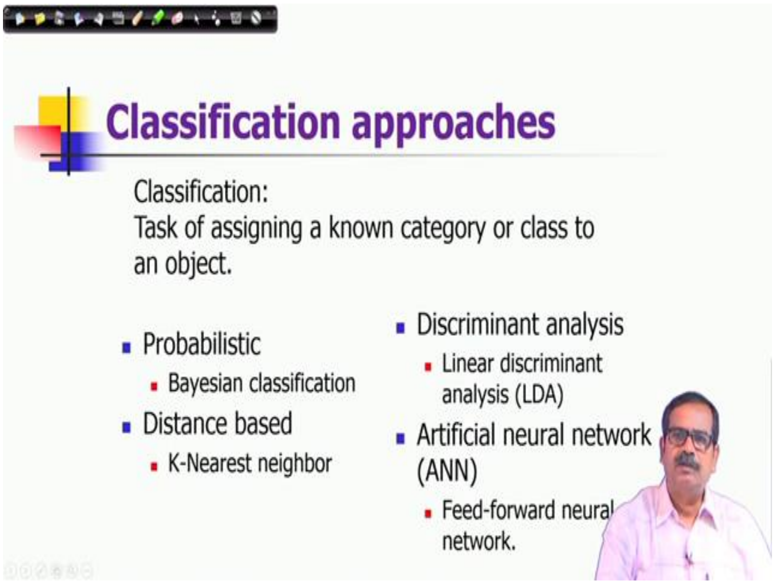
Computer Vision
Prof. Jayanta Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 47
Clustering and Classification Part – II

We continue our discussion on Clustering and Classification of data. And, in the last lecture we discussed various algorithms for clustering of data points and in this lecture I will start discussing on different classification approaches.

(Refer Slide Time: 00:37)



Classification approaches

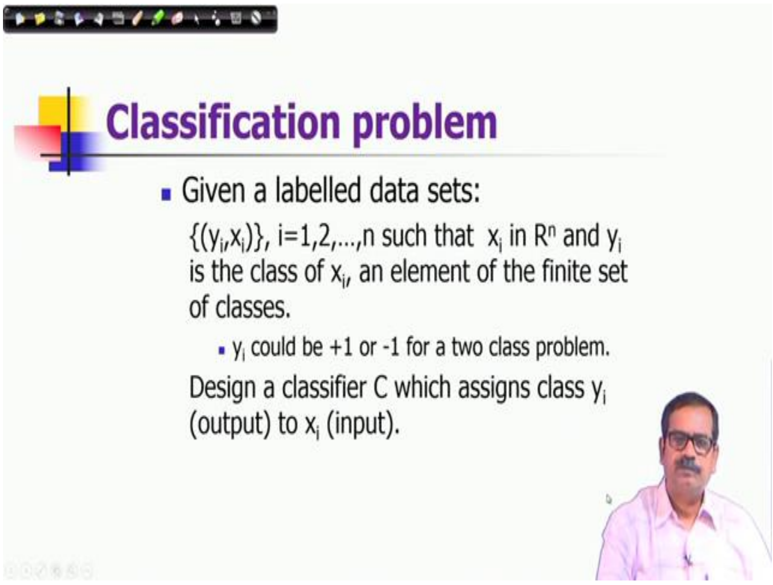
Classification:
Task of assigning a known category or class to an object.

- Probabilistic
  - Bayesian classification
- Distance based
  - K-Nearest neighbor
- Discriminant analysis
  - Linear discriminant analysis (LDA)
- Artificial neural network (ANN)
  - Feed-forward neural network.

Now, classification is a task of assigning a known category or class to an object. There are various approaches out of them few are shown here. It could be probabilistic approach, distance based, discriminant analysis best approach, you can use also artificial neutral network for classifying data points.

In this lecture or in this particular topic under this course I will be considering a few examples of each of this approaches. They are for the probabilistic approach will be considering Bayesian classification techniques, particularly will discuss about Naive based Bayesian classifier. For distance based approach will consider K nearest neighbour classifier and discriminant analysis we will discuss about linear discriminant analysis. And, then in artificial neutral network it is a feed forward neutral network that we will be discussing.

(Refer Slide Time: 01:42)



Classification problem

- Given a labelled data sets:

  $\{(y_i, x_i)\}$, i=1,2,…,n such that x_i in R^n and y_i is the class of x_i, an element of the finite set of classes.

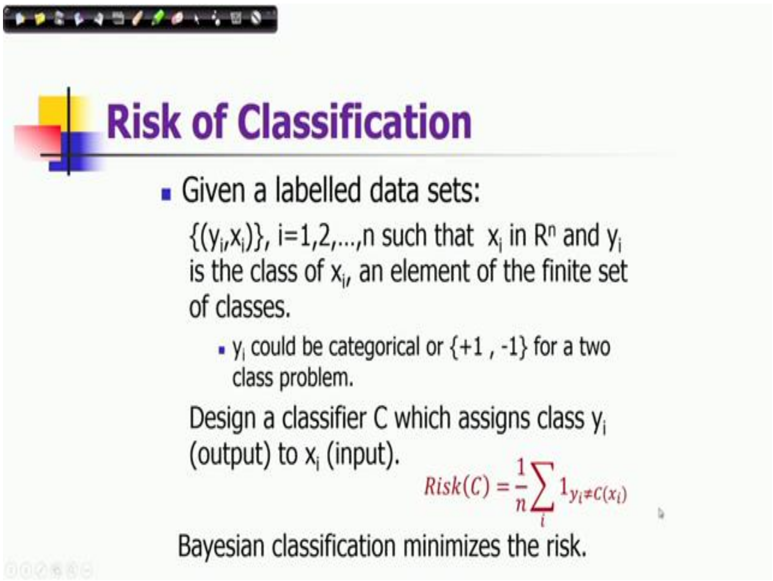
  - y_i could be +1 or -1 for a two class problem.

  Design a classifier C which assigns class y_i (output) to x_i (input).

So, let us start with the problem definition of a classification task. Here you have a labeled data sets given in the form of (yi, xi). Note here xi is the data point, this is a n dimensional vectors as for every object in an abstract way they represented as a feature vector in a n dimensional real space. So, xi is such a vector or such a point in a real dimensional in a n dimensional real space. And, yi is its class, it could be any categorical data. It could be some names of class or there could be numerical representation of those classes.

So, the numeric values distinctly identify these classes; however so, we consider of course, a finite set of classes. So, it is an element of a finite data set classes and we can design a classifier C; I mean that is the problem statement that you need to design a classifier C which assigns class yi to xi. So, which should be supported by which should supports its data set ah. So, as I mentioned that for a two class problem say yi could be denoted as + 1 or - 1.

