

Computer Vision and Image Processing - Fundamentals and Applications
Professor Doctor M K Bhuyan
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Lecture 33
Introduction to Machine Learning - IV

Welcome to NPTEL MOOCs course on Computer Vision and Image Processing - Fundamentals and Applications. In my last class I discussed the concept of Bayesian decision theory. In Bayesian decision theory, I have to estimate the probability of w_j given x that I have to determine. So for this I need two information one is the probability of x given w_j that is called likelihood and also the probability of w_j that is called a prior.

So, this information I need the probability of x given w_j that is the class conditional density or the likelihood and suppose, this information is available that means the density of the likelihood function or the class conditional density is available then this is called the parametric method. That means the density form the likelihood function or the class conditional density is available, but the parameters are not available.

So, suppose if I consider a Gaussian density, in Gaussian density there are two parameters one is the mean and other one is the variance and if I consider high dimensional case, then it is the mean vector and the covariance matrix. So, in case of the parametric method this density form is available I know the density of probability of x given w_j . So that information is available but I do not know about the parameters.

So I have to estimate the parameters. So there are two methods, very popular methods; one is the maximum likelihood estimation, another one is the Bayesian estimation. So by using these two techniques I can determine the parameters, the parameters are mean and the covariance. Another case is suppose, the density form is not available, the density form class conditional density is not available that is the probability of x given w_j . So that is not available we have to estimate the density.

So there are two popular techniques one is called the Parzen-Window technique and another one is called a k nearest neighbor technique. So by using these two techniques I can determine the density, the density of probability of x given w_j that is the likelihood. So in this class I will discuss the parametric methods first I will discuss the maximum likelihood estimation and after this I will discuss the Bayesian estimation. After this I will discuss the

non-parametric methods one is the Parzen-Window technique and another one is the k nearest neighbor technique. So let us discuss about this parametric and non-parametric methods.

(Refer Slide Time: 3:19)

Parameter Estimation

Introduction

- Data availability in a Bayesian framework
 - We could design an optimal classifier if we knew:
 - $P(\omega_j)$ (priors) ✓
 - $p(\mathbf{x} | \omega_j)$ (class-conditional densities) ✓
 - Unfortunately, we rarely have this complete information!
- Design a classifier from a training sample
 - No problem with prior estimation
 - Samples are often too small for class-conditional estimation (large dimension of feature space!)

Parametric approach → pdf form is known,

Non-parametric approach → Density estimated

x_n belongs to ω_j → Supervised Learning

$$P(\omega_j | \mathbf{x}) = \frac{P(\mathbf{x} | \omega_j) P(\omega_j)}{\sum_{j=1}^c P(\mathbf{x} | \omega_j) P(\omega_j)}$$

$$P(\mathbf{x} | \omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$P(\mathbf{x} | \omega_j) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right)$$

$\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$

θ_{i1}, θ_{i2}

$j = 1, 2, \dots, c$

So the first one is the parameter estimation. So I told you that what is the parametric method in case of the Bayesian decision theory. So I have to determine this the probability of ω_j given \mathbf{x} that I have to determine. And that is nothing but the probability of \mathbf{x} given ω_j that is the likelihood and the probability of ω_j and j is equal to 1 to c probability of \mathbf{x} given probability of ω_j . So in this case what information is available?

So in this case if I want to determine the probability of ω_j given \mathbf{x} that is the posterior probability. So suppose the density form of the class conditional density is available. So suppose the density form is available. So density form is suppose, it is the parametric form in case of the normal density I have two parameters one is the mean vector another one is the covariance matrix.

So this parametric form is available. If you see this formula, so for calculating probability of ω_j given \mathbf{x} , so what information I need? I need the information of probability of \mathbf{x} given ω_j that input information I need, that is nothing but the class conditional density. And the parameters are suppose the mean vector and the covariance matrix. So this information also I need, the mean vector and the covariance matrix also I need.

So that means I have two parameters, the parameters are θ_{i1} and θ_{i2} suppose. So two parameters I need and also I need the information of how many classes. So classes is 1 to c number of classes. So here you can see this is j , j is equal to 1 to c . So c number of classes, so

this information I need for calculating probability of w_j given x . So one is the class conditional density that is the probability of x given w_j that information I need.

And I have two parameters the one is the mean vector, another one is the covariance matrix and also I need one another information that is number of classes j is equal to 1 to c . So that information I need to calculate the probability of w_j given x . So this information is not available directly. So for this actually we have the training samples, training samples for all the classes that is nothing but the supervised training.

So for each and every classes we have the training samples. And after these from these training samples, we have to do the estimation, estimation is nothing but we have to estimate the parameters. So parameters are the mean and the covariance. So we have to estimate the parameters. And in this case in case of the parametric approach the PDF form is known that means, I know the PDF of the class conditional density in case of the parametric form.

So I can write this in parametric approach that is the PDF form known. But the parameters I have to determine the PDF form is known but the parameters I have to determine I have to estimate. That is a parametric approach. What is the non-parametric approach? In case of a non-parametric approach, the PDF form is not available. That means I have to estimate the density that is nothing but the density estimation.

So in case of the non-parametric method approach the density form is not available. So we have to estimate the density. So for this we need the training samples, the training samples for all the classes. So if I consider the training sample suppose x_n that belongs to a particular class, suppose class is w_j , that is available. That is the training samples are available corresponding to a particular class and that is nothing but the supervised learning.

