Practical Machine Learning with R

Lecture 3

Objectives:

- Introduction to Naïve Baye's, an eager learner
- Characteristics of Naïve Baye's and its comparison with KNN
- Fundamental knowledge of Probability and Conditional Probability.
- Application of Baye's Rule to create a Classifier.

Characteristics of Naïve Baye's Algorithm :

- This is a classification algorithm that has a probabilistic framework. This encapsulates a group of classifiers therefore is a family of algorithms. These algorithms entail both supervised and semi supervised learning systems. Gaussian Naïve Baye's, Kernel Density Baye's,Multinomial Naïve Baye's, Bernoulli Naïve Baye's as well as semi supervised parameter estimation are a few of the popular ones.
- Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. The distribution is not assumed to be normal. This can be advantageous in context to continuous features.
- This algorithm is based on the Baye's theorem which was formulated by Reverend Baye's in the 1800s.
- The foundational principle of this algorithm is based on the fact that the probabilities of events can be updated by additional information obtained.
- The theorem is termed "Naïve" primarily due to the inherent assumption that the prediction features/attributes are class conditionally independent of each other and are equally important.
- Despite the aforementioned assumption which is not realized realistically, compared to other ML algorithms the prediction accuracy is high for Naïve Bayes.
- This algorithm is a popular one in industry due to its high performance and simple implementation.
- Google uses this algorithm for Page Rank to implement search criterion in context of pulling and placing the pages from the underlying databases.
- Additionally, It is used for text classification, sentiment classification, spam classification, disease classification, target marketing, medical diagnosis and credit approval etc.
- It is easy to build this algorithm, it is easily trainable, requires a small data set to create the training set and is it is fast.
- It showcases good performance with a moderately large data set. It works well with many predictor feature vector as well and unlike KNN it is not sensitive to irrelevant features which might generate noise.
- The Complexity of Naïve Baye's is = Nd where N is the number of examples and d is the dimensionality of the features.
- KNN is slower if there is a lot of data but KNN is better if the prediction features are highly correlated rather than independent.
- KNN is non-parametric therefore the boundaries can take any form whereas Naïve Baye's can take only linear, elliptical and parabolic boundaries.
- KNN weighs every feature equally unless we are using weighted distances.
- Naïve Baye's just ignores missing valued feature whereas KNN cannot conduct classification if missing data exists.
- Naïve' Bayes works well with noisy data. Noisy data is averaged out.
- KNN is slow during prediction phase as compared to Naïve Bayes.
- KNN is lazy learner whereas Naïve Bayes is Eager learner. Lazy learning is instance-based learning since this algorithm stores the training data waiting to receive the training example to classify therefore it does not entail the phase of generalization since there is no model or classifier.

Eager Learning receives the training data followed by creating a classification model prior to receiving a test set.

- Correlated attributes/features degrade the performance of Naïve Baye's.
- Naïve Baye's does not work well with too many numeric features.

Fundamental Knowledge:

Probability, Joint Probability and Conditional Probability

Unconditional Probability:

Probability of any event is always between 0 and 1.

For a scenario there can be a set of possible events. If the exhaustive set of events are segmented into mutually exclusive partitions, then the summation of probability of the events is equal to one.

As an illustration if we are trying to determine whether a bank transaction is fake or not then these two events are exhaustive and mutually exclusive. This also applies to the scenario of ascertaining if an email is spam or not.

The symbolic notation used for Probability is as follows:

P(Fraud) = Probability of fraudulent transaction
or
P(F) = Probability of fraudulent transaction

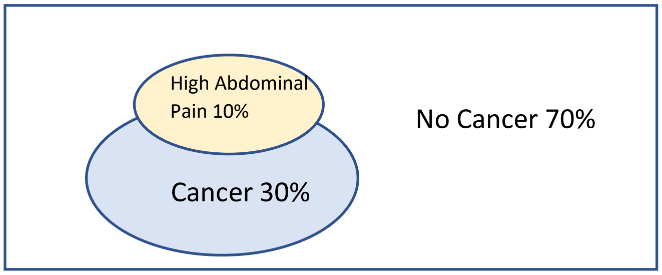
$P(F^c)$ = Probability of non-fraudulent transaction This is also termed as the complement of P(F)

$P(\neg F)$ = Probability of non-fraudulent transaction

Joint Probability:

Events that are not mutually exclusive are the ones that we wish to explore. Often the primary targeted event of interest happens concurrently with an event or a set of events. By using these set of events, we can endeavor to make predictions regarding the event of interest.