(Refer Slide Time: 03:10)



Risk of Classification

- Given a labelled data sets:

  $\{(y_i,x_i)\}$, i=1,2,…,n such that x_i in R^n and y_i is the class of x_i, an element of the finite set of classes.

  - y_i could be categorical or {+1 , -1} for a two class problem.

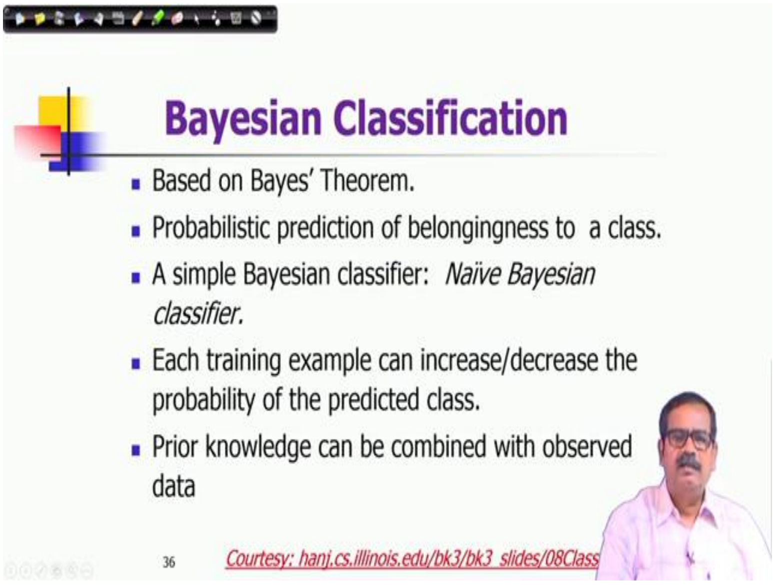
  Design a classifier C which assigns class y_i (output) to x_i (input).

  $$Risk(C) = \frac{1}{n}\sum_i 1_{y_i \neq C(x_i)}$$

  Bayesian classification minimizes the risk.

There is a risk associated with classification, particularly misclassification as you can see that risk is defined as the number of misclassified data. And, it is a fraction of number of misclassified data that is how the risk is defined and incidentally Bayesian classification which we are going to discuss next it minimizes this risk.

(Refer Slide Time: 03:36)



Bayesian Classification

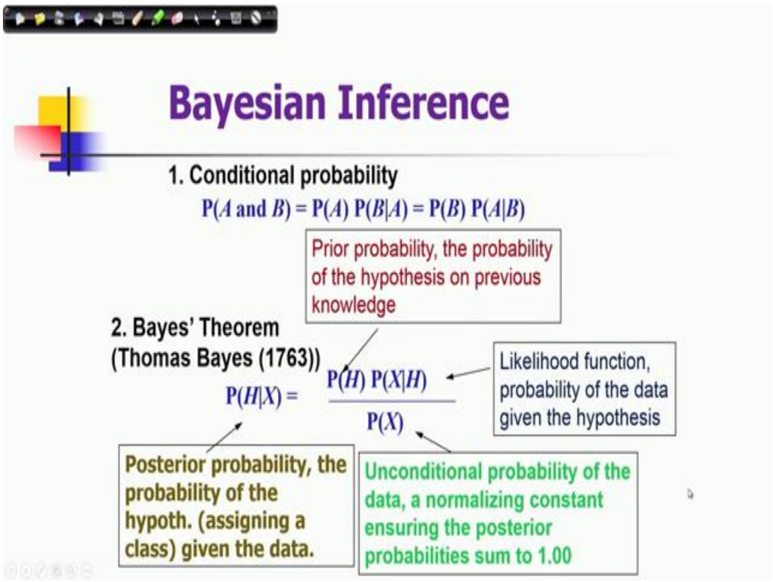
- Based on Bayes' Theorem.
- Probabilistic prediction of belongingness to a class.
- A simple Bayesian classifier: *Naïve Bayesian classifier.*
- Each training example can increase/decrease the probability of the predicted class.
- Prior knowledge can be combined with observed data

36 Courtesy: hanj.cs.illinois.edu/bk3/bk3_slides/08Class

So, let us consider this particular, thing Bayesian classification. So, it is the name comes from a theoretical framework which has been proposed by Bayes and there is a Bayes theorem. We will discuss about this theorem from which this classification rules are set and classifier set designed. It is a probabilistic prediction of belongingness to a class and a simple example Bayesian classifier that is an example of a simple Bayesian classifier is a Naive Bayesian classifier which will be considering in particular in this lecture.

Each training example in this case it can increase or decrease the probability of the predicted class. So, incrementally you can learn the classifier that is the advantage of this classification technique and prior knowledge can be combined with the observed data.

(Refer Slide Time: 04:37)



Bayesian Inference

1. Conditional probability

$$P(A \text{ and } B) = P(A)\, P(B|A) = P(B)\, P(A|B)$$

2. Bayes' Theorem (Thomas Bayes (1763))

$$P(H|X) = \frac{P(H)\, P(X|H)}{P(X)}$$

Prior probability, the probability of the hypothesis on previous knowledge

Likelihood function, probability of the data given the hypothesis

Posterior probability, the probability of the hypoth. (assigning a class) given the data.

Unconditional probability of the data, a normalizing constant ensuring the posterior probabilities sum to 1.00

So, let us consider the Bayes theorem and the inferacing that could be carried out using by applying this theorem. So, first we know the definition of a conditional probability it is from the joint probability of A and B, we can say that either no B occurs given A or A occurs given B. So, probability of A and B could be expressed as probability of a multiplied by probability of B given A or probability of B into probability of A given B. So, they are all equivalent, they should give you the same probability value.

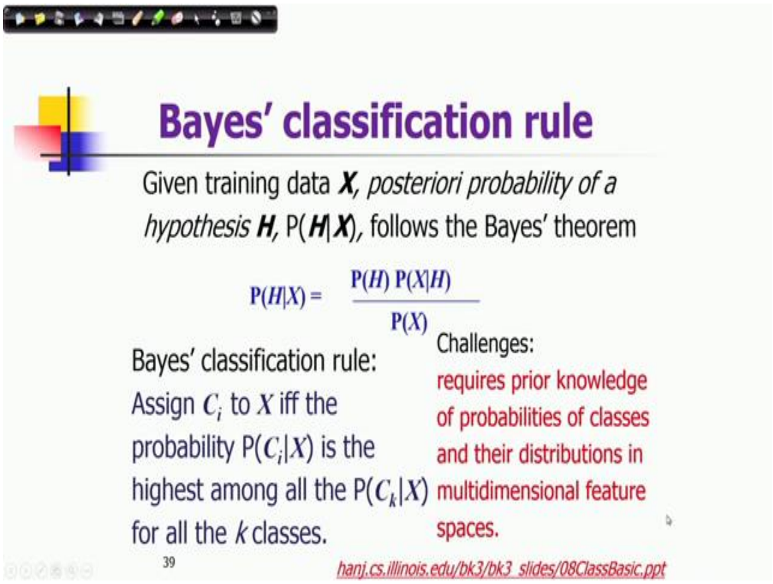
Now, Bayes theorem is derived from this observation that you can see that it has been shown that probability of a hypothesis H given a data X can be expressed as the ratio of this two quantity.

