Naive Bayes Ils a probability based classification algorithm Conditional probability P(A|B) = P(A=a|B=b)probability that A = a given B = bA, B -> Random Variables. let, D1: the outcome of die 1 D2: the outcome of die 2 Question: What is the probability that D1=2 given that ( € looks: P(Di=2 | Di+D2 ≤ 5)

O possible values for D1+D2 ≤ 5  $=\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3)\}$ (3,1), (3,2), (4,1) } outcomes D, = 2 'p (D1=7 D1+ D2 < 5) = Valid only  $P(A|B) = P(A \cap B)$ P(B) Note: P(A NB) are -> probability of both events occurring P(A1.B) - Individual probability of A occurring given that is has occurred tolisticand reinstant

Scanned by CamScanner

Gr2. Suppose an individual applying to collège determines that he has 80%. Change of being accepted 2 he knowthat dornitory housing will only be provided to 60% of all proof: the accepted students. Then P (Accepted and Dormstory housing). = P (Dosmitory housing | Accepted) \* P(Accepted) = 0.6 \* 0.8 = 0.48 Independent VIS Mutually Exclusive events > A & B are independent J. P(A|B) = P(A)

P(B|A) = P(B) Example 1:

A: Getting value of 6. In die 1 throw (D1=6)

B: H. II 3 In die 2 throw (D2=3)  $\Rightarrow P(D_1 = 6 \mid D_2 = 3) = P(D_1 = 6) = \frac{1}{6}$ P(A|B) = P(B|A) = 0 ) then 1 & B are events P(A) = D(B) = P(B) = D

P(A) = D

P(A) = D

P(A) = D

P(B) = P(B) = D

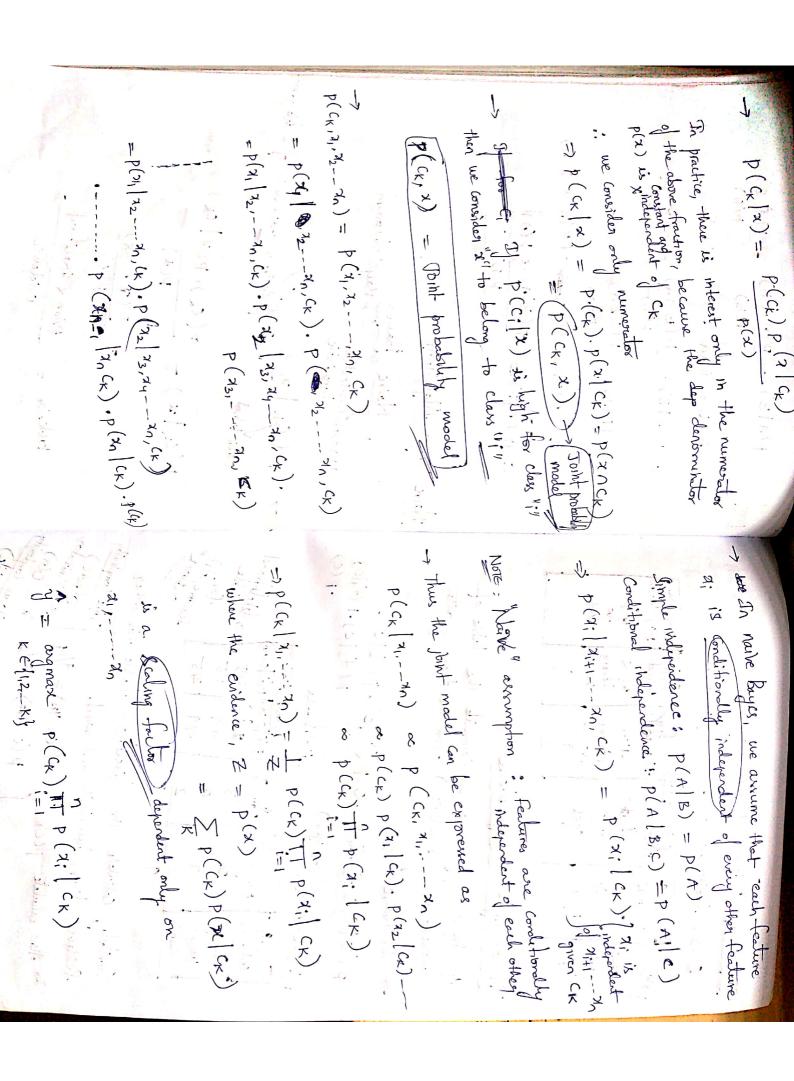
P(B) = D Bayes Theorem:

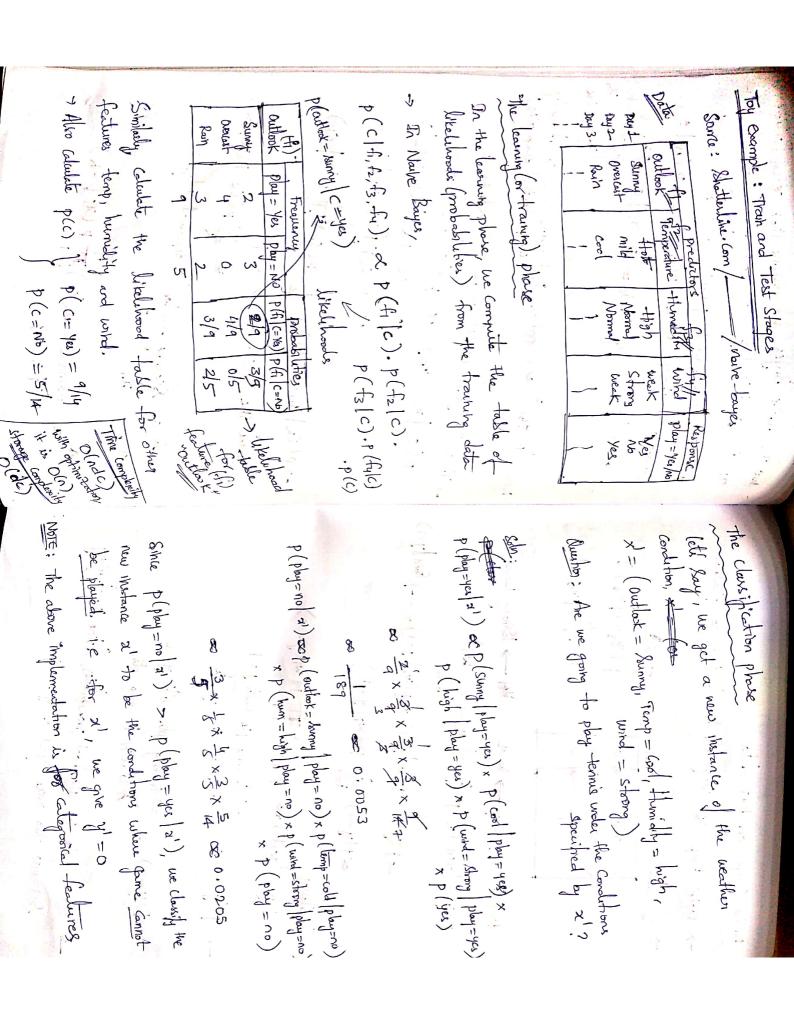
(P(A|B) = P(B|A)(P(A)) if p(B) = 0

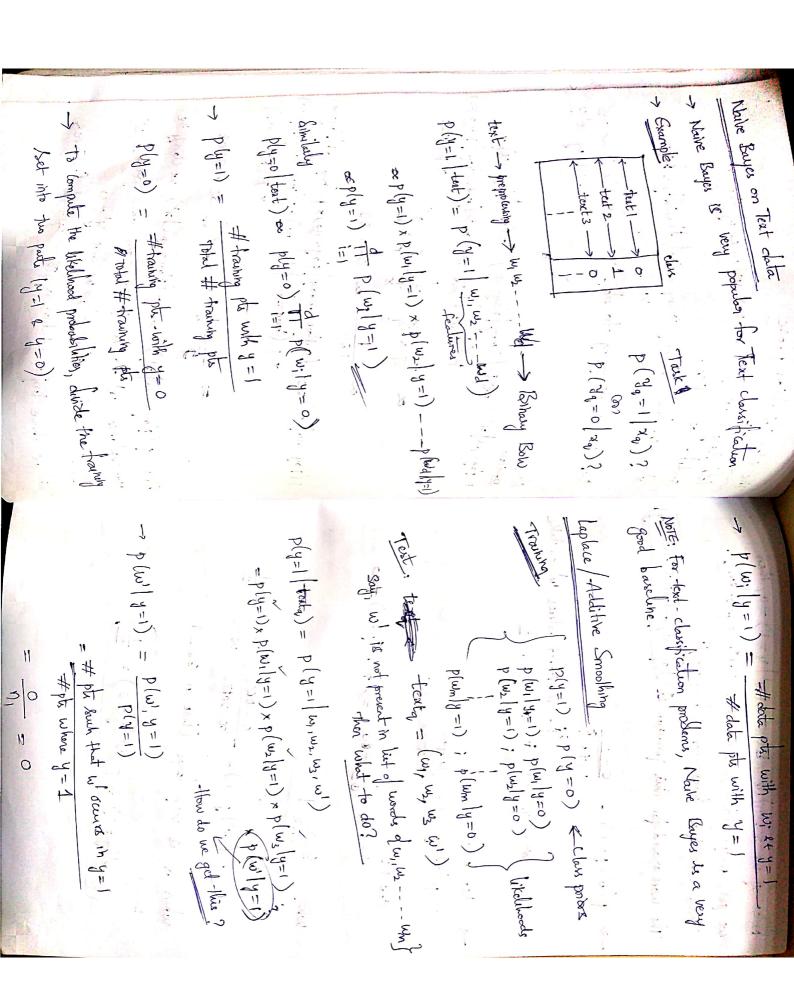
Postesior probability

Postesior probability

 $p(A|B) = \underbrace{P(A \cap B)}_{p(B)} = \underbrace{P(\overline{A}, B)}_{p(B)}$ ANB = BAA -> From Set theory  $p(A|B) = \frac{p(B \cap A)}{p(B)} = \frac{p(B, A)}{p(B)}$  $p(B|A) = \frac{p(B \cap A)}{p(A)}$ => p(B) = P(B) P(B) \* P(A)  $\therefore p(A|B) = p(B|A) * p(A)$  p(B)Naive Bayes Algorithm Naive Bayes Rayes theorem Simplishic  $p(c_k|x) = \frac{p(c_k) \cdot p(x|c_k)}{p(x)}$ Posterior = Prior x likelihood :