So for estimating the probability of w_j given x that is the posterior probability, I need the information of the prior, the priori probability, probability of w_i and also I need the information of the probability of x given w_i that is the class conditional density. So I need this information and also I need the information of number of classes. So we have c number of classes. So based on this I have explained two approaches one is the parametric approach another one is the non-parametric approach.

(Refer Slide Time: 9:13)

A priori information about the problem

Normality of $p(\mathbf{x} | \omega_i)$

$p(\mathbf{x} | \omega_i) \sim N(\underline{\mu}_i, \underline{\Sigma}_i)$

Characterized by 2 parameters

$\theta_1 \theta_2$

So in this case what I am considering suppose the density form is known that is the class conditional density. So density form is known but I have to estimate the parameters. So in this case there are two parameters one is the mean, another one is the covariance. So one is a mean vector, another one is the covariance. So suppose this is θ_1 and this is θ_2 .

So for this I can apply two popular techniques one is the maximum likelihood estimation another one is the Bayesian estimation. In case of a non-parametric method the density form is not available but I have to estimate the density.

(Refer Slide Time: 9:46)

Estimation techniques

Maximum-Likelihood (ML) and the Bayesian estimations

Results are nearly identical, but the approaches are different

So you can see so for estimation techniques that is the parameter estimation techniques, I may consider this popular method that is the maximum likelihood estimation. And also I may

consider the Bayesian estimation. So in both the cases results are nearly identical but the approaches are different. In case of the Bayesian estimation, the computational complexity is more as compared to maximum likelihood estimation.

In case of the Bayesian estimation I have to determine the multi-dimensional integration. So I will explain this, but in in case of the maximum likelihood estimation I have to determine the differentiation. So that is why if I consider the computational complexity then Bayesian estimation if I consider that it is more computational complex as compared to maximum likelihood estimation.

(Refer Slide Time: 10:43)

The slide features a dark blue header with the text "Maximum likelihood". Below the header, there are three bullet points. To the right of the first bullet point, there is a handwritten red note that says "maximize" above the mathematical expression $P(D|\theta)$.

- Parameters in ML estimation are fixed but unknown! $P(D|\theta)$
- Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Parameters are chosen in a way that they best support/ describe the training data.

So in case of the maximum likelihood estimation, parameters in maximum likelihood estimation are fixed but not known. And in this case what I have to consider, I have to maximize the probability of obtaining a given set suppose H is the training sample or maybe I can consider probability of D , D is the training set given θ . So I have to maximize this probability of D given θ that θ is fixed.

So parameter estimation that maximize a likelihood function, that I can consider, maximize the probability of obtaining the given training set. So I have to maximize this. So maximize the probability of D given θ in case of the maximum likelihood estimation. That means maximizing the probability of obtaining the given training set. And in this case the θ is fixed a θ is the parameter vector. Now let us see the mathematics behind the maximum likelihood estimation. So what is the maximum likelihood estimation?

(Refer Slide Time: 11:49)

$$P(w_j|x) = \frac{P(x|w_j)P(w_j)}{\sum_{j=1}^c P(x|w_j)P(w_j)}$$

$$P(x|w_j) \text{ Parametric form } \sim N(\mu_j, \Sigma_j)$$

$$\theta_j \Rightarrow \mu_j, \Sigma_j$$

$$P(x|w_j, \theta_j)$$

$$P(\theta) = \prod_{k=1}^n P(x_k|\theta)$$

Maximize $P(\theta)$

Training algorithm: $D_1, D_2, \dots, D_c \rightarrow \theta_j$

Samples: $\mu_1, \mu_2, \Sigma_1, \Sigma_2 \rightarrow w_1, w_2$

So in case of the Bayesian decision theory, we have to estimate this the probability of w_j given x we have to estimate. And for this you can see I need the information of this x given w_j that is the class conditional density. And also the priori and if you see the this is the evidence j is equal to 1 to c probability of x given w_j probability of w_j .

So in this case the probability of x given w_j so this parametric form, the parametric form that is available. That is the suppose if I considered a normal distribution that information is available that is the parametric form available. But the parameters are not available, but the parameters we have to estimate. So in case of a normal distribution I have two parameters so one is the mean, another one is the covariance.

So if you consider this one the θ_j that is the parameter vector. So I have two parameters one is the mean vector and another one is the covariance matrix that I have to determine. So if I want to show the maximum likelihood estimation. So suppose we have the training samples the training sets D_1, D_2, \dots, D_c that means we have the training sets and we have the training algorithm is available.

And we have to estimate the parameters, we have to estimate the parameter vector and what information is available? I know the density form that is the class conditional densities available. This dependence on the training set I can write like this, the dependence on the training set I can write like this the probability of x given w_j θ_j that is the dependence on θ_j . That is, I am considering the class conditional density and I am showing the dependence on θ_j .

So the problem is to determine unknown parameter vectors. So that means I have to determine θ_1, θ_2 from the information of that training dataset. Because we have the training data set for all the classes and from this training dataset, I have to determine the parameters. So that is the parameter vector I have to estimate. And in this case I am considering the independent data set.

So what do you mean by independent data set? So suppose I am considering the data set D1 that is for the class w1. Similarly, if I consider another data set that is D2 and that is for the class w2. Like this if I consider another class suppose, so if I consider the data set suppose the training data set is Di and suppose I have the samples, the samples are x1, x2 like this. These are samples and corresponding to these the class is wi.

And suppose the class is, another class is wj. So this training data set Di is for the class wi. This is not for class wj, this is not for a class wj. The training data set Di is for class wi, it is not for the class wj. That is the supervised training, that is the concept of the supervised training. So here x1 x2 these are the samples, these are samples of the data set, the data set is Di and the samples are drawn independently.