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

So, probability of H is the prior probability of the hypothesis based on our previous knowledge whereas, probability of X given H we call it as likelihood function. So, in the particular hypothesis if it is true what is the likelihood that X will occur under this hypothesis and probability of X is the unconditional probability of data.

So, it is a normalizing constant in this particular case which is ensuring posterior probability sum should be 1; that means, if X occurs there could be all different hypothesis which may generate this X. So, which are those hypothesis from which to which this occurring of this X belongs to are conditional. And, now if I sum all those possibilities that that should be equal to probability of X. So, posterior probability, probability of the hypothesis assigning the given a data.

(Refer Slide Time: 06:43)



Bayes' classification rule

Given training data ***X***, *posteriori probability of a hypothesis **H***, P(***H***|***X***), follows the Bayes' theorem

$$P(H|X) = \frac{P(H)\,P(X|H)}{P(X)}$$

Bayes' classification rule: Assign C_i to X iff the probability $P(C_i|X)$ is the highest among all the $P(C_k|X)$ for all the k classes.

Challenges: requires prior knowledge of probabilities of classes and their distributions in multidimensional feature spaces.

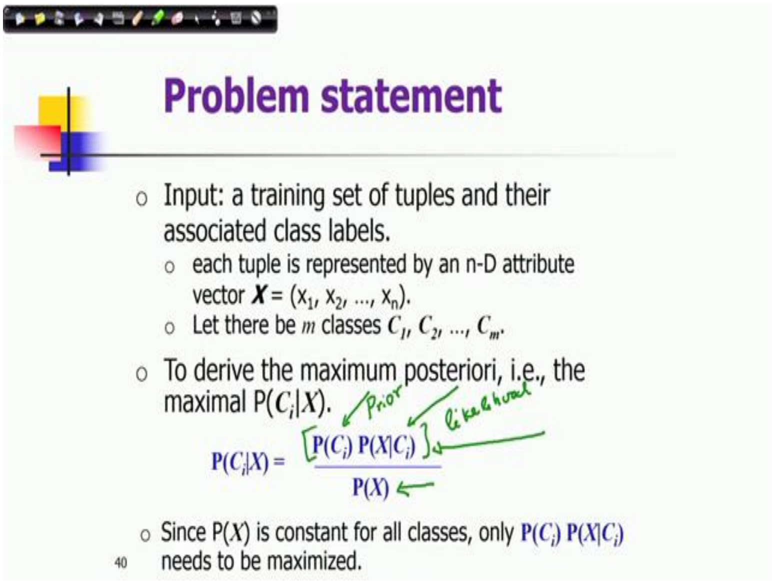
39

hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

So, given training data X posterior probability of hypothesis H, it is it can be computed by following this Bayes theorem. We will see that no it is it has significance, it is relevance in terms of designing a classifier; because here ah our objective is to maximize the probability of a class given the data and class is representing an hypothesis.

So, you need to compute this posterior probability here. So, that is what is the Bayesian classification rule or Bayes classification rule. It says that assigns Ci to X if and only if $P(C_i|X)$is the highest among all the other classes all, all other probabilities called all other classes. So, there are challenges of course, in this computation, it requires prior knowledge of probabilities of classes and their distributions in multidimensional feature space.

(Refer Slide Time: 07:39)



Problem statement

- Input: a training set of tuples and their associated class labels.
  - each tuple is represented by an n-D attribute vector $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$.
  - Let there be m classes $C_1, C_2, \ldots, C_m$.
- To derive the maximum posteriori, i.e., the maximal $P(C_i|X)$.

$$P(C_i|X) = \frac{\left[\overset{\text{Prior}}{P(C_i)}\,\overset{\text{likelihood}}{P(X|C_i)}\right]}{P(X)}$$

- Since $P(X)$ is constant for all classes, only $P(C_i)\,P(X|C_i)$ needs to be maximized.

40

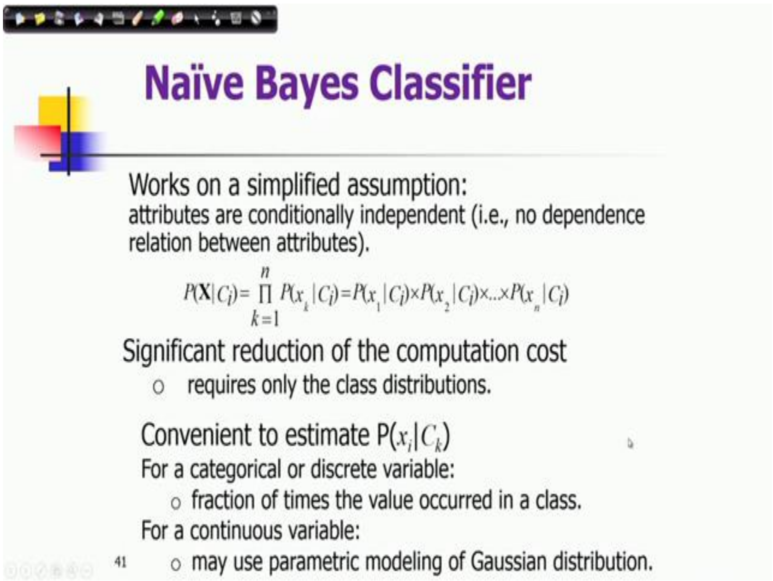
So, just to once again summarize the problem statement in a more concrete form, we have an input to the classifier. It is a training set of tuples and their associated class levels and each tuple is represented by an n dimensional attribute vector X. We have shown name of the attributes or the representation has been shown here. And, they are the data points what I am referring in an n dimensional space and let us consider there are m classes. So, as we discussed in the previous slide itself our objective is to derive the maximum posterior probabilities.

So, for a class the class which keeps the maximum posterior probability given the data so, that class should be no assigned to data. So, we can compute this posterior probability using Bayes theorem of conditional is in the conditional probability rule that you can compute the prior of class Ci. So, this is the prior probability. Now, this computation can be you can get it from the data point itself; some applied information can be obtained from other knowledge from a different source. But, given the data distributions among different classes we can compute that how many times for a data a data occurred or in a class, how many time class instances are occurring in the classes.