Medical Diagnosis:



The event of abdominal pain is a strong evidence of Cancer though sometimes it can occur when there is no Cancer. High Abdominal pain is not the only evident symptom for cancer. Hypothetically for an oncology facility that receive patients, say 30% have cancer and 70% do not have Cancer and of those that have Cancer 10% suffer from High abdominal pain.

Event Representation : Let A = Cancer B = Abdominal Pain

For prediction purposes we wish to evaluate:

P(Cancer) and P(Abdominal Pain) = P(Cancer) ∩ P(Abdominal Pain)
This is a joint probability. ∩ is the intersection

These are dependent events therefore we cannot use the following simplistic rule for independence

P(B and A)=P(B)*P(A) or P(B∩A)=P(B)*P(A)

Instead the multiplication rule transforms to:

$\boldsymbol{P(B\cap A)=P(B)*P(A|B)}$**……………………..**1

In this scenario we must use the fundamental theorem of conditional probabilities for dependent events:

$$P(A|B)=\frac{P(B\cap A)}{P(B)}\text{……………………..2}$$

This symbolic representation is interpreted as Probability of event A given the probability of event B equals the joint probability of A and B relative to the probability of B(Marginal Probability) happening.

$P(A\cap B)=P(B\cap A)$………………3

$P(A\cap B)=P(A)*P(B|A)$…………4

Transforming 2 by using 4

$$P(A|B)=\frac{P(B|A)*P(A)}{P(B)}\text{ …………………5}$$

Since A and A^c are mutually exclusive and exhaustive events:

$P(B)=P(B\cap A)\cup P(B\cap A^c)$………………..6

Using 4 we obtain:

$P(B)=P(A)*P(B|A)\cup P(A)*P(B|A^c)\ =P(A)*P(B|A)+P(A^c)*P(B|A^c)$ ………..7

Baye's Theorem: Is obtained by substituting 7 into 5

$$P(A|B) = \frac{P(B|A) * P(A)}{P(A) * P(B|A) + P(A^c) * P(B|A^c)} \ldots\ldots\ldots\ldots..8$$

Using 5:

$$P(Cancer|High\ Abdominal\ Pain) = \frac{P(High\ Abdominal\ Pain|Cancer) * P(Cancer)}{P(High\ Abdominal\ Pain)}$$

Probability of cancer without the data evidence can be obtained by the unconditional probability:

P(Cancer) : **Prior Probability: Prior probability of class**

Upon evaluating numerous patients, the $P(High\ Abdominal\ Pain|Cancer)$ can be evaluated.

$P(High\ Abdominal\ Pain|Cancer)$: **Likelihood of Cancer: Probability of the predictor attribute or feature given the class (target feature). This is the evidence.**

$P(High\ Abdominal\ Pain)$: **Marginal Likelihood: Prior probability of predictor attribute or feature.**

$P(Cancer|High\ Abdominal\ Pain)$: **Posterior Probability : Posterior probability of class (target feature) given the predictor attribute or feature.**

Baye's Theorem Implementation

The following table is a crosstabulation frequency table that enumerates the number of instances of class/targeted feature (Cancer/Non-Cancer) cross referenced with the predictor feature (Abdominal pain).

	Abdominal Pain		
Frequency	Yes	No	Total
Cancer	30	8	38
No Cancer	2	35	37
	32	43	75

This table showcases the probabilities pertaining the likelihood values.