So I have the training data set that is D1 D2 Dc then what is the probability of D given theta that is nothing but this product. So suppose I have this training samples x1 x2 up to xn suppose xn, so xn and theta. So I have to maximize the probability of the given theta. So that I have to maximize. So that is the concept of the maximum likelihood estimation.

(Refer Slide Time: 17:22)

The image contains handwritten mathematical notes on a light gray background. The notes are organized into several sections:

- Top Left:**
 - Probability density function: $p(x|\theta)$
 - Likelihood function: $L(\theta) = \prod_{k=1}^n p(x_k|\theta)$
 - Log-likelihood function: $\ln L(\theta) = \sum_{k=1}^n \ln p(x_k|\theta)$
 - Setting the derivative to zero: $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$
- Top Right:**
 - MAP equation: $MAP = L(\theta) \cdot P(\theta)$
 - Maximize: $P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$
 - Flat prior: $ML = MAP$
- Middle:**
 - Gradient vector: $\nabla_{\theta} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \frac{\partial}{\partial \theta_2} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$
 - Parameter vector: $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_p]^T$
 - Sample mean: $\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k \rightarrow \text{AM of the samples.}$
- Bottom Left:**
 - Global max in the parameter space

So that means, again I am writing so I have to maximize the probability of D given theta, the probability of obtaining a given training set for a parameter vector, the parameter vector is theta. Now I am defining one likelihood function, the likelihood function that is the log likelihood function I am considering. So $l(\theta)$ that is the log likelihood function that I am considering, that is the log likelihood function. And after this what I am considering?

I am considering the differentiation of this likelihood function and suppose I am getting the approximate value. So I am just doing the differentiation because I have to maximize this one. So it is equal to 0 because I have to find a maximum of this. So what is this operator? So if I consider this operator, this is the partial derivative I am considering with respect to θ_1 like this which respect to θ_2 .

So suppose I have these parameters $\theta_1, \theta_2, \dots, \theta_p$ like this. So I have to find a global maxima. Maximum I have to find in the parameter space, parametric space. So global maxima I have to find in the parametric space. So from this equation if I see this equation so from this equation I can determine the parameters, the parameters are like this, this is a parameter vector. So I can determine θ_1 , I can determine θ_2 .

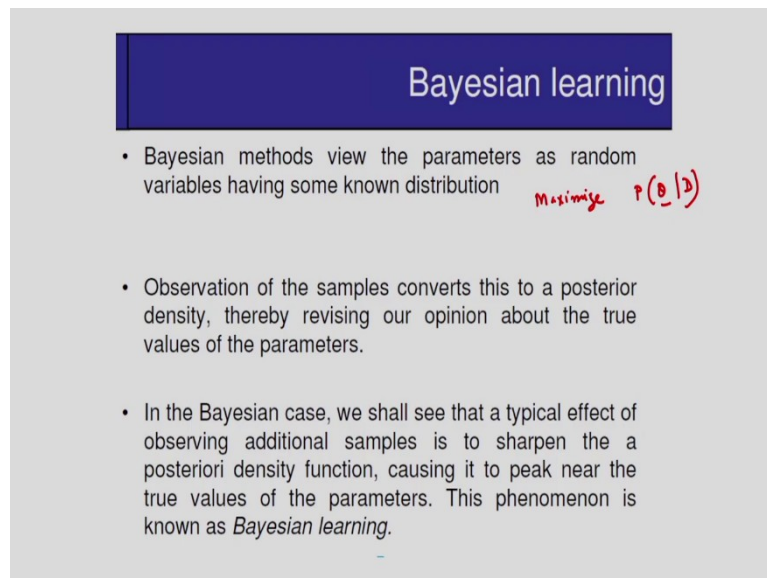
So in case of a normal distribution this is the mean vector and also this is the covariance matrix I can determine. So what is theta? The theta is nothing but this is the parameter vector $\theta_1, \theta_2, \dots, \theta_p$ transpose. So by considering this equation I can determine the parameters. So for example I can determine the mean of the samples. So it is the mean of the samples you can determine by this equation.

So it is the mean is something like this, this is nothing but the arithmetic mean of the samples. The arithmetic mean of the samples. And also I can consider the maximum a posterior probability that is the estimation maximum a posterior estimation that is the MAP I can consider. So MAP is nothing but the likelihood function I am considering that is the theta that is the likelihood function.

And the priori information also I am considering $p(\theta)$. So I have to maximize the probability of theta x, probability of x theta $p(\theta|x)$. So for flat prior, so if I consider the flat prior the maximum likelihood estimation will be same as that of the MAP maximum a posterior probability estimation. So if I consider a priori is flat, so this is the priori information. So suppose for the flat prior, the maximum likelihood estimation is equal to MAP, the MAP estimation.

So this is the basic concept of maximum likelihood estimation. I am not explaining how to determine the parameters but based on this equation, so if you see this equation I can determine the parameters all the parameters. So this is the fundamental concept of maximum likelihood estimation.

(Refer Slide Time: 22:27)



The slide is titled "Bayesian learning" in a dark blue box. It contains three bullet points:

- Bayesian methods view the parameters as random variables having some known distribution *Maximize $P(\theta|D)$*
- Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters.
- In the Bayesian case, we shall see that a typical effect of observing additional samples is to sharpen the a posteriori density function, causing it to peak near the true values of the parameters. This phenomenon is known as *Bayesian learning*.