So, that fraction can estimate this prior and by observing the data distribution within the class itself, we can compute this probability density function or probability of X given Ci also. So, this called likelihood. So, this term is likelihood and you can see this numerator probability of X actually need not compute. Because, while comparing the values of probability of C i given X or posterior probabilities; this numerator is constant given the data X this numerator is constant for every classes. So, it is sufficient sorry this denominator is constant for every classes the probability of X.

So, it is sufficient if I compute just the numerator for it or computer these two quantities that would give me the probability no that would give me the basis of comparison. I am not in absolute since computing posterior probability, but in relatively I can compare them by computing those factors. So, this is what that is the summary only product of probability of C i and probability X given C i needs to be maximized that is the problem statement.

(Refer Slide Time: 10:33)



Naïve Bayes Classifier

Works on a simplified assumption:
attributes are conditionally independent (i.e., no dependence relation between attributes).

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

Significant reduction of the computation cost

- requires only the class distributions.

Convenient to estimate $P(x_i|C_k)$

For a categorical or discrete variable:

- fraction of times the value occurred in a class.

For a continuous variable:

- may use parametric modeling of Gaussian distribution.

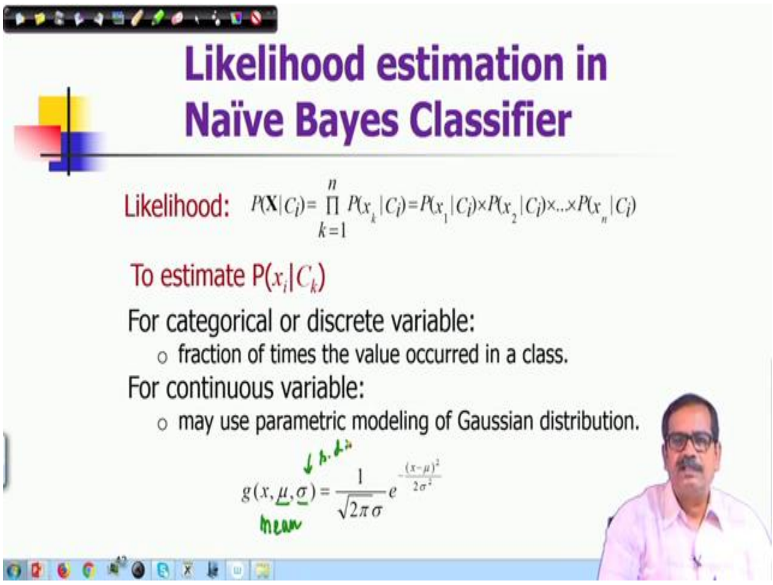
41

So, now will be discussing about Naive Bayes classifier. It works on a simplified assumption, where attributes are conditionally independent which means that there is no dependence relationship between attributes. And, in that if it is true then you can write the likelihood of X given a class C i as the product of all likelihood probabilities of individual attributes because, attributes are independent. So, that is the advantage in simply you can compute probability. So, all likelihood functions can be defined in one dimension instead of multi dimensional representation of probability density function that is the simplicity of Naive Bayes classifier.

Your handling of data becomes simpler or handling of multi dimensional data becomes simpler for modeling the probability density function. So, there is a significant reduction of the computational cost because, of this and it requires only class distribution that too in single dimensional feature space. So, it is convenient to estimate every attribute probability of every attribute given a particular class. And for example, for a categorical or discrete variables you can count how many times that values of attributes have occurred given all the occurrences of all other values of those attributes.

And, then you can find out the fraction of its occurrences over all occurrences and that can even estimate of this probability or if it is a continuous variable, then you can model it by any probability density function. Any kind of parametric modelling of density function you can do for example, Gaussian distribution you can and from there by estimating the parameters of Gaussian distribution you can compute this probability. So, you know that in Gaussian distribution there are only two parameters for a single dimensional, one dimensional feature space you have only its mean and standard deviation there.

(Refer Slide Time: 12:46)



Likelihood estimation in Naïve Bayes Classifier

Likelihood: $P(\mathbf{X}|C_i)=\prod_{k=1}^{n} P(x_k|C_i)=P(x_1|C_i)\times P(x_2|C_i)\times..\times P(x_n|C_i)$

To estimate $P(x_i|C_k)$

For categorical or discrete variable:

- fraction of times the value occurred in a class.

For continuous variable:

- may use parametric modeling of Gaussian distribution.

$$g(x,\underset{\text{mean}}{\underline{\mu}},\overset{\text{s.d.}}{\underline{\sigma}})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

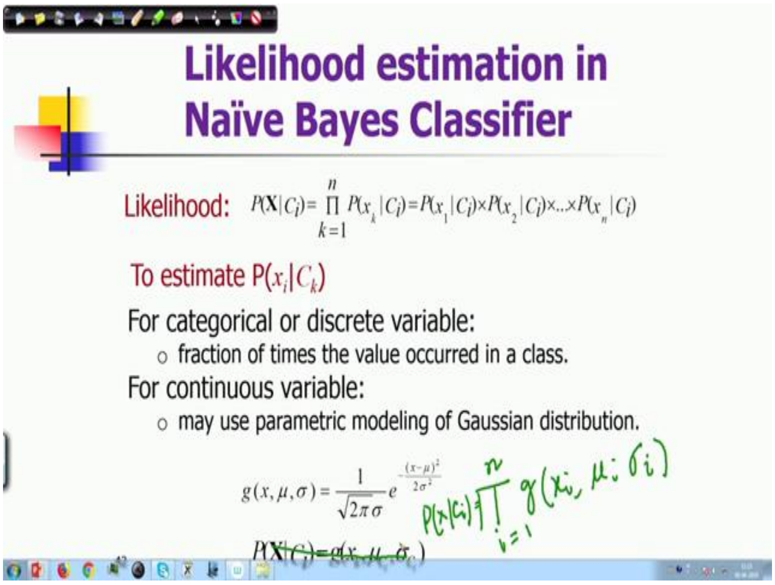
So, the likelihood is expressed in this form

$$P(X|C_i) = \prod_{k=1}^{n} P(x_i|C_i) = P(x_1|C_1) * P(x_2|C_2) *..* P(P(x_n|C_n)$$

So, there is the class condition probability of in attribute xi and as just I mentioned that no these are the methods by which you can do it. So, here this is just elaborating that how parametric modelling of Gaussian distribution could be done in the one dimensional feature space.

So, this is the distribution this is the probability density function of a variable X one dimensional random variable X which follows a Gaussian distribution and you can see that there are two parameters here. One parameter is mu which is called expectation of this distribution or you can in a simple layman’s term we can call it mean and this is sigma which is actually the standard deviation of this distribution. So, so given the data you can always estimate these values from mean and standard deviations, these parameters could be estimated.

