



Machine Learning for Engineering and Science Applications
Professor Dr. Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
Introduction to Probability Theory Discrete and Continuous Random Variables

(Refer Slide Time 00:16)




Machine Learning for Engineering and Science Applications

Introduction to Probability Theory
Discrete and Continuous Random Variables





In this video you will be looking at an introduction to probability theory and specifically will introduce the idea of discrete and continuous random variables.

(Refer Slide Time 00:26)



Introduction

- Probability -- Mathematical framework for representing uncertainty
- Multiple sources of uncertainty in Engineering and Science
 - Inherent randomness in system
 - Example : Quantum mechanics
 - Incomplete data/observability
 - Example: Macroscopic descriptions
 - Incomplete modeling
 - Example : Weather Models



So probability is a mathematical framework for representing uncertainty wherever you have some uncertain outcome. We tend to use probability as a mathematical representation of the uncertainty in the problem. So in engineering systems we have multiple sources that such

uncertainty occurs. Sometimes we have inherent randomness or stochasticity more specifically in the system. For example, in quantum mechanics the laws themselves actually lead to some amount of randomness or uncertainty.

Or let's say you are dealing with a pack of cards so you have a little bit of randomness thrown in there. So in such cases whenever you try to predict something you are going to have a probability theory coming in. Now we can look at a slightly higher level of abstraction where you might have deterministic systems. That is the laws themselves are actually not random unlike quantum mechanics but you might have incomplete observability which is you are not able to see all that is happening in the system. For example, if you have a macroscopic description of let's say flowing a room, we know that inherently there are molecules and within them atoms etc. You are not able to observe them.

And typically this leads to a little bit of uncertainty in the properties. We know that macroscopic properties are defined and derived from microscopic properties but this is done so actually probabilistic even though in real life we don't treat them as if they are probabilistic. You have multiple such examples. Whenever we have incomplete observability once again you can use probability theory.

A third level of abstraction is you might not have randomness. You might even have in some sense complete data plus you might have deterministic laws but still despite having full data you actually have an incomplete model. The model can be incomplete as in let's say weather models. So you do not deliberately you do not use all of the data in Article II that gets simple models are tractable markets in all these cases. In engineering you will typically use probability theory.

(Refer Slide Time 03:00)

Dual use of probability ideas in
Machine Learning

- Constructing Learning systems *Model*
 - Incorporate probabilistic algorithms by trying to mimic human reasoning about uncertainty
 - Probabilistic Models
- Analyzing Learning systems *Deterministic Model*
 - Even deterministic learning systems are only correct part of the time. Their output can, therefore, be analyzed probabilistically.
 - Probabilistic analysis of deterministic/probabilistic models

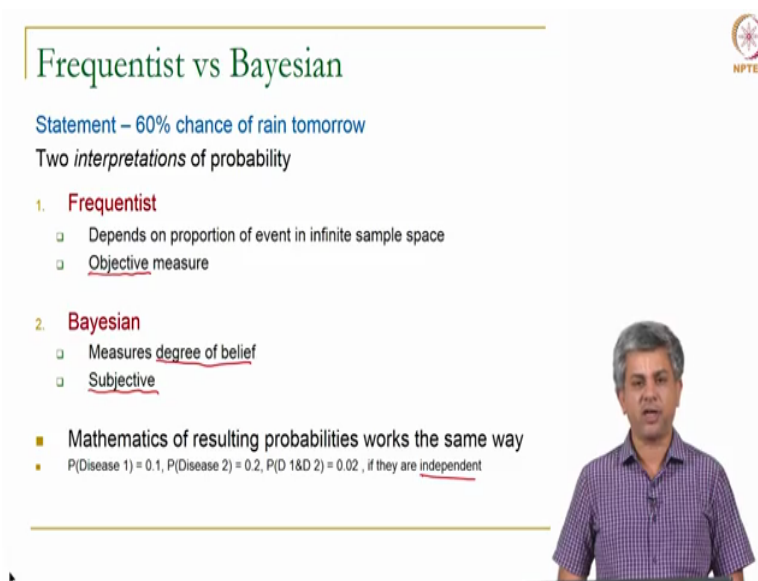
Now our interest of course is we want to use probability ideas in machine learning and there are two primary uses of that, first is in constructing learning systems themselves by a learning system I simply mean a machine learning model. Ok so you want to construct a model so if you try and mimic let's say human. Reasoning about uncertainty we say the probability of rain is probably sixty percent tomorrow.