The next one is the Bayesian estimation. In case of the Bayesian estimation we can consider the parameters as random variable having some known distribution. In case of the Bayesian learning I am repeating this, what I am considering the parameters are random variable with some known a priori distribution that is available that the information is available. These training samples allows the conversion of the priori information into posterior density.

So the training samples allows conversion of priori information into a posterior information. So in case of the Bayesian estimation, in case of the Bayesian learning what I have to consider I have to maximize the probability of theta given D. So D is the training set and theta is the parameter vector in this case the theta is random variable. So that means we have to determine the density that approximate an impulse. So briefly I will explain the mathematical formulation of the Bayesian learning. So what is Bayesian learning you can see.

(Refer Slide Time: 23:46)

$$P(w_i | x, D) = \frac{P(x | w_i, D) P(w_i, D)}{\sum_{j=1}^c P(x | w_j, D) P(w_j, D)}$$

Prior information is known.

$$P(w_i, D) = P(w_i)$$

$$P(w_i | x, D) = \frac{P(x | w_i, D) P(w_i)}{\sum_{j=1}^c P(x | w_j, D) P(w_j)}$$

Determine $\rightarrow P(x | D)$

$P(x) \rightarrow$ unknown

Parametric form $P(x | \theta) \rightarrow$ known.

$P(\theta) \rightarrow$ known $\xrightarrow{D} P(\theta | D)$

$$P(x | D) = \int P(x | \theta) P(\theta | D) d\theta$$

$$P(x | D) = \int P(x, \theta | D) d\theta$$

$$P(x | D) = \int P(x | \theta) P(\theta | D) d\theta$$

Maximize $P(\theta | D)$

$$P(x | D) \approx P(x | \hat{\theta})$$

Taking average of $P(x | \theta)$

In case of the Bayesian learning, again, I am showing the Bayesian decision theory that is the Bayes law I am writing, the probability of w_i given x, D ; D is the training data set. So that means I am considering the dependence on the training data set that is the dependence on D is equal to probability of $x / w_i, D$ that is the class conditional density. And I am writing D because this is the dependence on the training data set i is equal to 1 to c probability of x given w_i, D probability of w_i given D .

But in this case the priori information is known, the priori information is known. So I can write like this now the probability of w_i given x, D probability of x given w_i, D the probability of w_i and summation i equal to 1 to c probability of x, w_j, D probability of w_j probability of w_i and probability of w_i . So I can write like this. So in case of the Bayesian estimation I have to determine, what I have to determine? I have to determine the probability of x given D that I have to determine.

So what information is available the information is the probability of x that is not available this is unknown, that is not that is unknown and a parametric form is known, the parametric form is known. So that means the probability of x given θ that is known. And also the probability of θ and that is known. So this the training data set converts the priori information into the posterior information.

The $P(\theta)$ is available that is known but the training data set converts the priori information into the posterior information. So I have to determine the probability of x given D I have to determine that is nothing but probability of x given θ and probability of θ / D . So this actually I am obtaining like this the probability of x given D is nothing but probability of x given θ / $P(D; \theta)$ so that is available.

So this probability of x given D is nothing but probability of x given θ / $P(D; \theta)$ I can write like this. So that means from this actually I am getting this one, from this I am getting this one probability of x given D . Now because this is the important equation. So probability of x given D is nothing but probability of x given θ / $P(D; \theta)$ that is available. So I have to maximize the probability of θ / $P(D; \theta)$.

So maximize probability of θ / $P(D; \theta)$ I have to maximize and in this case the probability of x given D will be approximately equal to probability of x this is the approximate value of θ that is the parameter vector approximate value of θ . So that means I am taking average of probability $p(x; \theta)$. Because in this case I have to maximize this. So that means I have to maximize this, so it will be a Dirac delta function corresponding to the estimated value of θ .

So this is the approximate value of θ . So I have to maximize this probability of θ given D I have to maximize. So that means the probability of x given D is nothing but probability of x given θ . That means I am taking the average of $p(x; \theta)$. So this integration, if you see this integration that means I am taking the average and this is a multi-dimensional integration. So it is very difficult to determine because it is a computationally complex to determine the multi-dimensional integration.

So something like the Monte Carlo simulation technique I can use to determine this integration. And here you can see with the help of Bayesian estimation I can determine the probability of x given D . So that I can determine and after this I can determine the parameters, the parameters are θ , the parameter vector are θ . So this is the basic concept of the Bayesian estimation. So both the methods the maximum likelihood estimation and the Bayesian estimation they will give almost similar results.

So briefly I discussed the concept of the maximum likelihood estimation and the Bayesian estimation. So for more detail you can see and the book Pattern Classification by Duda and Hart, that book also you can see for this maximum likelihood estimation and the Bayesian estimation.

(Refer Slide Time: 32:06)

Non parametric Techniques

- Parzen Windows
- Nearest Neighbor classifier

Now I will discuss the concept of the non-parametric methods. In case of non-parametric methods that already I have explained that is the density form is not available but we have to estimate the density. And there are two techniques one is the Parzen-Window technique and another one is k nearest neighbor technique, these two techniques I can consider.