(Refer Slide Time: 14:07)



Likelihood estimation in Naïve Bayes Classifier

Likelihood: $P(\mathbf{X}|C_i)=\prod_{k=1}^{n} P(x_k|C_i)=P(x_1|C_i)\times P(x_2|C_i)\times\ldots\times P(x_n|C_i)$

To estimate $P(x_i|C_k)$

For categorical or discrete variable:

- fraction of times the value occurred in a class.

For continuous variable:

- may use parametric modeling of Gaussian distribution.

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$P(x|C_i)=\prod_{i=1}^{n} g(x_i, \mu; \sigma_i)$

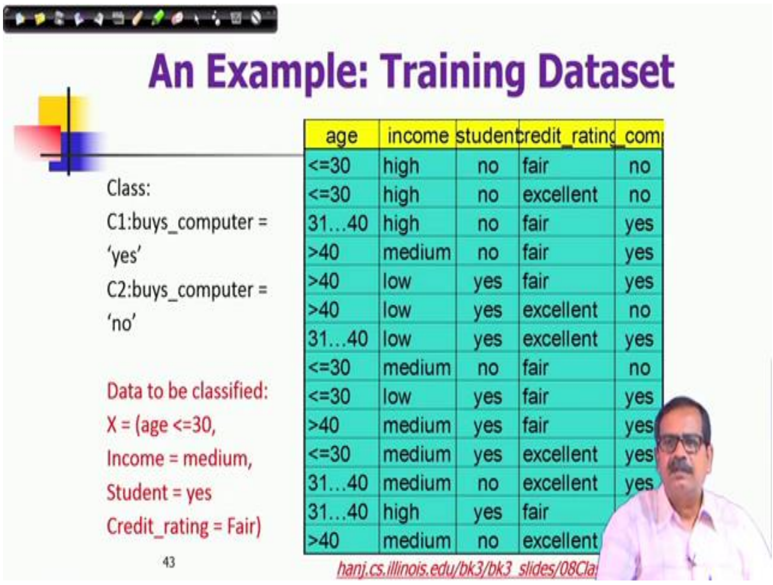
~~$P(\mathbf{X}|C_i)=g(x_k,\mu_{C_i},\sigma_{C_i})$~~

And so, this is what you are considering for it is the product of all this things which has been in this case actually I should write it in this way this is

$$P(X|C_i)=\prod_{i=1}^{n} g(x_i, \mathrm{µ}; \sigma_i)$$

So, this is what is the probability of this variable n dimensional feature given the class Ci this is this is how it should be written. So, there is some problem with this particular know expression.

(Refer Slide Time: 14:50)



An Example: Training Dataset

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data to be classified:
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	
>40	medium	no	excellent	

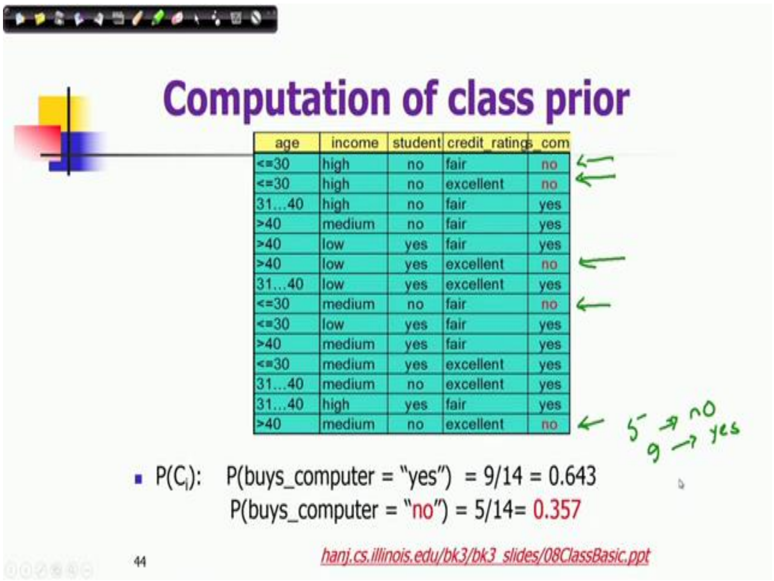
43

hanj.cs.illinois.edu/bk3/bk3_slides/08Cla

So, let me consider an example to explain to elaborate this computation, you can see this example here and here there is a data set which has been shown in this particular example. In this data set you have the statistics of the computer buyers I mean whether a person buys a computer or not, you have that kind of statistics. The person's age has been shown here, then the income and whether the person is a student and what is the credit rating and with that background whether that person particularly has bought computer or not. So, there are few distinct values for each variable.

So, these are the attributes age, income, student, credit rating and computer buyers. So, these are the attribute and computer buying is the class, you can see that there are two classes. One buys computer where the value would be yes and other class buys computer the value is and here the problem is that we need a data and a person has this kind of attributes or data associated with the person. What is the probability that the person buys a computer or whether not only probability actually in which class category, we can put this person that it is whether this person should be in the class of buys computer or not. So, this is what is this particular classification problem.

(Refer Slide Time: 16:43)



Computation of class prior

age	income	student	credit_ratings	_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

5 → no
9 → yes

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

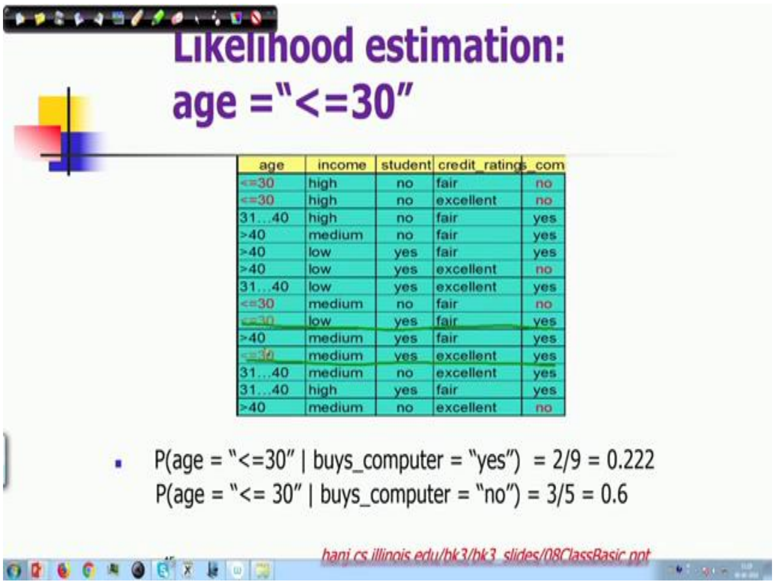
44

hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

So, first we need to compute the class prior in this case and you can see that here I have shown with color that how many instances of computer a person buying a computer is there and how many instances a person not buying a computer is there. So, these are the instances which in which cases a person did not buy a computer. So, there are 5 to 4, there are 5 such instances and if I count this number of record these are not actually 14 which means 9 instances there for yes.

