**Lab 08**

**Naïve Bayes**

1. What is Naïve Bayes Algorithm?

Naive Bayes is among one of the very simple and powerful algorithms for

classification based on Bayes Theorem with an assumption of independence among

the predictors. The Naive Bayes classifier assumes that the presence of a feature in

a class is not related to any other feature. Naive Bayes is a classification algorithm

for binary and multi-class classification problems.

2. Bayes Theorem

Based on prior knowledge of conditions that may be related to an event,

Bayes theorem describes the probability of the event

conditional probability can be found this way

Assume we have a Hypothesis(H) and evidence(E),

According to Bayes theorem, the relationship between the probability of

Hypothesis before getting the evidence represented as P(H) and the

probability of the hypothesis after getting the evidence represented

as P(H|E) is:

P(H|E) = P(E|H)\*P(H)/P(E)

Prior probability = P(H) is the probability before getting the evidence

Posterior probability = P(H|E) is the probability after getting evidence

In general,

P(class|data) = (P(data|class) \* P(class)) / P(data)

Bayes Theorem Example:

Assume we have to find the probability of the randomly picked card to be king given

that it is a face card.

There are 4 Kings in a Deck of Cards which implies that P(King) = 4/52

as all the Kings are face Cards so P(Face|King) = 1

there are 3 Face Cards in a Suit of 13 cards and there are 4 Suits in total so P(Face)

= 12/52

Therefore,

P(King|face) = P(face|king)\*P(king)/P(face) = 1/3

Types of Naïve Bayes:

1

Y

These three distributions are so common that the Naive Bayes implementation is often

named after the distribution. For example:

Binomial Naive Bayes: Naive Bayes that uses a binomial distribution.

Multinomial Naive Bayes: Naive Bayes that uses a multinomial distribution.

Gaussian Naive Bayes: Naive Bayes that uses a Gaussian distribution.

A dataset with mixed data types for the input variables may require the selection of

different types of data distributions for each variable.

Using one of the three common distributions is not mandatory; for example, if a real-

valued variable is known to have a different specific distribution, such as exponential,

then that specific distribution may be used instead. If a real-valued variable does not

have a well-defined distribution, such as bimodal or multimodal, then a kernel density

estimator can be used to estimate the probability distribution instead.

1 The Classifier

The Bayes Naive classifier selects the most likely classification V nb given the

attribute values a 1 , a 2 , . . . a n . This results in:

V nb = argmaxvj ∈V P (v j ) P

(a i |v j ) (1)

We generally estimate P (a i |v j ) using m-estimates:

where:

P (a i |v j ) =n c + mp (2)

n + m

1

|

n = the number of training examples for

which v = v j n c = number of examples for

which v = v j and a = a i p = a priori

estimate for P (a i v j )

m = the equivalent sample size

2 Car theft Example

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

2.1 data set

Example No. Color Type Origin Stolen?

1 Red Sports Domestic Yes

2 Red Sports Domestic No

3 Red Sports Domestic Yes

4 Yellow Sports Domestic No

5 Yellow Sports Imported Yes

6 Yellow SUV Imported No

7 Yellow SUV Imported Yes

8 Yellow SUV Domestic No

9 Red SUV Imported No

10 Red Sports Imported Yes

2.2 Training example

We want to classify a Red Domestic SUV. Note there is no example of a Red

Domestic SUV in our data set. Looking back at equation (2) we can see how to

compute this. We need to calculate the probabilities

P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,

P(Red|No) , P(SUV|No), and P(Domestic|No)

and multiply them by P(Yes) and P(No) respectively . We can estimate these

values using equation (3).

Yes: No:

Red: Red:

n = 5 n = 5

1

|

|

5 + 3 5 + 3

5 + 3 5 + 3

5 + 3 5 + 3

n\_c= 3 n\_c = 2

p = .5 p = .5

m = 3 m = 3

SUV: SUV:

n = 5 n = 5

n\_c = 1 n\_c = 3

p = .5 p = .5

m = 3 m = 3

Domestic: Domestic:

n = 5 n = 5

n\_c = 2 n\_c = 3

p = .5 p = .5

m = 3 m =3

Looking at P (Red Y es), we have 5 cases where v j = Yes , and in 3 of those

cases a i = Red. So for P (Red Y es), n = 5 and n c = 3. Note that all attribute are

binary (two possible values). We are assuming no other information so, p = 1 /

(number-of-attribute-values) = 0.5 for all of our attributes. Our m value is

arbitrary, (We will use m = 3) but consistent for all attributes. Now we simply

apply eqauation (3) using the precomputed values of n , n c , p, and m.

P (Red|Y es) = 3 + 3 ∗ .5 = .56 P (Red|No) = 2 + 3 ∗ .5 = .43

P (SUV |Y es) = 1 + 3 ∗ .5 = .31 P (SUV |No) = 3 + 3 ∗ .5 = .56

P (Domestic|Y es) = 2 + 3 ∗ .5 = .43 P (Domestic|No) = 3 + 3 ∗ .5 = .56

We have P (Y es) = .5 and P (No) = .5, so we can apply equation (2). For v = Y

es, we have

P(Yes) \* P(Red | Yes) \* P(SUV | Yes) \* P(Domestic|Yes)

= .5 \* .56 \* .31 \* .43 = .037

and for v = No, we have

P(No) \* P(Red | No) \* P(SUV | No) \* P (Domestic | No)

= .5 \* .43 \* .56 \* .56 = .069

Since 0.069 &gt; 0.037, our example gets classified as ’NO’

1

Task

ABOUT DATASET: It is for non-functional requirement analysis. 5 different classes.

Explore the dataset carefully

1. Plot the class count

2. Encode the labels

3. Count the words in each row

4. Convert the text to lower case and split into words

5. Remove the alpha-numeric

6. Remove the stop words i.e. the, is, an, a, here, their, there etc. (without nltk)

6. Split the dataset to 75 25

7. Use Bag of Word for vectorization(feature extraction)

8. Implement the models( variations of naive bayes).

9. Predict the accuracy in case of class imbalance f1-score

10. Comparison of different variations