Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 03

Lecture - 13

Lecture 13 : Uncertainty Modeling

Hello everyone. Welcome to this course on Artificial Intelligence for Economics. I am Adway Mitra, an Assistant Professor in Indian Institute of Technology, Kharagpur. And today we are on lecture 13 of this course. The topic of today's lecture is Uncertainty Modeling. So, in the last few lectures, we have been discussing the general idea how we can learn from data so that we can make predictions on new data.

So, we have already discussed the concept of unsupervised and supervised learning. We have also learned some of the like fundamental models and algorithms related to supervised learning such as decision trees, linear classifiers, linear regression and so on. Now, whenever we are making a prediction, we must understand that the prediction is never expected to be perfectly accurate. There is always going to be some degree of uncertainty whenever we are making a prediction.

Now, where does this uncertainty come from? There are two kinds of uncertainty primarily, they are called the aleatoric and epistemic uncertainty. The first one kind of uncertainty is due to our imprecise observations. So, you would remember one thing that when we were making any kind of prediction we first need our x that is the feature vectors and on the basis of which we try to predict y which is our level. Now, when we are the feature vectors where do these feature vectors come from typically they may come from some kind of observations. So, when we are making those observations there may be some amount of noise in the measurements.

So, that already or they or it might be that there are some observations which are missing that that missing this missing observations may either be because our instruments are faulty they were not able to observe certain things or it may be that we ourselves did not know that certain things needed to be observed. that is when we are making the prediction of any variable y. Let us say I want to predict what is going to be the GDP growth rate of India in the coming quarter. So, like they of course, this depends on many things on like on a variety of economic and extra economic factors. Now, it is possible

that there are some crucial factors which will impact the GDP growth rate, but which I am not aware of or which I am for some reason not able to measure.

So, the feature vectors will be insufficient to make the prediction. So, that bring so, the extra factors which were not observed for whatever reasons they adds a degree of uncertainty in our predictions. Like if we had been able to observe them maybe we could have made an accurate prediction, but since we are not able to observe them our predictions become a bit uncertain. The other type of uncertainty comes from the limitations of the model itself. So, like when we are using the linear regression or the linear classifier, you would remember that our task was to like learn a function f which maps from the feature space to x to the label space y.

But, now in case of linear regression or linear classification we have made an assumption that that function f is a linear function, but maybe in truth it is not maybe it is a non-linear function or maybe it is some kind of a kind of function which like we cannot express through mathematical parameterization that is we cannot write a mathematical expression for that function. So, in that case uncertainty of our predictions will be coming because of the insufficientness or insufficiency or like incapability of the function which we are using even though the observations of the feature vectors themselves may be perfect. So, these are the two primary sources of uncertainty which come in when we are trying to make a prediction. So, this is why when we are making a prediction it is usually not enough to just give a number that this is what we are going to this is the value which we are going to observe. In general or in most situations that is not going to be the accurate answer.

So, when we are like making a prediction along like if we are at all to give a single number that will probably be something like the. most likely scenario and along with it we also have to specify some other scenarios which are less likely, but still somewhat likely. So, this suggests that we use the concepts of probability theory somehow because probability theory it allows us to quantify uncertainty. So, in this lecture we will be focusing on the relevant concepts of probability theory and how they can be used for supervised learning. So, today we will revise the concepts of random variables as well as the well known probability distributions.

We will talk about expectations and how they can be used for risk analysis in economics as and we will also discuss the Bayes theorem and Bayesian networks which are also very important in many economic applications. So, like first of all, so many economic decisions which we like may need to make are based on estimates and projections of the future. Like for example, should I invest in a particular company or not, will it do well or not or if yes, if we can predict that it is going to do well then also we have to somehow decide how much I should invest. or how much budget should be allocated for a

particular purpose and how can may the expenditures change over time. So, all of these kinds of decisions we frequently have to make in the domain of economics and all of them involve some kind of prediction and the predictions as I already said may be possible to make using concepts of supervised learning.

