CATEGORICAL NAIVE BAYES EXPERIMENT REPORT

GANESHREDDY ANNAPAREDDY

50442295

ganeshre@buffalo.edu

1. NAIVE BAYES

Naive Bayes is a conditional probability model. That is, given a feature vector x it produces the conditional probability $P(y|x)$ that the dependent variable y has a given value, given that the feature vector is x. To turn this into a classifier, we simply take the value of y which maximizes $P(y|x)$. A Naive Bayesian model is easy to build with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity the Naive Bayesian classifier often does surprisingly well and is widely used.

" Sometimes the attributes of a base classifier are not continuous and will have to deal with categorical values and that's where the categorical features come in. "

DESIGN OF A NAIVE BAYES CLASSIFIER

⇒ We will have to consider that our dataset has '*n*' features.

⇒ Now considering that we have '*m*' classes C_1, C_2, C_3, C_4 ..... C_m just compute for the unknown sample ***X***, the following for each class C_i

$$— P(C_i|\boldsymbol{X}) = P(C_i)_p(\boldsymbol{X}|C_i) \,/\, P(\boldsymbol{X})$$

⇒ Now I am going to assign the sample ***X*** to the class *I* which has the highest $P(C_i|\boldsymbol{X})$

⇒ The naïve assumptions :

$— P(\boldsymbol{X}|C_i) = p(x_1|Ci) * ... * p(x_d|Ci)$

$— p(x_1|Ci) = 1*(e^{-1/2(x-\mu x/\sigma x)^2})/ \sigma_x \sqrt{2\Pi}$

DATASET

I have taken the tennis dataset from Kaggle (https://www.kaggle.com/

	day	outlook	temp	humidity	wind	play
0	D1	Sunny	Hot	High	Weak	No
1	D2	Sunny	Hot	High	Strong	No
2	D3	Overcast	Hot	High	Weak	Yes
3	D4	Rain	Mild	High	Weak	Yes
4	D5	Rain	Cool	Normal	Weak	Yes
5	D6	Rain	Cool	Normal	Strong	No
6	D7	Overcast	Cool	Normal	Strong	Yes
7	D8	Sunny	Mild	High	Weak	No
8	D9	Sunny	Cool	Normal	Weak	Yes
9	D10	Rain	Mild	Normal	Weak	Yes
10	D11	Sunny	Mild	Normal	Strong	Yes
11	D12	Overcast	Mild	High	Strong	Yes
12	D13	Overcast	Hot	Normal	Weak	Yes
13	D14	Rain	Mild	High	Strong	No

COLUMN	DESCRIPTION
day	Tells us what day it is
outlook	It gives us the information on how the weather is on that day
temp	It tells if the temperature is hot, cool or mild on the respective day
humidity	It tells us about the humidity level on the specific day
wind	It tells us on how the wind is flowing in what force
play	It tells us if the play can be continued or not

Training :

I used the play tennis dataset to get the Categorical Naive Bayes model and see the evaluation result. Python was used for this task on the Jupiter notebook. This experiment is mostly based on the probability. I have used NumPy, pandas and sk-learns libraries for the experiment. I have used two of python functions like

COLUMN	DESCRIPTION
value_counts	To get the data type of the column in a dataset
pd.crosstab	To get the table comparing two values

- Got the probability of both "YES" and "NO"
- Then I got the table view probability of

py — (YES)	0.6428571428571429
pn — (NO)	0.35714285714285715

Outlook & play

play	No	Yes
outlook		
Overcast	0	4
Rain	2	3
Sunny	3	2

Temperature & play

play	No	Yes
temp		
Cool	1	3
Hot	2	2
Mild	2	4

Humidity & play

play	No	Yes
humidity		
High	4	3
Normal	1	6

Wind & play

play	No	Yes
wind		
Strong	3	3
Weak	2	6

Outlook Probability :

overcast - yes = 4/9; overcast - no = 0; rain - yes = 3/9; rain - no = 2/5; sunny - yes = 2/9; sunny - no = 3/5

Temperature Probability :

cool - yes = 3/9; cool - no = 1/5; hot - yes = 2/9; hot - no = 2/5; mild - yes = 4/9; mild - no = 2/5

Humidity Probability :

high - yes = 3/9; high - no = 4/5; Normal - yes = 6/9; Normal- no = 1/5

Wind Probability :

strong- yes = 3/9; strong- no = 3/5; weak- yes = 6/9; weak - no = 2/5

OUTPUT :

I am going to find the probability of "YES" using the formula pyes = py * photy * phighy * pweaky , and the output I got was

P(YES)	**0.031746031746031744**

I am going to find the probability of "NO" using the formula pno = pn * photn * phighn * pweakn , and the output I got was

P(NO)	**0.04571428571428573**

I considered the maximum probability which is no as output This is how internally naive bayes classifier works.