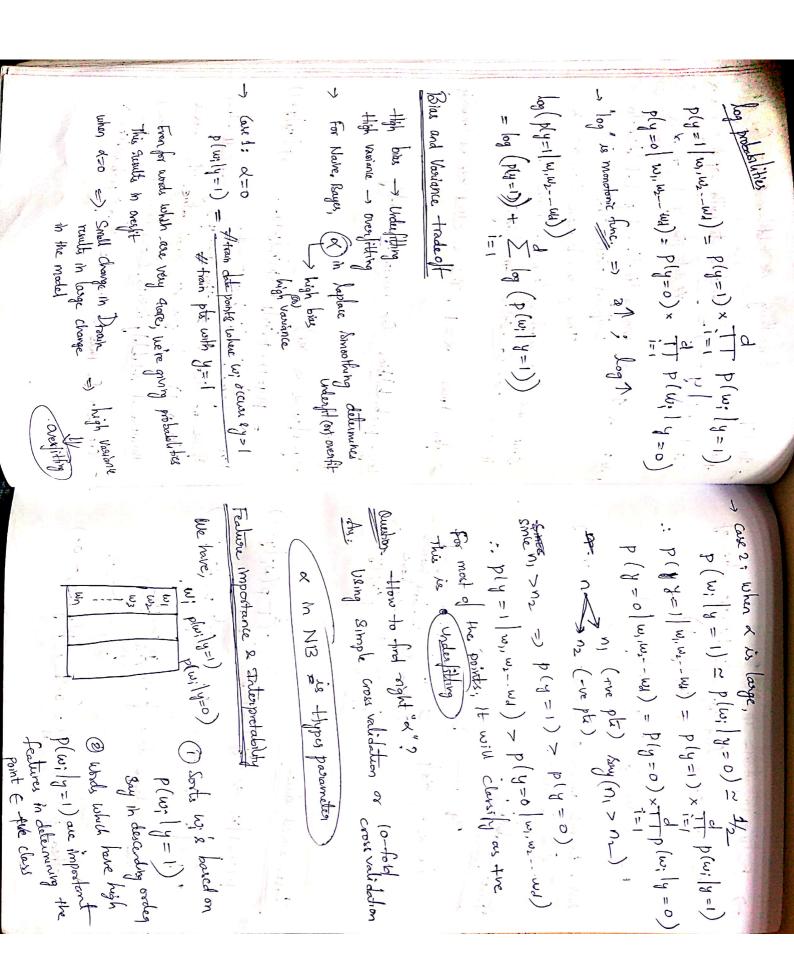




> p(w| | y=1) = 0+a > let n/=0, .... K=#district values w Gar Tuke

K=#district values w Gar Tuke

K=1 Typically | In this Case K=2 we get p(y=1/tenty) also zero. That a possen! In Caplace Smoothing, we add a R ak for all Why Otal 1 The work around for this possbern is (Layslace) Addh her Sinsoff when a is very large (like in Gaz 2) 1 p(wly=1)= 0+α Cex 1: (p(w|y=1) = p(w|y=0) = = ] The state of Note: a 1 -> moning likelihood probabilities to p(y=1/w, w2, w3--wd) - One solo for this is to use ( hop to probabilities) -> often times, d=1, this is called add-one smoothing Log-probabilities for runnesical stability How all probabilities like blu 0 2 1. since we are adoling teems in numerator a denom, The mo let d = 100, then multiplying 100 small values the words. | = 1) = # date pts . y=1 + & K it is called additive Smoothing. numerical stability issues. (ex: numerical underflow) (0<p<1), will give very small value. This causes = p(y=1) x p (m, y=1) x p (m=/y=1) -- - x p (md/y=1) the prob will move close to 1/2 Pic a when Troit More



-> Feature importance in NB : Moto: In NB, important features are obtained directly b(1=1/m, m==m) = b(1=1), I=1 b(m: [1=1) Since imbalance data } => (tre) (pt) for -ve closs b --- p(w; ly=0) for the class of find m; with highest value of Then ply=1) = 1 = 0-9, ply=0)=0.) because La Contains words dw, we --- why I am isombouling you = } from the model. 24 - And I want of the state of high value of p(w, |y=1) when we have imbalanced deta then for an imbalanced detaset, - while aprilying taplace smoothing, the effect of a is more on @ Calegorial features -> Good Consider New as category Note Outlier occurs very few times in the R-re closes (3) Numerical features .--> Inquitation 1) Text-data -> No case of missing data. Mixing Values: (1) 29 29 W1 W2 W3 W Outries -> Drop outliers or apply laplace smoothing w' & dw, w2 -- wm} = set of words in the profession of the same say (1) Due to class prissi - may beity / donn hant class Solo: O upsampling (on Down Sampling minority dans probability. \$ (w| (y =1) = 0 = 0 = ) w to an outlier. 1 make p(y=1)=p(y=0)=4 or 1/2

(5) Runtime Complexity, train time complexity)
3 Used for tent classification (5) Extensively used when we have categorical features (6) Interpretability
Best & hask Gares for NIS  (1) Conclitational independence of (continued)
Large chmensionality  I large dimensionality is large, we log probabilities  When dimensionality is large, we log probabilities
Similarity or distance natural or distance material
Michila classifichion
Handling Nimenic 1 Features (Gaussian NB)