But there is always uncertainty involved with these predictions and our aim here is to quantify such uncertainty with the help of probability theory. So, let us come to the the necessary concepts of probability theory. So, we all know that a probability theory is based on the concept of random experiment which is a process that can have several possible outcomes, but we cannot say that for one particular run of the experiment which outcome will going to happen. For example, in case of coin toss we know that the possible outcomes are head and tail, but when a coin is tossed once we cannot say exactly what will be the outcome, will it be the head or will it be the tail. Now, it is possible that if all the necessary information was presented to us means when we are tossing the coin exactly how we are tossing it, exactly which finger we are like on which we are applying the force, how much force we are employing to toss through that coin up in the air, then what is what are the properties of the wind around that place and so on and so forth.

So, maybe if all of these information we could have then maybe it could be possible to work out some equations of physics and come up with the definite answer that yes it is going to fall on the head or it is going to fall on the tail, but since we do not have those things we just assume that it is a noisy process or it is a random experiment. Same case goes with the dice throw also. So, the set of all possible outcomes like head or tail in case of a coin or 1, 2, 3, 4, 5, 6 in case of a dice, these are what is known as the sample space of the random experiment and event is like defined as some subsets of the sample space. For example, in case of the dice throw we can say that like we are getting an even outcome from the dice. So, obviously that consists of 3 outcomes 2, 4 and 6 these are all these outcomes are all part of that event call even outcomes for the dice throw.

Now, like based on this we define a probability measure. A probability measure is a mapping as a function from the event space to the interval $[0,1]$. That is we have a random experiment, we have its sample space that is all possible outcomes. We can define all possible subsets of it. that is the power set of the sample space the each of them is an event and to each such event we can associate a measure between $[0,1]$.

However, that has to satisfy certain criteria for example, $P(\Omega)$ the P of the entire sample space should be equal to 1. Furthermore, if A is any event we can define A complement as $\Omega \setminus A$ that is Ω is the sample space from that you remove all the outcomes associated with that event A and whatever are remaining is called A complement. So, we this relation must be satisfied $P(A)=1-P(\bar{A})$. Now, what is a random variable? A random

variable is a mapping from this event space which we already defined to the space of real numbers. Now, we say that like X is a random variable when we are saying X is a random variable although it is called a variable, but actually it is a function as we already said it is a function from the event space to real numbers.

So, when I say what is $P(X \leq a)$. So, this basically means it is the total probability or the sum of the probability measures of all the events for which the random variable X maps it to values less than a. or less than or equal to a right a is some number which we have some real number which we have specified then the probability that the random variable takes less than equal to a if the definition of this is that the it is defined as the sum of the probability measures P of all the events for which this condition is satisfied. Now, the range of the random variable X or the set of unique values which the random variable takes right. Remember that X is a function from the event space to the space of real number.

So, it has a range. It can take different values which are all real numbers, but what are the unique values that it takes. That set of unique values taken by the random variable or mapped to by the random variable is called as the support of that random variable. So, the random variables can be either discrete in which case their support set is discrete or it can be continuous for in which case its support set is continuous. So, like here are two examples. So, this is what a typical DRV or discrete random variable looks like.

So, you can see that it takes only three unique values 1, 3 and 7 each of which is associated with a probability measure. And as you can see these all add up to 1. And in this case like what we are seeing is it is a continuous random variable that is it can take all possible real numbers. or maybe all real numbers within a particular range. Now, if you integrate the area under this curve in this case also you will get 1.

So, why is that? So, now first of all like any random variable it is characterized by certain functions. The most fundamental of them is called as the cumulative decision function which is represented by $f(x)$. So, like so in this case x stands for a particular real number. That is the we are trying to evaluate this cumulative distribution function or CDFF at this particular value of a small x and the it is defined as this function is defined as $P(X \leq x)$. So, we have already defined in fact, sorry for a typo this will be $X \leq x$.