So, that is why the probability of prior probability of class of buys computer equally yes would be 9 by 14, that is the fraction that is how this probability is considered and similarly the probability for the other class would be 5 by 14 ok. So, these are the 2 prior probabilities that we can we have estimated from this data, then will compute the likelihood of each attribute value for each class.

(Refer Slide Time: 18:11)



Likelihood estimation: age ="<=30"

age	income	student	credit_rating	s_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6

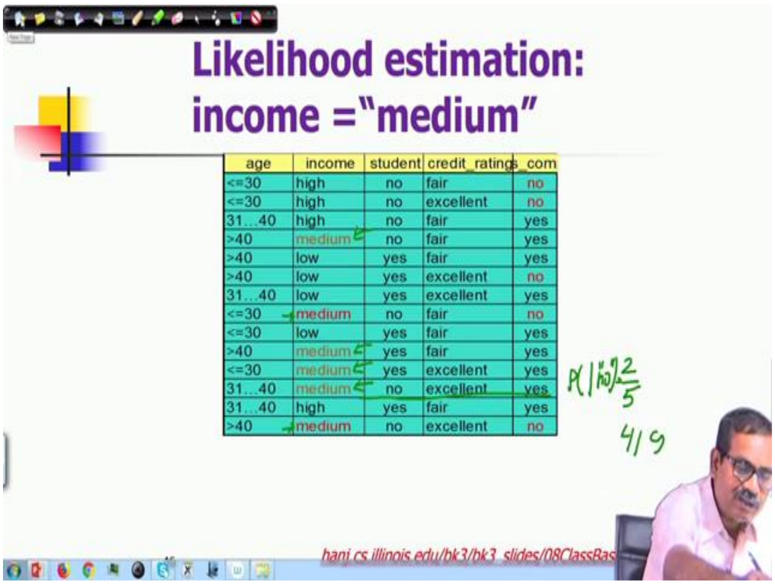
hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

So, now will consider the likelihood of age given the class of buying computer and also not buying computer. So, there what we need to consider here that for each class we have to see that distribution of attribute values; that means, how many times what is the fraction of times age less equals 30 occurs out of all possible values of attribute of age.

So, if I consider say no will see that this is one occurrences of this fraction, this is also another occurrences of this fraction, this is not for no this is for yes. So, there are 4 times this age less than less equals 30 occurs and actual occurrences of no is 5. So, the class likelihood probability of age equals 30 is quite high here, it should be 4 by 5 whereas, the other cases when the class of years. There we can see that there are so, many instances 1 2 3; that means, out of 9 cases only 1 instance of buying a so, this is called yes is there out of 9. So, which means it is likelihood probability is 1 by 9.

So, that is how we can compute these probabilities so, this is yes and this is yes. So, that I 4 is actually 3, it is not 1 is reduced. So, yes has 2 so, 2 by 9 instead of 1 by 9 because 1 9 is to I consider that was no in the class one no. So, it is 2 by 9 that correction should be there for this data 2 by 9 for yes and 3 by 5 for no. So, the probability is 0.222 likelihood probability given the class yes and 0.6 given the class no. So, let us proceed again for the other cases.

(Refer Slide Time: 20:30)



Likelihood estimation: income ="medium"

age	income	student	credit_ratings	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$P(\,|\,no) = \frac{2}{5}$

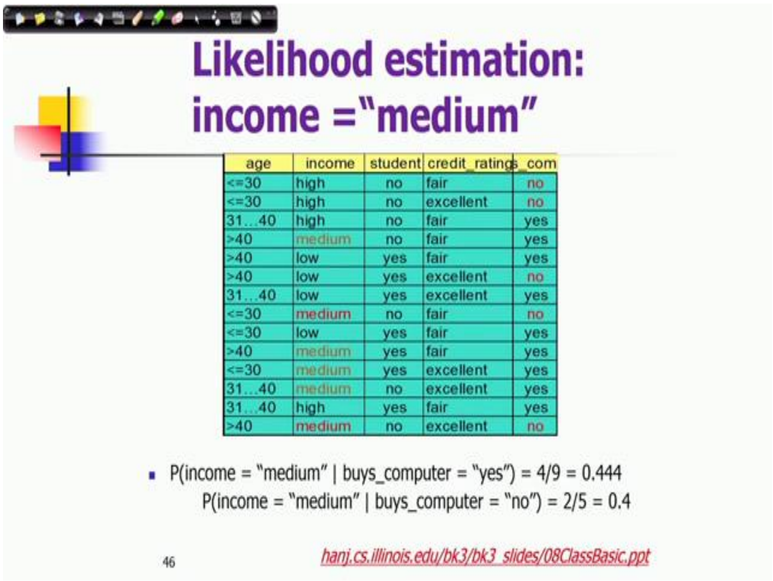
$4/9$

hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBas

So, now we are considering the income equals medium, once again will see how many times medium has occurred. So, 1 2 3 here actually I have given colors. So, from the color we can find out for no medium has occurred twice for no and we know that for the class no; that means, the persons those] they are not buying a computer there are 5 instances.

So, the likelihood probability should be 2 by 5 that is the probability of this attribute given the class no, in short if I write and for yes again it has occurred so, this is yes; so, 2 3 4 times. So, out of 9 no instances of yes 4 instances of were of income medium so, it should be 4 by 9. So, that is no value, let us check once again whether I made mistake in the case also let us see.

(Refer Slide Time: 21:48)



Likelihood estimation: income ="medium"

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

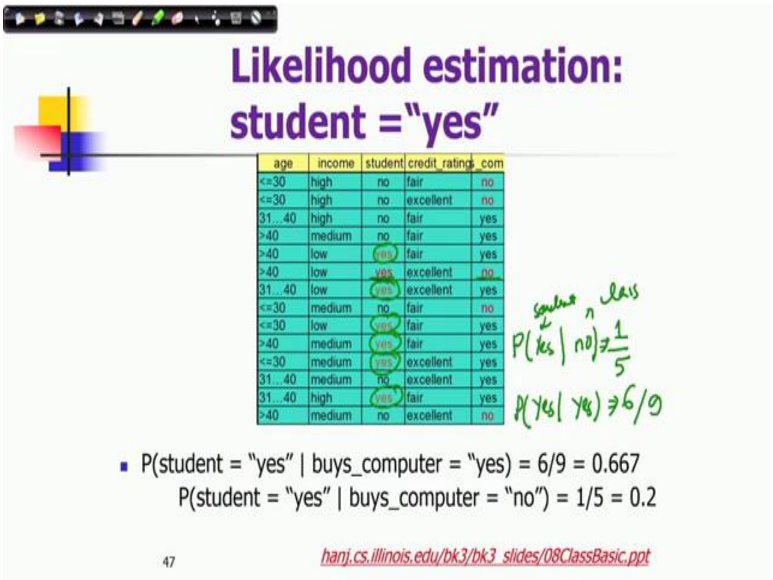
- P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4

46

hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

Yes 4 by 9 and 2 by 5; so, these are the values, these are the likelihood values.