(Refer Slide Time: 32:28)

Non parametric generative classifiers

- Generative models assume data to come from a probability density function. $p(x|y)$
- Parametric learning assumes we know the form of the underlying density function, which is often not true in real applications.
- All parametric densities are either unimodal (have a single local maximum), such as a Gaussian distribution, or multi-modal (example: GMMs)

So in case of generative models, we assume that the data to come from a probability density function. That means we have this information the probability of x given w_j for the generative models. But in this case sometimes this density is not available that information is not available. So we have to estimate the density. So how to estimate the density? So I will

explain these two techniques one is the Parzen-Window technique another one is the k nearest neighbor technique.

(Refer Slide Time: 33:03)

- Nonparametric procedures can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known.
- They are data-driven (or are estimated from the data).
- There are two types of nonparametric methods:
 - Estimating $p(\mathbf{x}|\omega_j)$ \rightarrow Parzen Window
 - Bypass class conditional probability estimation and go directly to *a-posteriori* probability estimation, $P(\omega_j|\mathbf{x}) \rightarrow$ Nearest neighbor

So in case of the non-parametric procedure what we can consider, we have to estimate the probability of x given w_j that means we have to estimate this. But in case of the k nearest neighbor technique we can directly estimate this density. Because what is the ultimate objective? The ultimate objective is we have to determine the probability of w_j given x that is the posterior probability I have to determine.

(Refer Slide Time: 33:31)

- The basic idea in density estimation is that a vector, \mathbf{x} , will fall in a region R with probability:
$$P = \int_R p(\mathbf{x}') d\mathbf{x}'$$
- P is a smoothed or averaged version of the density function $p(\mathbf{x})$.

So the basic idea in density estimation is there a vector \mathbf{x} will fall in a region R with a probability, that probability I am determining. P is a smooth or average version of the density function. So density function is $p(\mathbf{x})$.

(Refer Slide Time: 33:47)

- Suppose n samples are drawn independently and identically distributed (i.i.d.) according to $p(\mathbf{x})$. The probability that k of these n fall in R is given by:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}$$

- The expected value for k is: $E[k] = nP$.
- The ML estimate, $\max_{\theta} (P_k | \theta)$, is $\hat{\theta} = \frac{k}{n} \cong P$.
- Therefore, with large number of samples, the ratio k/n is a good estimate for the probability P and hence for the density function $p(\mathbf{x})$.

Suppose n samples are drawn independently and identically distributed that is the i.i.d according to $p(\mathbf{x})$ that probability $p(\mathbf{x})$ the probability that k of this n the k number of samples out of n number of samples fall in the region R is given by this distribution. That is the probability of p_k I can determine that is nothing but a binomial distribution.

So what I am considering the probability that the k number of samples out of n fall in the region R . And from this I can determine the expected value of k . So I can determine expected value of k $E[k]$ is equal to n into P . And if I consider the maximum likelihood estimation so by the maximum likelihood estimation I can determine or I can maximize the probability of probability p_k for the given θ that I can maximize.

And corresponding to this I will be getting the probability the probability is nothing but p is equal to k divided by n . So n is the total number of samples I am considering. So therefore with large number of samples the ratio k divided by n is a good estimate for the probability P and hence for the density function the density function is $p(\mathbf{x})$. So that means this ratio k by n gives the estimate of the probability the probability is P . So here n is the total number of samples and k is the number of samples within this particular region that I am considering. And that this ratio k divided by n it gives the probability the probability is P .

(Refer Slide Time: 35:33)

- Assume $p(\mathbf{x})$ is continuous and that the region R is so small that $p(\mathbf{x})$ does not vary significantly within it. We can write:

$$\int_R p(\mathbf{x}') d\mathbf{x}' \cong p(\mathbf{x})V$$

where \mathbf{x} is a point within R and V the volume enclosed by R , and

$$p(\mathbf{x}) \cong \frac{k/n}{V}$$

- Suppose n samples are drawn independently and identically distributed (i.i.d.) according to $p(\mathbf{x})$. The probability that k of these n fall in R is given by:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}$$

The expected value for k is: $E[k] = nP$.

- The ML estimate, $\max_{\theta} (P_k | \theta)$, is $\hat{\theta} = \frac{k}{n} \cong P$

- Therefore, with large number of samples, the ratio k/n is a good estimate for the probability P and hence for the density function $p(\mathbf{x})$.

And if I consider the $p(\mathbf{x})$ is continuous and the region is so small that $p(\mathbf{x})$ does not vary significantly within it. So I am considering this case that is the $p(\mathbf{x})$ is continuous and that the region R is very small, very, very small, then in this case the $p(\mathbf{x})$ does not vary significantly within it. So for this I can write like this. So integration $\int p(\mathbf{x}') d\mathbf{x}'$ is approximately equal to $p(\mathbf{x})V$.

So I will be getting this and from this expression, from the previous expression because if you see the previous expression is the probability is nothing but k divided by n . And now I am getting the this probability that $p(\mathbf{x})$ is nothing but k divided by n divided by V . So by using this expression I can determine that density, the probability $p(\mathbf{x})$ that is the density is equal to k divided by n divided by V .

So V is the volume and close by the region R . So R is the region I am considering and V is the volume enclosed by the region the region is R . So by using this expression I can determine the density the density ρ I can determine.

(Refer Slide Time: 36:51)

- However, V cannot become arbitrarily small because we reach a point where no samples are contained in V , so we cannot get convergence this way.



- Alternate approach:
 - V cannot be allowed to become small since the number of samples is always limited.
 - One will have to accept a certain amount of variance in the ratio k/n .

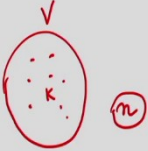
Now in this case however V cannot become arbitrarily small because we reach a point where no samples are contained in V . And in this case we will not get the convergence the V cannot be very, very small. Suppose if I consider the volume is very, very small, then in this case what will happen it may not enclose any samples. Then in this case you cannot determine that density.