So, we have already defined what is meant by this thing probability of $X \leq a$. So, basically the sum of all the events in which the like they or rather I should say the sum of all the mutually exclusive events for which this condition is satisfied. each discrete random variable is characterized by one more thing which is a probability mass function. This CDF is it like it we can define it for both continuous and discrete random variables,

but for discrete random variables we have one more thing which is called as the probability mass function. So, it is a function $f(x)$ which is equal to $P(X=x)$.

So, like as I already said since it is a discrete random variable X its support is discrete. It takes only a finite or I should not say finite it takes only discrete values like in this case it took these three values 1, 3 and 7. Unlike this case in which X is taking like all possible values in this range which is of course, not discrete which is continuous. So, like so, we can like when we are saying $P(X=x)$. So, like this is like we can again go back to the definition of events and we can like calculate the probability measure associated with each of those events for which which are mapped to this value of X by this random variable and we can add them up.

So, like one important property which this cumulative this probability mass function it satisfies is that this $f(x)$ is it is always non negative it can never be less than 0, it should lie it must lie between 0 and 1 and when we add up this function $f(x)$ over all possible values of x that is over the support of X then it is always equal to 1. So, for different distributions these effect I mean it are they are characterized by a particular functional form of this probability mass function. And each of these probability these functions they are associated with certain parameters which are also like have certain characteristics. For example, the Bernoulli distribution, this is the most basic discrete random variable or discrete distribution.

It basically mimics the coin toss. Its support set is just 0 and 1 and its PMF is defined like this and it has a single parameter p. Similarly, we have other discrete random variables like binomial, Poisson, geometric, categorical, multinomial each of which have a PMF, but this is not a class on probability theory. So, I am not going to describe all of these PMFs in detail. On the other hand, when we come to the continuous or continuous random variables. In that case the CDF still holds, but the like it is no the PMF or the probability mass function does not work in that case.

Instead what we have is probability density function. So, what is that? So, it is characterized by $f(x)$ a function f measured at a real number x. So, it is equal to P of $x-\delta$ less than equal to the random variable X is less than equal to $x+\delta$. That means, it is the cumulative the total probability of all the events which are mapped to values between $x-\delta$ and $x+\delta$ by the random variable x. So, the like as I already mentioned the like the support set in this case is continuous like in case of uniform distribution the support set can be any interval on the real axis let us say a like any two numbers a and b.

In case of beta distribution the support set is only the interval $[0,1]$. In case of the Gaussian distribution also known as the normal distribution which we are all familiar

with the support set is the entire real line. the in case of Γ distribution it is non negative reals and so on and so forth. So, like just like every discrete random variable is characterized by a probability mass function every continuous random variable is characterized by its probability density functions which have a well known mathematical parametric form and the associated with it are certain parameters also like this. For example, in case of the Gaussian distribution we all know the parameters are the μ and the σ which are known as the mean and the variance parameters.

Now, what do we do with these distributions? So, like they are like we can define. joint probability distributions for 2 or more events that can occur simultaneously. Say for example, when I say like let in case of a dice throw let x be the event that we get an even outcome and y be an event where we get an outcome which is less than 3. So, in this case both can happen simultaneously the intersection of these 2 events is the outcome 2. because that is both even and it is less than 3.

So, like so let us say X and Y are the two random variables. So, then we can define what is known as the joint distribution of it. So, like we write it in this particular way. So, its joint distribution $f(X,Y)$. So, this is can be either a PMF or the joint PMF or joint PDF depending on whether they or continuous random variables or discrete random variables.

If it is continuous random variable, so it just means the probability of the event that x capital X takes the value small x and the random variable capital Y it takes the value small y. So, that is the definition of this joint distribution or joint PMF at the point x comma y. Similarly, in case of continuous distribution we can define the joint PDF, so that is also defined in this way. that is basically the probability of the event that the random variable X takes values between $x-\delta$ and $x+\delta$ and the random variable Y it takes values between $y-\delta$ and $y+\delta$. Now, like we say that like these random variables X and Y we call them independent if this condition is satisfied.