(Refer Slide Time: 21:58)



Likelihood estimation: student ="yes"

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

student class

$P(yes \mid no) = \frac{1}{5}$

$P(yes \mid yes) = 6/9$

- P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2

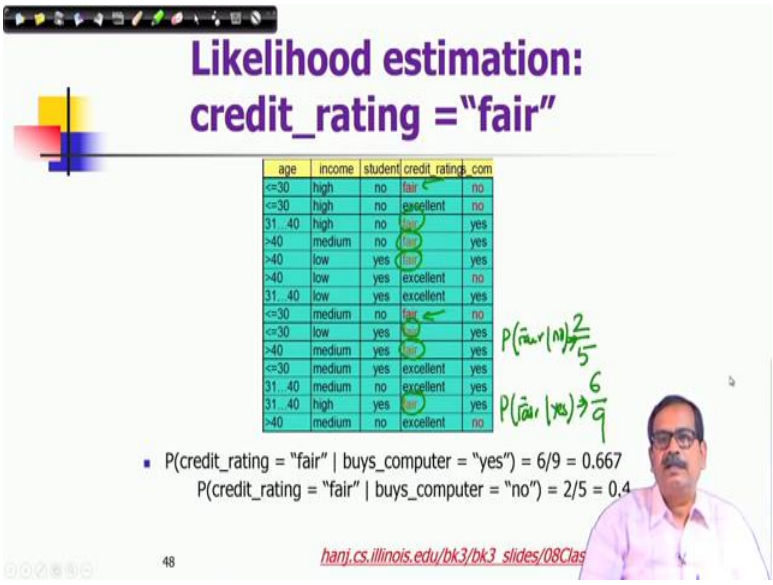
47

hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

Will proceed further; so, this time it is student, student equals yes. So, which means the person is a student or whether the person is not a student here. So, here we observe that for one instances in the class no the it is the student and which means that likelihood probability given the class no means not buying computer for a person is student is 1 by 5. So, this is probability yes let me write it in short, this yes means this is for student and this is for class.

So, this is 1 by 5 and the other cases we have shown that how many times student has occurred; that means, persons who have bought computer how many of them are student or there are 9 person who have brought computers in this data, out of them this is one, this is another, this is another, this is another quite a few. So, this is 6 6 by 9. So, probability a student yes given the class yes so, 6 by 9. So, it is let us see once again yeah. So, this is how the likelihood of student equals yes that attribute value given the classes could be computed.

(Refer Slide Time: 23:42)



Likelihood estimation: credit_rating ="fair"

age	income	student	credit_rating	s_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

$P(\text{fair} \mid \text{no}) \rightarrow \frac{2}{5}$

$P(\text{fair} \mid \text{yes}) \rightarrow \frac{6}{9}$

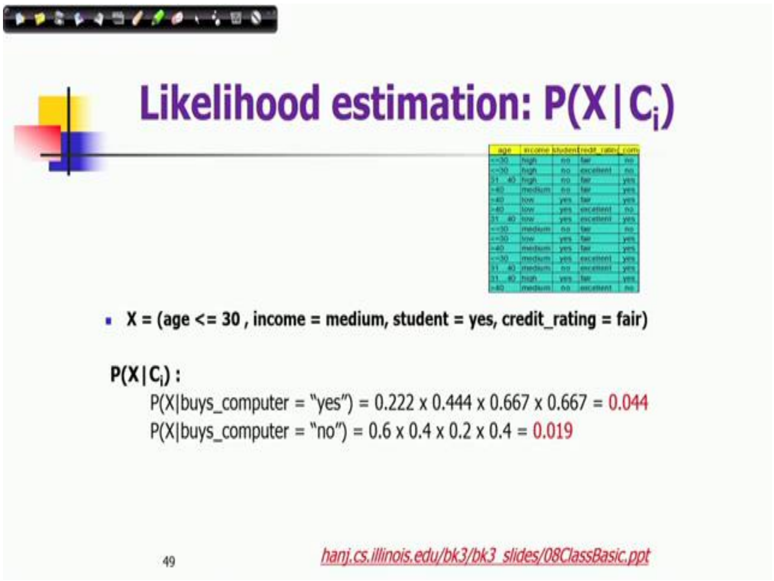
- P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

48

hanj.cs.illinois.edu/bk3/bk3_slides/08Clas

So, the next attribute will be considering credit rating where the given the data we have credit rating equals fair. So, in this case we have these are the instances when persons are not buying computer, they have these attribute value of fair credit rating which means that you will have 2 by 5 when it is fair given the class no and for yes you have 3 4 5 6 so 6. So, that is 6 by 9 that is a credit rating here given person’s buys a computer. So, these are the values that we can compute estimate likelihood estimates right. So, you have 6 by 9 by 2 by 5.

(Refer Slide Time: 24:47)



Likelihood estimation: $P(X|C_i)$

age	income	student	redit_ratin	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$:

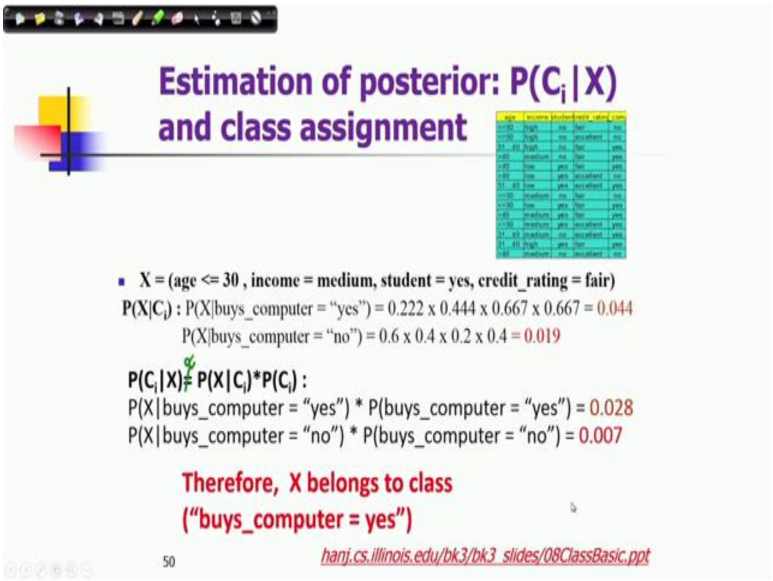
P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

49 hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

So, we have computed no all the likelihood values of age less equals 30, income equals medium, student equals yes and credit rating equals fair. So, the likelihood of X given any class any particular class C i for example, the class could be person buying computer means it is the product of likelihood of all those given classes. So, we can compute in this way finally, likelihood of this data given this two classes can be computed.