So inherently even with on our models there is some probability built in. Ok so in order to incorporate such probabilistic thinking you have probabilistic models. So notice this. You can have probability right built in right into the model itself. Ok so that is one way of using probability. Another idea is you might actually have a deterministic Model for example as we see many neural network models are almost by design they are deterministic so you could have a deterministic Model.

That is how the input relates to the output is actually a deterministic process. Nonetheless the output itself can be analyzed probabilistically because the learning system is only correct part of the time. It's not correct all the time. So for example if you might see a Google Image analyzer or any other image analyzer. Typically, the actual output in the algorithm as we'll see later on in the course will not be a specific class.

It will not say this picture as it got deterministic. What we can say typically something like this picture is a cat with probability point nine. So this would be something like an analysis of the learning system probabilistic. It might go wrong 90 percent of the time. Stuff like that can be analyzed probabilistic. This is a probabilistic analysis of even deterministic or even probabilistic motives.

(Refer Slide Time 05:12)



The slide is titled "Frequentist vs Bayesian" and features the NPTEL logo in the top right corner. The main text on the slide reads: "Statement – 60% chance of rain tomorrow" followed by "Two interpretations of probability". It lists two points: 1. Frequentist, which includes "Depends on proportion of event in infinite sample space" and "Objective measure"; 2. Bayesian, which includes "Measures degree of belief" and "Subjective". A third point states "Mathematics of resulting probabilities works the same way" with an example: "P(Disease 1) = 0.1, P(Disease 2) = 0.2, P(D 1 & D 2) = 0.02, if they are independent". A speaker overlay is visible on the right side of the slide.

Then become we come probabilistic analysis. There are two interpretations of what a particular probability means. Now all of us know probability lies between 0 and 1 but there are two large schools of thought. They often come into philosophical fight. You will not get to mention that in this course at least. But just for a brief, Introduction if you take a statement such as that a sixty percent chance of rain tomorrow.

This can be interpreted in two distinct twists. So one way is what we are usually used to. This is called a frequent stock model frequent. This model would be something like Ok so the temperature rise this much the pressure today so much it's slightly cloudy. And in all such cases before then such things happen in sixty percent of the times it rained, So such a statement would depend on let's say something like what you have observed so far. And that's a frequent based approach to probability.

It says it depends on the proportion of events in an infinite sample space. I see and see shortly. It's typically an objective measure. So if I say what is the probability of a Faraday throwing up to you will say if I throw this dice let's say millions of time two will come up about one sixth of the times. So the probability is one six, This is an object to measure. The second measure is called the Bayesian approach, Preferred typically by economists or even philosophers. Ok so this actually measures degree of belief.

So if somebody says that there is a sixty percent chance of rain tomorrow what they mean typically is it looks a little bit more than you know fifty fifty. So it looks kind of likely that I am going to get about rain a little bit more than I but I am not really sure. So something of that sort it's a rough estimate. That's something like a base. Yeah of course there are more


technical meanings it's not as bad as what I'm making it out to be but it's a subjective measure. So there is a certain amount of degree of belief in the statement that's incorporated in a Bayesian.

Now for all purposes you do not really strongly get in a couple of places. We'll make this distinction but other than that we do not really strongly care about which approach we are taking because whatever the probabilities result out of this the mathematics works exactly in the same way. For example, if a doctor says to a person your probability of getting a disease one let's say heart attack is point one and probability of getting disease two, let us say foot ache is point two. Ok let's say that these two diseases are independent okay. This is important in the example I am using. Now regardless of which interpretation of probability you choose.

It is always true that probability of disease one and disease two will leave them even side independent. You are going to get point one multiplied by point two, which is point zero two. This is regardless of whether it's a frequent test approach although there is a Bayesian approach. So the mathematics of probabilities work exactly in the same way regardless of which approach we choose. So we'll stick to you know choosing between frequent tests and Bayesian depending on what makes sense and we'll only look at what the mathematics of the resulting probabilities.