That is if they are joint distribution at any point x comma y can be factorized or not at any point, but at every point x and y it can be factorized by the product of their individual distribution. So, f x this is an individual PMF or PDF of the random variable X. This one is the individual PMF or PDF of the random variable y. If we multiply them together then we will get the joint PMF or PDF at that point provided these two random variables X and Y they are independent of each other. That is one does not influence the other in any particular way.

Now, however in general like they can influence each other like it might happen that random two random variables they have some bearing on each other. Say for example, like so based on this concept we have conditional probability that is the probability that

one event will happen when we already know that the another event has happened. That is the probability that outcome of a dice throw is 2 when we already know that the result is even. So, accordingly we can define the conditional PDF or conditional PMF also in this way and it is defined as the joint PMF or PDF divided by the PMF or PDF of that random variable on which we are conditioning. So, in this case when I write y given x that means, x is the condition that is x is something we already know and y is something which we are trying to predict.

So, we write the so, this can be calculated as the joint distribution joint PMF or PDF of these two at this particular point divided by the individual PMF of or PDF of X at like at the corresponding point. Now, there is a law of total probability and the Bayes theorem which are defined in this way and we are which we are all familiar with. So, like the law of total probability it allows us to reduce the joint distribution to the individual distribution by a process which is known as marginalization that is we either add over or integrate overall possible values of the other variable which we are trying to eliminate. And similarly in case of Bayes theorem we can like we are trying to we establish this kind of a relation between the conditional and the individual or marginal distributions. Now based on the concept of Bayes theorem we have something known as Bayesian networks for complex systems which involve many uncertain, but interdependent variables.

So, these variables they have some dependence relations between them like the just like the like we talked about the conditional probability. So, this means that they are the two random variables they are not independent of each other. If one happens then we know the about something or we have we may have at least partial information about the other event. If we know that like it is a sunny day today, then we have some knowledge about the let us say X is a variable that is which says it is the weather is sunny or not sunny and Y is another variable which is stands for the temperature. So, if we already know that today is a sunny day that is X equal to sunny, then we can expect that the temperature variable Y it is likely to be on the higher side.

Even though we may not know exactly what value it will take, but we have some idea about it. For example, I may think that like a temperature of 30 degree Celsius today is a bit more likely than a temperature of say 5 degree Celsius. Now, like when let us say that we have a complex system with many random variables which are dependent on each other. So, we build this kind of a representation which is known as a Bayesian network. So, it is basically a directed graph where every vertex represents a random variable and we have these kinds of edges between like different pairs of these variables.

So, what does the edge represent? So, conditional distribution is defined at each node over its parent nodes. So, these edges like as you can understand like whenever we have

these directed edges it induces some kind of parent child relation between the different vertices. For example, we can about vertex e we can say that b and c are the are its parents. for vertex f we can say that it has one apparent e and similarly about vertex a we can say that it has no parent, but it has two children b and c and so on. So, now we can whenever or at every variable we can define its joint distribution.

Sorry, we can define its conditional distribution conditioned on its parents. So, whenever I am writing the conditional distribution of B, I must condition on its sole parent A or whenever I am considering a distribution on E, I must consider or condition on both of its parents B and C. So, like you can take a look at the example which I have said here, here the assumption is that all of these variables they are binary variable. Now, for a like since it has no parent I am just defining its marginal distribution P of a probability of a equal to 1 is 0.7 which automatically means that probability a equal to 0 will be 1 minus 0.

7 equal to 0.3 because there are only two possibilities whose probabilities must add up to 1. Now, in case of b when I want to specify its distribution I must specify the condition. So, the condition in this case is a because it is parent. So, the it a can take two possible values 1 or 0.