So what a process I can consider V cannot be allowed to become small since the number of samples is always limited. Because we have the limited number of samples. So V should not be very, very small. Otherwise we cannot expect that, the samples will be available within this particular volume, the volume is very small. And one another case we have to consider the certain amount of variance in the ratio k divided by n .


So we can consider this that means the volume should not be very, very small because we have the limited number of training samples. And also we have to consider the certain amount of variance in the ratio, the ratio is k divided by n that we can consider.

(Refer Slide Time: 38:05)

$$p(\mathbf{x}) \cong \frac{k/n}{V}$$



- Fix the volume of region V and count the number of samples k (out of n) falling in $V \rightarrow$ Parzen Window
- Vary V in a way so that to enclose k samples around \mathbf{x} and make a decision for the label of $\mathbf{x} \rightarrow$ k - Nearest neighbor



So you know this expression the probability of \mathbf{x} that is the density k divided by n divided by V . So in case of the Parzen-Window technique fix the volume of the region V and count the number of samples k out of n number of samples falling in V . So that means the volume is fixed. And we have to count the number of samples within this particular volume the volume is V . So k number of samples within this particular volume out of n number of samples.

So total number of samples is n and I am counting how many samples are within this particular volume. So k number of samples within this particular volume. So from this information you can see k divided by n divided by V from this information I can determine the density that is called a Parzen-Window technique.

In case of the k nearest neighbor technique the volume is not fixed. But I can consider the k number of samples so suppose it is fixed suppose I am considering 5 number of samples. So I have to increase the volume so that it encloses the 5 number of samples that means that k number of samples. So first volume it encloses one sample suppose the second volume it considered suppose two number of samples and third volume if I consider, it considered, it encloses two samples. So how many samples total samples 5 number of samples enclosed by this volume.

So that means I am increasing the volume, I am growing the volume so that it encloses the k number of samples. In case of the Parzen-Window technique the volume is not; the volume is fixed. And I have to count the number of samples k within this particular volume. In case of the k nearest neighbor technique the volume is not fixed and we have to increase the volume. So that it encloses k number of samples. And from this information I can determine the

density, the density is p_x is equal to k divided by n divided by V . So by using this expression I can determine the density.

(Refer Slide Time: 40:14)

▪ To estimate the density of \mathbf{x} , we form a sequence of regions R_1, R_2, \dots containing \mathbf{x} : the first region contains one sample, the second two samples and so on.

Let V_n be the volume of R_n ,

k_n the number of samples falling in R_n and

$p_n(\mathbf{x})$ be the n^{th} estimate for $p(\mathbf{x})$: $p_n(\mathbf{x}) = (k_n/n)/V_n$.

Theoretically, if an unlimited number of samples is available, we can show $p_n(\mathbf{x}) \rightarrow p(\mathbf{x})$.


So for estimating the density what I am considering. So to estimate the density of \mathbf{x} we promise sequence of regions, we can consider a regions $R_1 R_2$ like this containing \mathbf{x} . The first region contains one sample, the second contains two samples and like this and V_n be the volume corresponding to the region R_n . And k_n is the number of samples falling in the region R_n . So I am considering the k number of samples falling in the region R_n .

And in this case I can determine or I can estimate the probability or the density p_x . So the p_n be the n^{th} estimate of the probability the probability is p_x . So I can determine the density, the density is nothing but k_n divided by n divided by V_n . And if I consider suppose the unlimited number of samples. So many, many samples if I consider then what convergence I can get? The p_n converges to p_x . So if I consider a large number of samples then the p_n approaches p_x .

(Refer Slide Time: 41:30)

Parzen Window

- Three necessary conditions should apply if we want $p_n(\mathbf{x})$ to converge to $p(\mathbf{x})$:

- 1) $\lim_{n \rightarrow \infty} V_n = 0$ 
- 2) $\lim_{n \rightarrow \infty} k_n = \infty$ ✓
- 3) $\lim_{n \rightarrow \infty} k_n / n = 0$ ✓

The estimate for $p_n(\mathbf{x})$ is:

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

But in this case, in case of the Parzen-Window technique we have to consider these three conditions one is the limit $\lim_{n \rightarrow \infty} V_n$ should be equal to 0. So that means what is the meaning of this the volume may be very, very small because I am considering large number of samples, because n tends to infinity. Then in this case you may expect some samples within this particular volume the volume is suppose V_n .

The volume V_n may be very, very small because I am considering very large number of samples, large number of samples we are considering. So that means you may expect some of the samples within this very small volume, the small volume is V_n . So this is the first condition that is the limit n tends to infinity V_n is equal to 0. The second condition is $\lim_{n \rightarrow \infty} K_n$ is equal to infinity.


Since we have large number of samples, so that means the k_n will be also very large. Since the N is very large, so that means the k_n is also very large. And if I consider a ratio K_n divided by n , n tends to infinity that is the limit, then in this case this ratio will be 0, tends to 0, because n is very, very high as compared to k_n . So that is why this K_n divided by n that ratio limit n tends to infinity should be equal to 0.

And there are some mathematical derivations in the case of the Parzen-Window. So I can determine the density, the density can be estimated by using this equation. So you can see the book by Duda and Hart; The Pattern Classification by Duda and Hart and you can see the derivation of this equation. So by this equation you can determine the density.