	Abdominal Pain		
Likelihood	Yes	No	Total
Cancer	30/38	8/38	38
No Cancer	2/37	35/37	37
	32/75	43/75	75

For illustration purposes Likelihood of Cancer = P(High Abdominal Pain| Cancer)=30/38 =.79

This informs us that Probability that a patient has abdominal pain given the patient has cancer= 79% chance

P(High Abdominal Cancer and Cancer)=P(Cancer)*P(High Abdominal Pain| Cancer) =(38/75)*.79=.40 which can be interpreted that 40% of the cancer patients had Abdominal pain.

$$P(Cancer|High\ Abdominal\ Pain) = \frac{P(High\ Abdominal\ Pain|Cancer)*P(Cancer)}{P(High\ Abdominal\ Pain)}$$

Posterior probability $P(Cancer|High\ Abdominal\ Pain) = \frac{(.40)}{\frac{32}{75}} = 93\%$

Prior probability P(Cancer) = 38/75 = .50

Therefore, an evidence of abdominal pain updates the Prior probability to result within the Posterior probability. Posterior probability is updated using the prior probability and the evidence.

Pragmatically the posterior probability cannot be dependent on just one evidence-based predictor feature . Multiple Predictor features must be used since in this scenario Abdominal pain cannot be the only determinant of the posterior probability.

	Abdominal Pain		Blood Clots		Migraine		
Likelihood	Yes	No	Yes	No	Yes	No	Total
Cancer	30/38	8/38	32/38	6/38	29/38	9/38	38
No Cancer	2/37	35/37	3/37	34/37	10/37	27/37	37
	32/75	43/75	35/75	40/75	39/75	36/75	75

If a new patient comes in and we wish to evaluate the probability of him/her having the cancer. The patient's biomarker scores inform us that he has abdominal pain, he has blood clots but no migraine.

Now we will build his/her Posterior probability.

If there are other predictor features/attributes like blood clots (x_1) and fatigue (x_2), we can rewrite the Baye's theorem as follows

$$P(c|x_1 \cap x_2 \cap \neg x_3) = \frac{P(x_1 \cap x_2 \cap \neg x_3|c)*P(c)}{P(x_1 \cap x_2 \cap \neg x_3)} \ldots\ldots\ldots\ldots\ldots\ldots 9$$

As the number of features increase the above formulation become very complex and computationally intensive therefore, we leverage the naïve Baye's assumptions of class conditional independence.

Baye's theorem has an assumption that all the predictor features are independent of each other therefore the denominator can be adapted as follows:

$$P(x_1 \cap x_2 \cap \neg x_3) = P(x_1) * P(x_2) * P(\neg x_3) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 10$$

This is also called the class conditional independence meaning that the predictor features are independent of each other if they are conditioned on the same class.

The equation 5 can be rewritten as follows:

$$P(c|x_1 \cap x_2 \cap \neg x_3) = \frac{P(x_1 \cap x_2 \cap \neg x_3|c)*P(c)}{P(x_1)*P(x_2)*P(\neg x_3)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 11$$

The denominator are unconditional probabilities for predictor feature and do not depend on the class, specifically probability of blood clots, probability of abdominal pain and probability of no fatigue does not depend on the class (Cancer or no Cancer) therefore can be considered as a constant and left out .

The equation 7 can be generically written as

$$P(c|x_1, x_2, \neg x_3) \propto P(x_1|c) * P(x_2|c) * P(\neg x_3|c) * P(c) \ldots\ldots\ldots\ldots\ldots\ldots 12$$

Overall Likelihood of Cancer: $P(x_1|c) * P(x_2|c) * P(\neg x_3|c) * P(c) \ldots\ldots\ldots\ldots\ldots\ldots 13$

Overall Likelihood of Cancer = (30/38)*(32/38)*(9/38)*(38/75)= .79*.84*.24*.51=.0812

Overall Likelihood of no Cancer = (2/37)*(3/37)*(27/37) *(37/75)= .05*.08*.73*.49=.00143

Likelihood of Cancer/ Likelihood of no Cancer = .0812/.00143 =56 Therefore this patient is 56 times more likely to have cancer than not to have cancer.