So, I must specify the distribution of b for both conditions. When a equal to 1 I am specifying what is the probability that b equal to 1 and when a equal to 0 then also I am specifying what is the probability that b equal to 1. So, note that in this case these two probability 0.9 and 0.2 they must they need not add up to 1. Because, I have not specified the full distribution thus to specify the full distribution I should have also I have written that probability b equal to 0 given a equal to 1 that is equal to 1 minus probability b equal to 1 given that a equal to 1 that would have been 0.

1. Similarly, for the variable like say let E, you can see that it has two parents B and C and so I have considered four conditions. So, B equal to 1, C equal to 1 is one condition, B equal to 1, C equal to 0 is another condition, B equal to 0, C equal to 1 and finally, B equal to 0, C equal to 0. So, for all possible conditions on its parents I have specified what is the distribution on E. So, now the joint distribution of all of these variables can be factorized as the product of all these conditional distributions. So, like when we have a Bayesian network like this, it basically is trying to helping help us to define the joint distribution of these like all these random variables, but in a efficient way.

That is I do not have to consider the entire joint distribution, I can break it up into small small factors, one factor per variable and each of these factors is a conditional probability distribution which is base like. based on one particular variable conditioned on its parent variables. So, when we have a Bayesian network like this what do we do with it? One thing which we can do is probabilistic inference that is let us say some of the we know

the values of some of the variables for sure, then based on that we can try to predict what can be possible values of other variables. Like for example, let us say I know that c equal to 0 or d equal to 1 these are my observations. Let us say these two variables I am able to observe based on which I am trying to predict possible values of b.

Now, the model by itself does not specify this distribution, it specifies the distribution of B in terms of its parent which is A, but not in terms of C or D, but we can if we can use the Bayes theorem in a smart way, then we can calculate this probability distribution. So, for this there are there are algorithms which we can run on these Bayesian network which like those algorithms are primarily based on the Bayes theorem which we have already discussed which helps us to calculate this. And it is because of this reason that this network is called as the Bayesian network. Now, one more thing give like let us say that we are given observations of the of a particular variable. Let us say like there is a variable x which we have observed n number of times which means that we have done n random experiments and the results of those experiments with respect to the variable x are x 1, x 2 up to x n.

So, then like what do we say about x that is we can now consider x as a random variable, but what probability distribution will it follow. So, like we so, the idea is that we like we try to approximate this like x with one of the standard probability distributions which we are aware of for which we have a standard parametric form of the PMF or PDF. So, like we may never be able to choose a perfect distribution for x, but we can approximate it with some of the well known distributions because we know how to work with those distributions. So, like we need to consider first of all a known distribution family whose support set is consistent with the observations. So, if the observations are mostly binary then I may choose the Bernoulli distribution.

If the like observations I see they are all being real numbers any real numbers which can be either positive or negative I can choose the normal or Gaussian distribution. If I see that it is only positive real numbers I may choose the gamma distribution and so on and so forth. However, the every distribution is characterized by a PMF or a PDF. So, we like using the observation which we have we can construct a histogram and the histogram it like from the law of large numbers it follows that like if we have enough observations then the histogram converges to the PMF or PDF of the of that of a particular probability distribution. So, we will observe the histogram of the from the observations which we have got and see the it resembles most strongly the PMF or PDF of which known distribution and accordingly we approximate these observation or this random variable by that particular distribution.

However, what is left is to calculate the parameters. That is even if I have decided that this x must follow the Gaussian distribution what will be the value of mu and sigma. So, I

just write down this function $P(X)$ as the joint density of joint mass function all the individual observations because we have considered that the like the experiments which we have done are independent of each other. So, the outcomes of these experiments x_1, x_2 etcetera they we can consider them as independent random variables. So, in case of that as we know the joint distribution is simply the product of the individual distributions.