We can find out that the it the value is 0.044, when the class is yes that is buying a computer and value is 0.019 when class is you know buys the computer is no. And then finally, the posterior has to be computed which means each likelihood has to be computed no has to be multiplied with their prior. And then you get the posterior I mean proportional value of course, we are not dividing with probability of X; it is not the absolute posterior value, but you can get the proportional posterior value.

(Refer Slide Time: 25:51)



Estimation of posterior: $P(C_i|X)$ and class assignment

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(C_i|X) \overset{\propto}{=} P(X|C_i)*P(C_i)$:

P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028

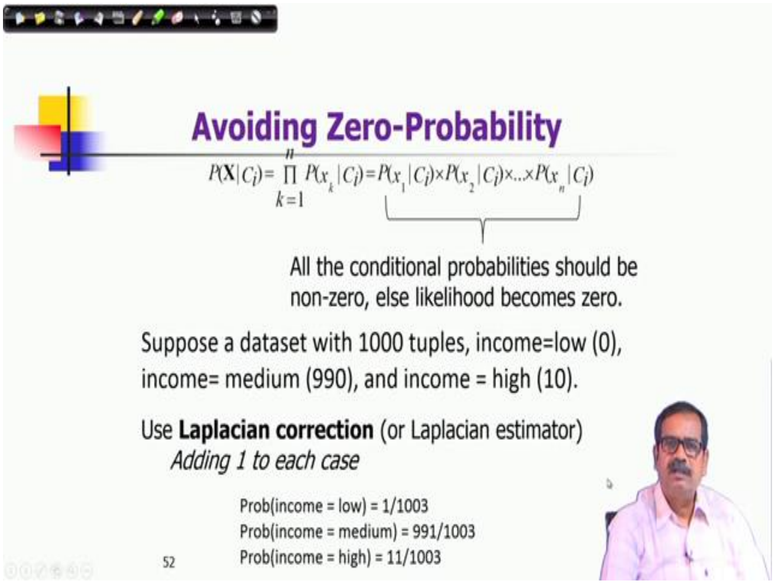
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

50 hanj.cs.illinois.edu/bk3/bk3_slides/08ClassBasic.ppt

So, this is how we are getting. So, in this case it is not exactly this, it is proportional in particular this constant when X is given it is a proportional value. So, the posterior of this I should say once again this is yeah, this is the product of this two that is 0.028. So, this value is 0.028 and the other value is 0.007. So, naturally 0.028 is the know greater value and this is out of these two classes this is maximum. So, which means that I can assign this attribute this variable to a class of computer buying class. So, that is the inferencing so, therefore, X belongs to class of buys computer equals yes.

(Refer Slide Time: 26:52)



Avoiding Zero-Probability

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = \underbrace{P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)}$$

All the conditional probabilities should be non-zero, else likelihood becomes zero.

Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10).

Use **Laplacian correction** (or Laplacian estimator)

Adding 1 to each case

Prob(income = low) = 1/1003

Prob(income = medium) = 991/1003

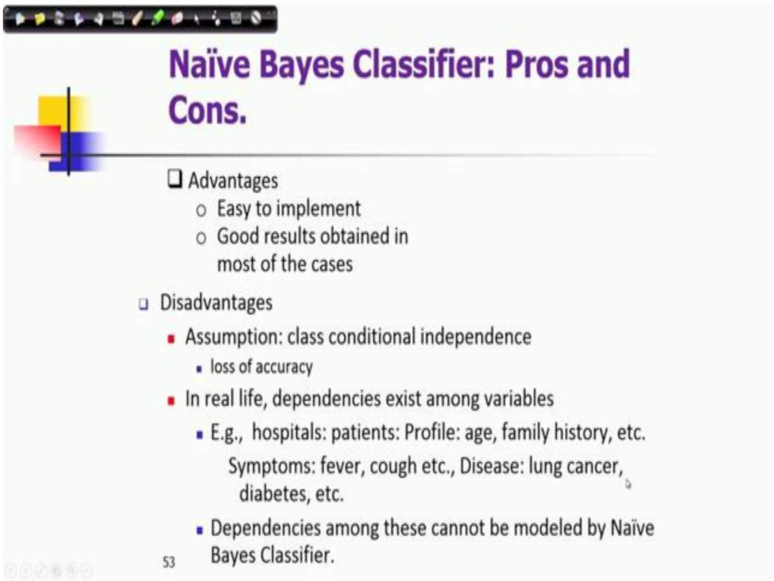
Prob(income = high) = 11/1003

52

So, this is the example of learning. So, one of the problem of the computations of this likelihood is that if there is any zero probability; suppose there is no instance at all. It may happen in your data point for a particular class then the likelihood becomes zero. So, or very small of like that is one of the problem. So, to avoid this we need to do certain because, you know it is a observation and observation are noisy. So, we do not except that it would be absolutely zero for every class the likelihood.

So, but it is low, it indicate it should be low ah. So, in that case so, this is one example that now data set with 1000 tuple, income equal to low 0, income equal to 990 and income equal to high 10. So, we use a correction which is called Laplacian correction or Laplacian estimator by adding one to each case. So, instead of 0 we consider the value is 1. So, you will get a very low probability and we avoid this problem of zero probability in computing likelihood.

(Refer Slide Time: 28:00)



Naïve Bayes Classifier: Pros and Cons.

❑ Advantages
- Easy to implement
- Good results obtained in most of the cases

❑ Disadvantages
- Assumption: class conditional independence
  - loss of accuracy
- In real life, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history, etc. Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
  - Dependencies among these cannot be modeled by Naïve Bayes Classifier.

53

So, what are the pros and cons of Naive Bayes classifier? It is the advantage is that it is easy to implement, it good results obtained in most of the cases. There are disadvantages like first it is the assumption is it that class conditional independence which may not be true for a realistic data and if it is not true, then there will lose also accuracy. So, that is what in real life dependencies exist among those attributes or variables.

For examples: hospitals, patients, profile, age, family, history etcetera; symptoms: fever, cough etcetera, disease: lung cancer, diabetes. So, many of the information are related and this cannot be modeled by Naive Bayes classifier. So, let me take a break here, we will continue this discussion on different approaches of classifications in in the lectures so.

Thank you very much for your listening.

Keywords: Classification, Bayesian inference, likelihood estimation, clustering