To find the posterior probability of Cancer

For multiple features

$$P(c|x_1 \cap x_2 \cap \neg x_3) = \frac{P(x_1 \cap x_2 \cap \neg x_3|c) * P(c)}{P(x_1 \cap x_2 \cap \neg x_3)} \ldots\ldots\ldots\ldots\ldots 14$$

$$P(x_1 \cap x_2 \cap \neg x_3) = P(x_1 \cap x_2 \cap \neg x_3|c) \cup P(x_1 \cap x_2 \cap \neg x_3 c^c)$$
$$= P(x_1 \cap x_2 \cap \neg x_3|c) + P(x_1 \cap x_2 \cap \neg x_3 c^c) \ldots\ldots\ldots\ldots\ldots\ldots 15$$

$$P(c|x_1 \cap x_2 \cap \neg x_3) = \frac{P(x_1 \cap x_2 \cap \neg x_3|c) * P(c)}{P(x_1 \cap x_2 \cap \neg x_3|c) + P(x_1 \cap x_2 \cap \neg x_3|c^c)} \ldots\ldots\ldots\ldots\ldots 16$$
$$P(c|x_1 \cap x_2 \cap \neg x_3) = \frac{P(x_1 \cap x_2 \cap \neg x_3|c) * P(c)}{P(x_1 \cap x_2 \cap \neg x_3)P(c) + P(x_1 \cap x_2 \cap \neg x_3)P(c^c)} \ldots\ldots\ldots\ldots\ldots\ldots 17$$

Probability of cancer given the pattern of abdominal cancer, blood clots and no migraine predictor feature

$$\frac{.0812}{.0812 + .00143} = \frac{.0812}{.08263} = 98.2\%$$

To find the probability of the class cancer given the predictor features can be represented more formally by the following formula;

$$P(C_L|x_1, x_2 \ldots\ldots\ldots\ldots x_n) = \frac{1}{Z} p(C_L) \prod_{i=1}^{n} p(x_i|C_L)$$

C_L is the class (targeted feature Cancer no Cancer)

Z is used to convert likelihood to probability. Z is called the normalization constant.

Laplace Estimator: Sometime if the frequency of a certain feature is zero, the probability of Cancer computes to zero which causes this feature itself becomes dominant over all others. For instance, if Abdominal pain = 0 then it will controvert the effect of blood clots. To deal with this issue Laplace estimator adds a small number to all frequency so that none of the probabilities is zero.

Numerical Data: Naïve Baye's requires the data to be categorical for which the frequencies can be calculated. The numerical data can be segmented into various equitable ranges or logical ranges as per the scenario and then converted into categories. The number of categories should be chosen in order to balance the tradeoff between overfitting and underfitting the data.

Bayes Algorithm's model building phase which implements abstraction and generalization entails the calculation of Prior and conditional probability for training data set whereas the model evaluation phase entails calculation of posterior probability for test data examples and assignment of class label to the example as per the majority class.

Advantages:

If the features are conditionally independent (assumption of this algorithm) then this is a classifier generalizes well.

Naïve Baye's requires a relatively small data to train data.

Naïve Baye's is interpretable and easy to implement.

Naïve Baye's is a generative algorithm , probabilities of each class is calculated .It learns over time evolving with newer data generating newer probabilities being incorporated in the model.

The performance of Naïve Baye's degrades slowly as complexity increases as compared to other algorithms like k-NN or Decision Trees.

Disadvantages:

The implicit assumption that the features are conditionally independent is not true realistically. Naïve Baye's does not perform well with correlated features. This implies that give a class for instance given a diagnosis of Heart attack risk(yes or no) ,prediction features like weight(high or low) is independent of cholesterol(high or low).

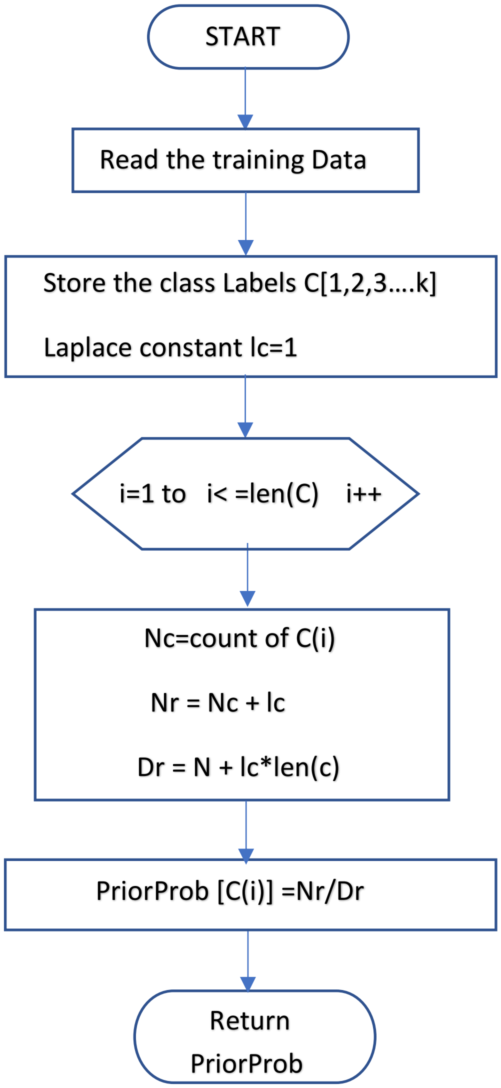
If a specific category does not occur within the training data then its probability is zero for test data .This can be handled by using Laplace Estimator.