(Refer Slide Time: 08:33)

Definitions



- **Random experiment** – Experiment that results in different outcomes despite being seemingly similar conditions.
 - Example – Tossing of a coin, throwing of a dice, rainfall amount

- **Sample space** – Set of all possible outcomes of a random experiment. Coin toss $S = \{HH, HT, TH, TT\}$
 - Example : Tossing of a coin. $S = \{H, T\}$
 - The sample space we choose depends on the purpose of analysis
 - Example : Diameter of a manufactured pipe. S could be
 $S = \mathbb{R}^+ = \{x \mid x > 0\}$ OR
 $S = \{low, medium, high\}$ OR
 $S = \{satisfactory, unsatisfactory\}$



So let's come to a few definitions which we'll be using. The first is the definition of a random experiment the simple definition of random experiments you do an experiment and it results in difficult different outcomes each time. Despite you having similar conditions for example I toss a coin, it seems to me that I am keeping the coin exactly the same way on my thumb in

exactly the same way, and yet sometimes you get heads and sometimes you get tails, such an experiment is called a random experiment. So rainfall amounts are throwing off dice infinite examples of this. The second definition is that if a sample space ok so suppose you do a random experiment all the possible outcomes the set of all possible outcomes of this random experiment is called the sample space. For example, if you toss a coin the set of all possible outcomes is you either get a heads or you get a tails. This would be if you are tossing a coin once. Now suppose I toss a coin twice, Then the sample space is heads heads. Heads tails, tails heads or tails tails. So this would be my sample space.

Now one of these four should have a card when I toss the coin four times or two times. Now what's important is that the sample space which we use for determining probabilities depends on actually the purpose of analysis. So the same event can be described in many different ways. So let's say we have manufactured a pipe and we want to and knowing that manufacturing has certain an uncertain base built into it you are not always going to get a pipe of the same size each time.

So we can call this an random experiment. So what I want to describe is the sample space of what kind of pipe did I get. Now depending on the purpose of my analysis our sample space S could be either $x \in \mathbb{R}^+$ means the positive half of the real number line. So this is simply S is some number which is positive. All of us know this. So all we are saying is the diameter could lie anywhere between zero and infinity.

This is one sample space. Another possible sample space is the diameter of the pipe was low or it was medium or it was high. Suppose we are only interested in whether it is in one of these three. Is it too small. Is it kind of okay or is it too big. If these were our only three plus quantities of interest are qualities of interest in an analysis. Our sample space would be simply this okay. Or we could be basically interested in is it a satisfactory pipe for my purposes or unsatisfactory. So my sample space simply has two elements satisfactory or unsatisfactory.

The point is that you can describe the outcome of the same event in many different ways depending on which way you wish to analyze it and as an engineer often later on when we make machine learning models this becomes an important part of what role you play.

(Refer Slide Time 11:46)

Random Variables

- Useful to denote outcomes of random experiments by number
- Can be done even for categorical outcomes
- The variable that associates a number with an outcome of a random experiment is called a random variable $\rightarrow R, Z$

Notation – The random variable is denoted by a capital letter (e.g. X) and its value is denoted by a small letter (e.g. x).

Example : The rainfall on a particular day is a random variable R . We can ask "What is the probability that the rainfall is greater than 10mm?" by the mathematical notation $P(R > 10) = ?$ and $P(X = 3)$.

Describing the sample space. We come to a very important quantity. This is a fundamental quantity often while using probability theory, The idea of a random variable. So it is useful typically to denote the outcome of a random experiment by a number. Now notice for example if I toss a coin my outcome my sample space for example was either heads or tails heads, By itself it's not a number. Tails by itself It's not a number. You could assign a number to it. For example, you could assign the number one two heads and zero to tails.

Ok so then if you assign one number for example you could even have categorical outcomes. For example, I take an image I ask is this a cat or dog horse or a picture of a cow. So then you have four possible outcomes again cat dog horse cow horse cow are by themselves not numbers but you can assign numbers for example zero one two three or one two three four etc.


Ok so you can even assign numbers numerical to a categorical outcomes Ok so the variable that associates a number with an outcome is called a random variable. Ok so please notice this random variable by itself is something that is mapped, to either the yellow number or the integers etc. etc. notation, this sometimes gets confusing for students so please remember this the variable itself is denoted by a capital letter. For example, Captain X would denote a random variable the variable itself its value is denoted by a small number. For example, if I say X equals to point 5. X is the random variable and point five is the value that it takes.

Let's take another example. Suppose we want to find out the rainfall on a particular day. So this is a random variable as you know we cannot say for sure what that exact amount of rainfall would be. So let's call this random variable capital R . The amount of rainfall would

be actually denoted by small R . So suppose I want to make the statement. What is the probability that the rainfall is greater than ten MM, So I want to see how to denote this or what is the notation I would use for this. Remember probabilities denoted by P the amount of rainfall as a variable is capital R .