Now each of these like we since we have already chosen a parametric form. So, like this is a function of those corresponding parameters. So, I just try to maximize this function with respect to each of those parameters. So, if I can maybe I will try to differentiate with respect to the parameters and equal to 0 and so on. So, this is called as the maximum likelihood estimate of the parameters. So, why do we study these in economics? So, there are many applications of Bayesian networks in economics say risk assessment in management that and management that is the most important task the task of risk assessment.

So, these Bayesian networks they can complex they can model complex dependencies between economic variables and events which allowing allows for more accurate risk assessment that is like let us say that certain economic variables exceeding some some limit may be considered as a risky situation for us. So, what is the probability that such a thing will happen conditioned on other economic variables. So, that can be these kinds of things can be calculated easily using the Bayesian network. Similarly, in case of market analysis we may try to like represent the market dynamics by modeling the relations between different market variables such as supply demand prices consumer behavior etcetera. Also in supply chain management we Bayesian networks can be used to model the dependencies between the different stages of the supply chain.

Now, one more important concept which I should mention here is the expectation of the random variable. So, it is basically a probability weighted average value of a random variable which is defined in this way like if it is the DRV that is we just sum every possible value every possible value taken by the random variable multiplied by its corresponding probability mass function. In case of the CRV we cannot of course, add up because it is the support set is like continuous. So, we instead we do the integration and we while every value is multiplied by its corresponding pdf. So, like we have standard value we can easily calculate the expected values of x assuming if we assume what kind of parametric distribution x follows.

Like if it is this Bernoulli distribution then it can it turns out that its expected value is simply p. like in case of a coin toss. If it is it follows a Gaussian distribution then its expected value is simply the mean parameter. Also like these are some of the well known and famous properties of it is called as linearity of expectation and along with it there is

also a concept of variance. So, like while E of x expected value of x it tells us what is the likely value of x the variance it tells how far it can deviate from that likely or the mean value.

So, for example, in like again the variance of X can also be calculated easily if we know the parametric form of its decisions. Now, this is like the we in economics we often need the concept of expectation based risk analysis in many cases that is to understand the what is to be like the average case scenario and then accordingly we may take some decisions. Say for example, consider this case where I have 100 rupees to bet on a cricket match Now, I may bet x rupees on a particular team if they win then I will get twice the what I invested, but if I if that team loses then I will lose 20 times what I have invested. Now, suppose that the I somehow know what is the probability that this team A is going to win. So, in that case I can try that what is my expected gain if I bet any amount that is in the average case scenario how much money do I can I expect to gain back.

On other hands I can also ask these kind of more specific questions that is if I want to be left with at least 10 rupees that is I am willing to risk some losses, but the losses should not be too much. In that case I should be left with at least 10 rupees even if I lose. So, in that case what is the maximum amount I would like to bet or on the other hand I may be like optimistic also that I want to I have a target of winning 150 rupees. So, I may want to bet a high amount even if there is a loss there is a risk of a loss. So, how much should I bet? So, these questions like these are can be answered with the help of concepts of expectation.

Now, similarly the concepts of variance can also come in I want to predict the stock price y of a company based on some economic predictors. Now, suppose those economic predictors have some value x and let us say that I have also fit some kind of a model which predicts the possible value of Y given x. So, what is the expect first of the first question which I may ask is what is the expected value of the stock price that is the expectation or it we can also get a point estimate using the like a standard machine learning. At the same time I may also ask what is the risk that it may the stock price may follow below a particular threshold y.

So, this is something like a risky event. So, in conclusion uncertainty quantification is very important in economic decision making. We represent various factors as random variables and consider them to follow a known family of distributions. The family and the parameter values can be chosen based on the observations of each variable The relations between different variables can be represented through conditional distributions. The expectation captures the average case scenario while the variance suggests how much deviation from the expectation expected value is possible.

So, with this we come to the end of this particular lecture. In the coming lectures we will discuss about other aspects of supervised learning in particular how we can build a stronger prediction models using neural networks. So, till then Please take care and stay well. See you again. Bye.