Day 6

Agenda :

1. Naive Bays Algorithm.
2. Performance matrix for classification
    1. Confusion Matrix
    2. Precesion, Recall, F1 score, AUC, ROC
3. Sentiment analysis Project
4. Imbalancity treatment
5. Churn Prediction Project
6. Predict if a person will purchase iPhone or not.
7. MCQs

Naive Bays Algorithem

Conditional Probability :

Conditional probability is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred.

Probability of event A occured and event B occured

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of event A given B has occured

Probability of event B

P(B) != 0

Bays Theorem

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities.

Probability of B occurring given evidence A has already occurred

Probability of A occurring

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Y: class of the variable

X: dependent feature vector (of size n)

Naive Bays

-- Bayes' rule provides us with the formula for the probability of Y given some feature X. In real-world problems, we hardly find any case where there is only one feature.

-- When the features are independent, we can extend Bayes' rule to what is called Naive Bayes which assumes that the features are independent that means changing the value of one feature doesn't influence the values of other variables and this is why we call this algorithm "NAIVE"

$$P(Y=k|X1, X2....Xn) = \frac{P(X1|Y=k) * P(X2|Y=k)...... * P(Xn|Y=k) * P(Y=k)}{P(X1) * P(X2).... * P(Xn)}$$

This formula can also be understood as

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

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

Frequency Table

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

Frequency Table

		Stolen?	
		Yes	No
Type	Sports	4	2
	SUV	1	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

Frequency Table

		Stolen?	
		Yes	No
Origin	Domestic	2	3
	Imported	3	2

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Origin	Domestic	2/5	3/5
	Imported	3/5	2/5

Color	Type	Origin	Stolen
Red	SUV	Domestic	?

As per the equations discussed above, we can calculate the posterior probability P(Yes | X) as :

P(Yes | X) = P(Red | Yes) * P(SUV | Yes) * P(Domestic | Yes) * P(Yes)

$= \frac{3}{5} * \frac{1}{5} * \frac{2}{5} * 1$
= 0.048

and, P(No | X):

P(No | X) = P(Red | No) * P(SUV | No) * P(Domestic | No) * P(No)

$= \frac{2}{5} * \frac{3}{5} * \frac{3}{5} * 1$
= 0.144

Since 0.144 > 0.048, Which means given the features RED SUV and Domestic, our example gets classified as 'NO' the car is not stolen.

Example :

Frequency and Likelihood tables of 'Color'

Frequency and Likelihood tables of 'Type'

Frequency and Likelihood tables of 'Origin'

So in our example, we have 3 predictors X.