The actual value it takes which is ten is denoted by small R . So we would write something like the mathematical notation is what is the probability that the amount of rainfall is greater than ten MM. So S are greater than ten. Ok so suppose I ask what is the probability that the dice give me the number three. I would and if X was denoting the random variable which gives you the output of the dice you would say something like probability that X is equal to three. So suppose you have a uniform random variable a uniform random variable is one that all outcomes are equally likely. For example, you have an unbiased coin. If you throw it you either get a heads or a tails with a probability point five. So the probability distribution is called a uniform distribution.


(Refer Slide Time 15:16)



Probability Distributions

A **probability distribution** tells us how likely a random variable is to take each of its possible states.

- **Discrete Random Variable (RV)**
 - Has finite (or countably infinite) range
 - Example – No. of typographical errors, no. of diagnostic errors, etc
 - Probability measured by Probability Mass Function (PMF)
- **Continuous Random Variable (RV)**
 - Has real number interval for its range.
 - Example – Temperature, Pressure, Voltage, Height, Current, etc
 - Probability measured by Probability Density Function (PDF)



Similarly, for a Dice you could have a uniform distribution again, ok so let's come to probability distributions, A probability distribution tells us how likely a random variable is to take each of its possible states. So remember a random variable can pick any state depending on what the sample space is if the sample space has 10 members. Then the random variable can take all 10 values any of the 10 values not all 10 values simultaneously. Any of the 10 values the probability distribution tells us that not all of them might be equally likely. Some of them might be less likely. Some of them might be more likely. So that probability distribution is what tells you how likely each one of these values is.

So depending on what kind of variable we are dealing with we might have two different types of probability distributions. So very common probability distribution is that of a discrete random but a very common random variable type. It's a discrete random variable. This has a finite accountability infinite number of possibilities. For example, if we look at the number of errors in a particular page or the number of errors i make while speaking or the number of errors doctor made in diagnosis all these are actual numbers. Ok so these are finite.

The range of the random variable we can take is actually finite. Ok so these are discrete random variables. Now more importantly the probability is measured by what is called a probability mass function as we see in the next slate. So we can have a continuous random variable also which has a real number interval for its range. An example would be any real number random variables such as temperature pressure voltage current etc. In such a case the probability is measured by probability density function. Please notice the difference for a

discrete variable it's a probability mass function for a continuous variable it's a probability density function.

(Refer Slide Time 17:17)

Probability Mass function

- Discrete Variable -> Probability Mass Function (PMF)
 - PMF -- List of possible values along with their probabilities
 - Example

X : Number that comes up on throw of a biased die

$P(X=1) = 0.1$ $P(X=2) = 0.1$ $P(X=3) = 0.2$
 $P(X=4) = 0.2$ $P(X=5) = 0.2$ $P(X=6) = 0.2$

■ To be a PMF for a random variable X, a function P satisfies:

- Domain of P is the set of all possible states of X → Sample space must be covered
- $0 \leq P(X=x) \leq 1$
- $\sum_{x \in X} P(X=x) = 1$ $P(S) = 1$

■ Uniform random variable: $P(x=x_i) = \frac{1}{k}$ ← k Outcomes
 $i=1, \dots, k$

■ Analogous to a point load

So let's come to a probability mass function. Once again it's done for a discrete variable. We denote it by PMF standing for probability mass function. All it this is a list of possible values.

Okay so if the random variable takes multiple values it's simply a list of those values along with their probabilities. So let's say you have a bias dice a bias means not all six sides are equally likely. So one is somewhat likely two has a different probability etc.

So let's say we have these six probabilities. Notice how I am denoting this. All I'm saying is P of X equal to one equal to point one etc you will see this should be actually true. Please excuse me. So P of X equal to on this point one P of X equal to his point one etc.. You have to give a probability for each possible outcome in the sample space. So for example if I take a graph and have this six possibility strong here which is the sample space please remember this is X. Then probability that x equals to one is point one. Point one. Point two Point two point two so this essentially is the probability mass function ok.

Some of you might notice that this looks like a point load which is exactly true. Instructors who might have seen something of the sort a point force applying at a single point.