Flowchart for Running the Algorithm

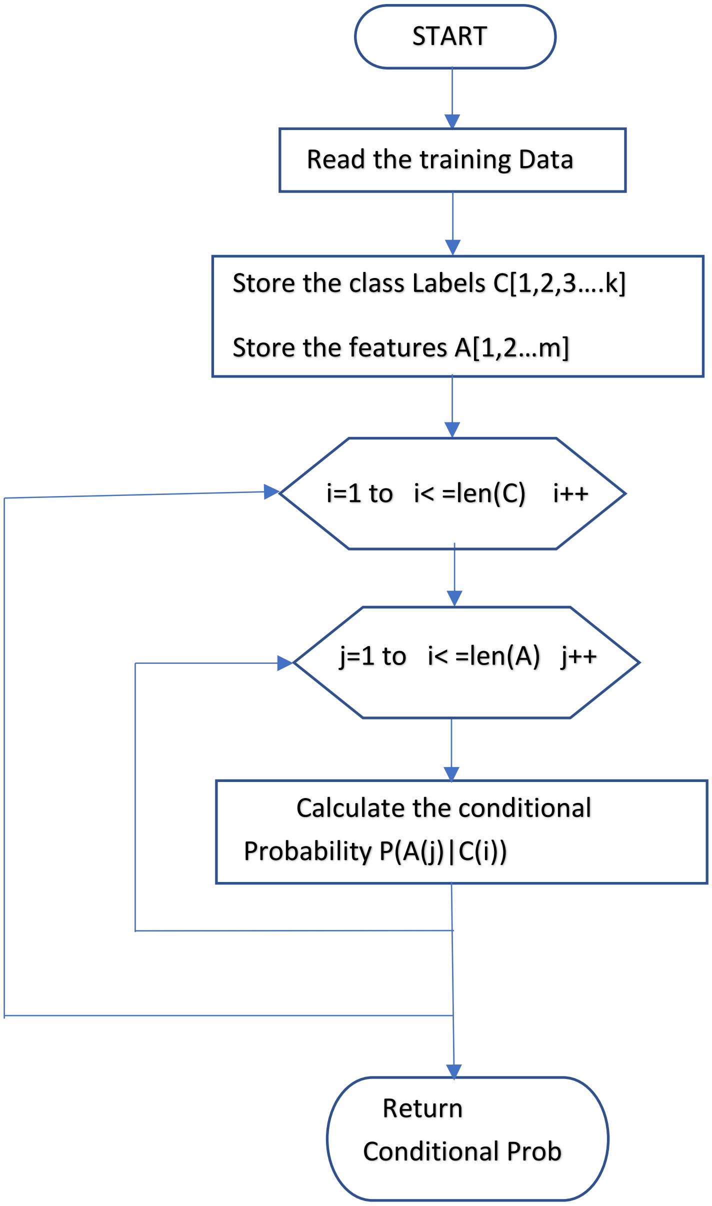
Algorithms Steps:

Training:

Step1) Calculate Prior Probability of the Class Labels:

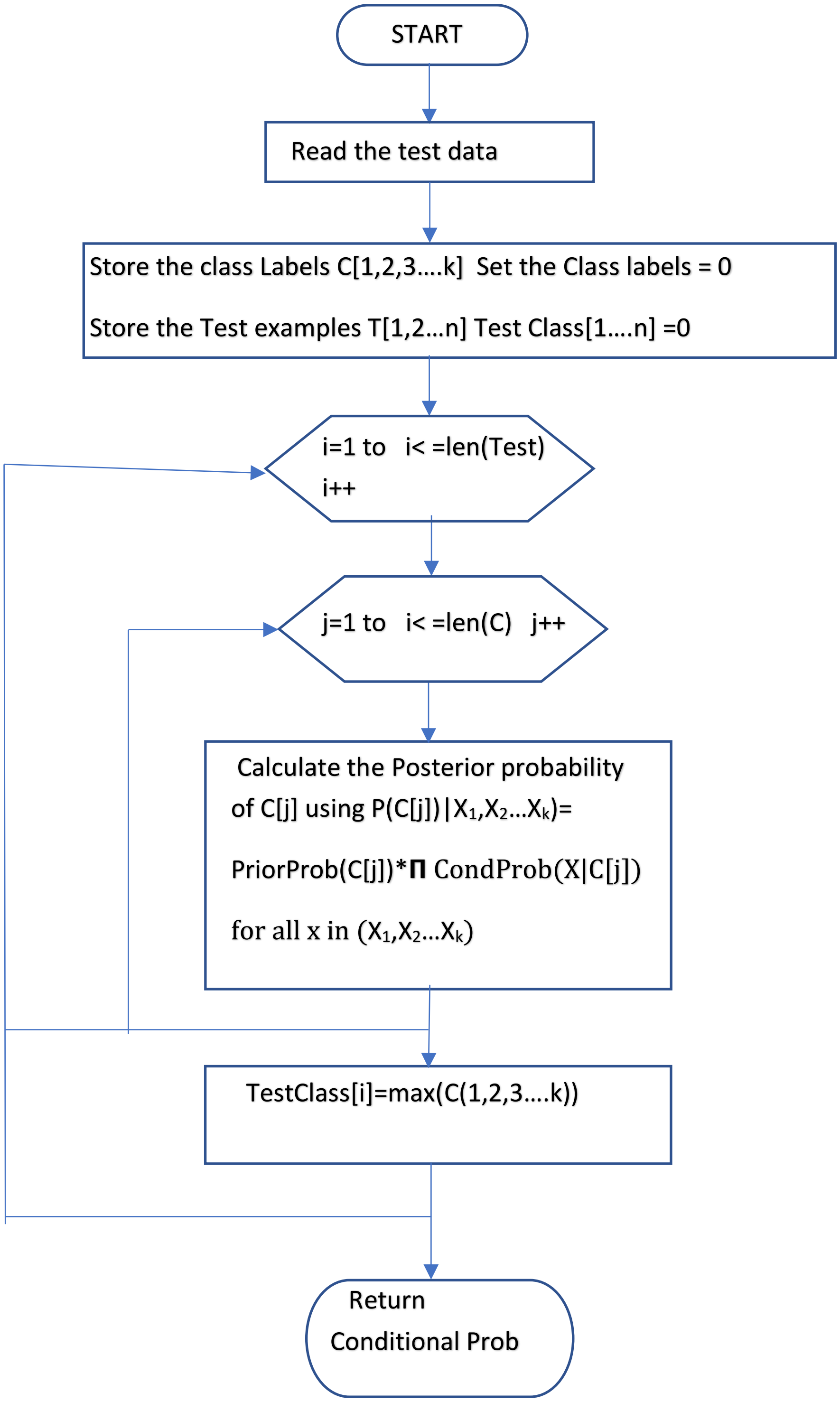


Step2) Calculate Conditional Probability of the Class Labels:



Testing:

Step2) Calculate Conditional Probability of the Class Labels:


START
Read the test data
Store the class Labels C[1,2,3….k] Set the Class labels = 0
Store the Test examples T[1,2…n] Test Class[1….n] =0
i=1 to   i< =len(Test)
i++
j=1 to   i< =len(C)   j++
Calculate the Posterior probability
of C[j] using P(C[j])|X_1,X_2…X_k)=
PriorProb(C[j])*Π CondProb(X|C[j])
for all x in (X_1,X_2…X_k)
TestClass[i]=max(C(1,2,3….k))
Return
Conditional Prob