(Refer Slide Time: 43:20)

K Nearest neighbor

- **Goal:** a solution for the problem of the unknown "best" window function.
- Approach: Estimate density using data points.
- Let the cell volume be a function of the training data.
- Center a cell about \mathbf{x} and let it grow until it captures k_n samples:
$$k_n = f(n)$$
- k_n are called the k_n nearest-neighbors of \mathbf{x} .




Now let us consider the k nearest neighbor technique. So in this case we have to estimate the density using the data points and let us consider the cell volume to be a function of the training data. And in this case the center a cell about \mathbf{x} and let it grow until it captures k_n number of samples. So here you can see a cell volume be a function of the training data.

And what I am considering? I am growing the region, that means I am increasing the volume, so that it encloses the k_n number of samples. So it encloses the k_n number of samples. So k_n is a function of n , n the total number of samples. So k_n are called the K nearest neighbor of \mathbf{x} that is the k_n .

(Refer Slide Time: 44:10)

K Nearest neighbor

- Two possibilities can occur:
 - Density is high near \mathbf{x} ; therefore the cell will be small which provides good resolution.
 - Density is low; therefore the cell will grow large and stop until higher density regions are reached.



And two possibilities can occur the density is high near \mathbf{x} therefore the cell will be small which provide good resolution. So density may be high near the feature vector, the feature vector is \mathbf{x} . And in this case therefore the cell will be small which provide good resolution. In the second case the density maybe low therefore the cell will grow large and stop until higher density regions are reached.

So that means I have to increase the volume so that it encloses k_n number of samples. So density is low, therefore the cell will grow large and stop until higher density regions are obtained. So this is the two conditions of the k nearest neighbor technique.

(Refer Slide Time: 44:59)

K Nearest neighbor

- **Goal:** estimate $P_n(\omega_i | \mathbf{X})$ from a set of n labeled samples.
- Let's place a cell of volume V around \mathbf{x} and capture k samples.
- k_i samples amongst k turned out to be labeled ω_i then:

$$p_n(\mathbf{X}, \omega_i) = \frac{k_i / n}{V}$$
- A reasonable estimate for $P_n(\omega_i | \mathbf{X})$ is:

$$P_n(\omega_i | \mathbf{X}) = \frac{p_n(\mathbf{X}, \omega_i)}{\sum_{j=1}^c p_n(\mathbf{X}, \omega_j)} = \frac{k_i / nV}{\sum_{j=1}^c k_j / nV} = \frac{k_i}{k}$$

$\frac{k_i}{k}$ is the fraction of the samples within the cell that are labeled ω_i .

So mathematically you can say, so directly we can estimate this posterior probability P_n given \mathbf{X} directly we can determine from the n labeled samples, n number of samples are available. And there are labeled samples that means I am considering the supervised training. Let us place a cell volume V around \mathbf{X} and capture k samples. And k_i samples amongst k turned out to be labeled w_i .

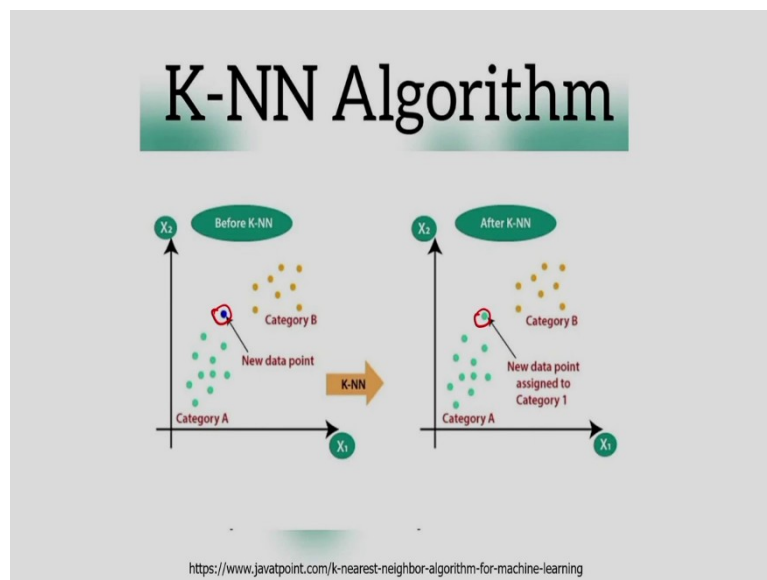
So that means I am considering k_i number of samples corresponding to the class the class is w_i . So total number of samples are n and I am considering k_i number of samples corresponding to the class, the class is w_i . So from this you can determine the density. The density is $P_n(\mathbf{X}, w_i) = k_i$ divided by n divided by V that you can determine. And from this you can see I can determine the posterior density $P_n(w_i | \mathbf{X})$ that I can determine.

So if you see this equation, so what I am determining already $P_n(\mathbf{X}, w_i)$ that I have determined, that is nothing but k_i divided n into V . And you can see the summation I am

taking that is the evidence. And I will be getting the ratio, the ratio is k_i divided by k . So what is the k_i samples amongst k ? So k number of samples I am considering but k_i samples corresponding to the class the class is w_i . And I am considering the total number of samples, the total number of samples is n .

