(Y) \times P(f_1|Y) \times P(f_2|Y) \ldots$

let's say this prob is very small $= \frac{1}{500}$

- This indicates f_1 value doesn't contribute too much when output col have yes only

- It can also be zero $\Rightarrow \frac{0}{500}$ when data is changed a little bit

- Let's say we changed the data again now $\Rightarrow \frac{3}{500}$

- We can see, if training data is changed little bit, there is too much variation in O/P. as when one term is zero, whole prob was getting zero

- This indicate high variance, that means if α value is too small, there is a chance of overfitting (high variance)

② α value is very large

- $\alpha = 1000$
- and particular feature probability was $\frac{1}{500}$
- Applying LAS $\Rightarrow \frac{1 + 1000}{500 + 2(1000)}$ (Assume $n = 2$)

New Prob of that feature $= \frac{2}{5}$

- $\alpha = 10{,}000$
- $\frac{1 + 10000}{500 + 20000} = \frac{1}{2}$

- As you increasing α, all the probabilities are reaching to half.

M T W T F S S

Page No.: | Date: | YOUVA

Let's check again

$P(Y|X) = P(f_1|Y) \times P(f_2|Y) \cdots \cdots P(f_n|Y) \times P(Y)$

- As you increase α, all of these values tend to reach half
- same goes for $P(N|X)$ → all of it's prob also reaches half.
- Since both of their all prob are half. The product of $P(Y|X)$ values is $(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \cdots)$
  the " " $P(N|X)$ " will also be $(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2})$
- This means both $P(Y|X)$ and $P(N|X)$ will be same
- The thing that will make $P(Y|X)$ and $P(N|X)$ different will depend on data ie which class is more Yes or No
- If there are more yes points
  $P(Y|X) = \underline{P(Y)} \times P(f_1|X) \times P(f_2|X) \cdots$

  this value will be more than $P(N)$
- and now $P(Y)$ Yes class will be selected
- What does it mean?
- It means if there is a new query point, we will assign that class whichever is more, your result will always be yes.

which indicates underfitting (High Bias)

Conclusion

- high alpha(α) $\Rightarrow$ high Bias / underfitting
- low alpha(α) $\Rightarrow$ high variance / overfitting

- α can be tuned, which means we can toggle between overfitting and underfitting

- Reasons for using laplace Additive Smoothing

→ Preventing Prob to become zero
→ can reduce underfitting or overfitting using α

→ n value will be understand when we will study different types of Naive Bayes