So that's what the probability mass function is analogous to. So in order for a probability mass function to be a valid probability mass function it has to satisfy certain criteria. One thing is you are to make sure that the domain of P is the set of all possible states of X. All that means is that P should have some that of some valid value for each output for X. For example

if I don't give let's say these two values and say that P is valid only from one to four then this is not a valid probability mass function.

My whole sample space must be covered. Next of course all of the individual probabilities since they are probabilities how to lie between zero and one they have to be non negative. As well as they have to be less than one. Finally we know that the whole sample space, So since one of these X should definitely occur the probability summation of individual probabilities has to be one. So using these laws you can immediately come up with the fact that for a uniform random variable which is a random variable where all the outcomes are equally likely.

If there are K outcomes so X I, I goes from one to K. Then a uniform random variable each of those probabilities will be equal to 1 by K. Ok so as I said earlier this is analogous to a point Load.

(Refer Slide Time 20:53)

Probability Distributions (contd)

- Continuous Variable -> Probability Density Function (PDF)
 - PDF - F
 - Like a d
 - R: Amount
 - $P(10 < R < ;$
 - To be a F satisfies:
 - The Dor
 - $\forall x \in X, f$
 - $\int_X p(x)$
- Normalized histogram approximates a probability density function

Diagram: A histogram with a normal distribution curve overlaid. The area under the curve is shaded. A small diagram shows a rectangle with width 0.5 and height 2, labeled 'Area ≤ 1 ' and 'Probability'.

So let's now come to continuous random variables. Remember that instead of a mass function you now have a density function ok. D stands for density. What it is effectively is a probability per unit length, once again you can make an analogy so instead of a point Load. You now have something like a distributed load since this is a continuous function. We don't have gaps between any two random variables.

What we have is a continuous distribution and instead of giving probability at the point what we will give is probability but unit length or in other words probability density. This is like a distributed load. So let's say R is the amount of rainfall and I want to find out probability that

the rainfall lies between ten and twenty. As it turns out the probability of any particular point is irrelevant. What you look for is probability in a range. So let's say you want to find out the probability that the rainfall is between ten and twenty ok. In that case we simply denote it by $P(10 \leq R \leq 20)$. These two are equivalent because probability of any exact specific value is effectively zero.

The probability of any exact value is effectively zero. So this probability can be given by any area under this curve just like for distributed Load, Load can be given by area under the curve ok. So if you don't understand that analogy even then you can immediately see that each of these probabilities when they sum up if it is probability per unit length then this area is given by integral of $P(X) dx$ between ten and twenty.

So in general the probability that A will be given by integral between A to B $P(X) dx$. So in order for P to be a probability distribution function the domain once again just like last time has to be all possible states of X ok. And probability density has to be positive. Notice that it's not necessary for the probability density itself to be less than equal to one because this is a density.

So let me show an example, So let's say my probability density function looks like this it's a heart function let's say this is $(0, 0.5)$. Then this value this area has to be one therefore this height has to be two ok. All I'm interested in is in making sure that the area of any sub portion has to be less than equal to one because it's the area which is the actual probability. This small p affects on the other hand is the density. what it is probability per unit length.

So I can arbitrarily make this large by simply reducing the length for the same probability. Our condition here is of course that integral of $P(X) dx$ has to be equal to one. due to way of thinking about the probability density function is to think of it as a normalized histogram. Okay so suppose I take a random number which lies between let's say minus five and five and I take ten thousand such random numbers. So let's say here is the histogram that's drawn. Notice these values here this value is eight hundred. This values one hundred. So some that I don't zero you get a lot of hits.

And everywhere else we get a low number of hits. We can draw a histogram. Now suppose I normalize this by what I mean by normalization is instead of looking at number of times I got a zero or a number of times I got a value between that's a minus point one and zero instead of doing that. I start looking at what fraction I got. So all these numbers if I divide them by ten thousand what they now will get here is point zero eight. So I had ten thousand tosses which

went between minus five and five. So now I divide all these numbers and I look at fractions instead of actual numbers. You see that they form a curve of this sort.

Now suppose I keep on increasing my tosses. You might kind of guess that slowly but surely it'll start converging to some nice bell curve even see this later. Later on this week such a curve will usually be a glossy anchor but it will usually converge to some sort of curve that curve would be the probability distribution function not for a finite number of throws but for an infinite number of throws a normalized histogram tends to a probability distribution function, Thank you.