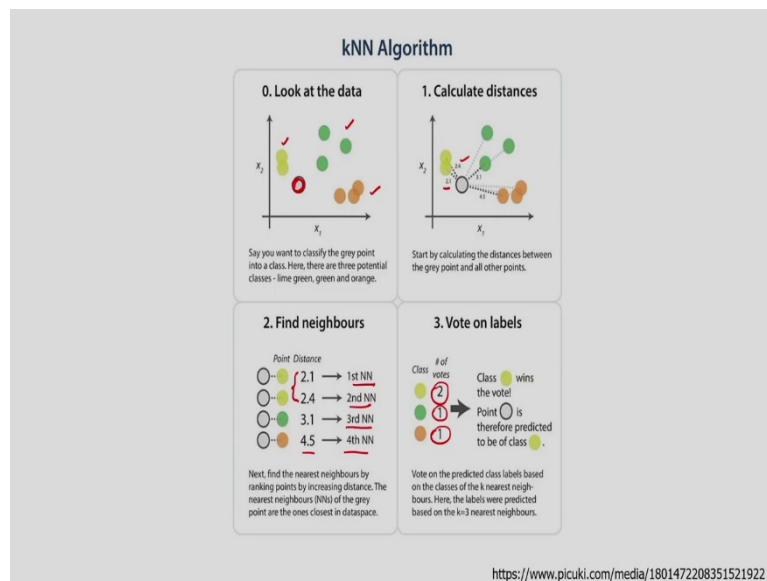
So this ratio the k_i divided by k that gives the information of the density that is the posterior density. So it gives the information of this. So we have to determine the k_i number of samples corresponding to the class, the class is w_i . So if I can determine the ratio k_i divided by k determine this ratio. Then I can determine the density, the density is $P_n w_i$ given x that I can determine.

(Refer Slide Time: 47:12)



So this process the k nearest neighbor algorithm that is the classification algorithm I can show pictorially like this. Suppose I want to classify a new data point, so in new data point is this. And I have to classes the category A and category B and this is before k nearest neighbor. And after this based on the minimum distance the new data point is assign to the class, the class is A. So that is the category 1, category 1 means the class A. So based on this you can see this new data point is assign to the category 1. So this is the k nearest neighbor algorithm.

(Refer Slide Time: 47:51)



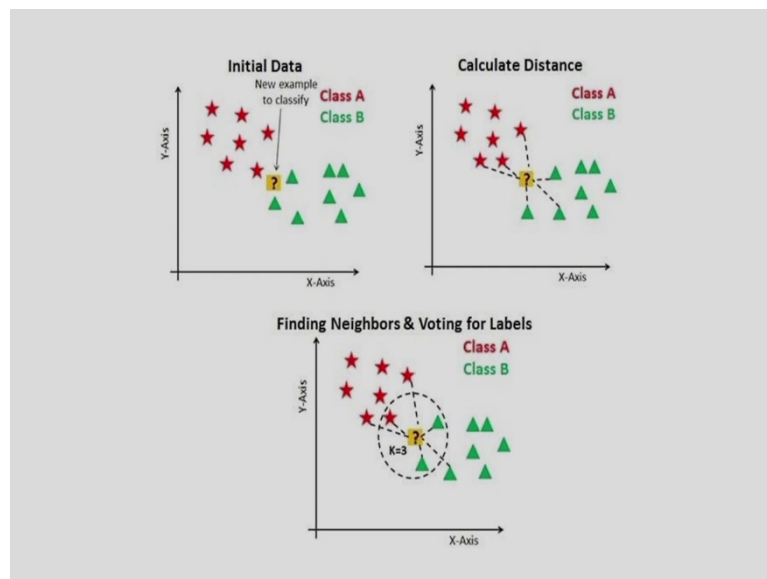
Again I can show what is the k nearest neighbor algorithm. Here you can see I am considering three classes one is the yellow, one is the green and one is the orange. And I am considering one new data point so this is the new data point I am considering. And I am finding the distance between this data point, the new data point and the other samples corresponding to different classes.

So one class is the yellow class, another class is the green class, another class is the orange class or a red class. So you can see I am finding the distance, the distance is 2.1, 2.4 like this I am determining the distance. So corresponding to this yellow one, you can see the distance is 2.1 that is the first nearest neighbor a distance is minimum. And again corresponding to the second one, the second data point that is the yellow one, the distance is 2.4 that is the second nearest neighbor I am getting.

And again if you see this green one. So corresponding to this green one the distance is 3.1 that is the third nearest neighbor. And again you can see the distance between the grey point and the orange or the red point, so it is 4.5 that is the first fourth nearest neighbor. And in this case we have to determine the number of votes.

So corresponding to this yellow how many votes, because two times it is the neighbor. So 2 votes I am getting, corresponding to the green I am getting 1 vote, corresponding to the orange I am getting 1 vote. So I can count the number of votes and based on this the new data point can be assigned to a particular cluster.

(Refer Slide Time: 49:35)



Again I am showing this one, so this new data point I have to classify and for this you can see I have to classes the class A and the class B. And based on the minimum distance this new data point can be assigned to a particular class, the class A and class B. So in this class I discuss the concept of the Bayesian distance theory and after this, I discuss the concept of the concept of the parameter estimation.

In the parameter estimation I discuss two algorithms very popular algorithms; one is the maximum likelihood estimation, another one is the Bayesian estimation. So briefly I explained these two concepts, after this I considered the case of the non-parametric estimation. In case of the non-parametric estimation I have to estimate the density. So for this I considered two algorithms one is the Parzen-Window technique another one is the k nearest neighbor technique.

So briefly I explained the basic concept of the maximum likelihood estimation, Bayesian estimation, the Parzen-Window and the k nearest neighbor technique. For more detail you can see the book Pattern Classification by Duda and Hart that you can see. So let me stop here today. Thank you.