Computational Neuroscience
Dr. Sharba Bandyopadhyay
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Week – 06
Lecture – 29

Lecture 29: Basics of Information Theory - II

Welcome. So we have introduced the idea of uncertainty of random variables. We have introduced the idea of reduction in uncertainty of a random variable by knowing another random variable which is also called mutual information. So we have extended that idea to call to in the case of continuous random variables by using differential entropy as a measure of quantifying the dependence between two random variables that is the difference in differential entropy of a random variable X and the differential entropy of a random variable X given another random variable Y. In our case stimulus and response. So using this mutual information we can now quantify how much dependence there is between a random variable with another random variable although we cannot explicitly model the relationship between the stimulus and response or two random variables X and Y.

So we cannot have a predictive model per say directly using mutual information but we can quantify how much dependence there is so that we can test models by knowing that okay the dependence can be up to this much. So does the model capture that much dependence. So in that way we can get an idea of how much our model works or how well the model works in the ideal sense because that is or if there is more room for improvement. Also we will later on talk about how we can take the help of mutual information in order to identify the possible features that a neuron encodes without directly being able to model it but empirically deriving a possible model from the data.

So another way of quantifying mutual information is through Kullback-Leibler divergence or Kullback-Leibler distance that is Kullback-Leibler distance. It is also called relative entropy. We define we simply write it as D_{kl} between two distributions p and q. Let us say p is a distribution of a particular random variable X and q is another description of the distribution of p of the same random variable. I mean basically the two distributions are defined on the same support that is the possible values taken by X.

What $D_{kl}(p, q)$ provides is how different or quantifies how different the two distributions p and q are or how far away one distribution is from the other in different kind of space. Here it is not truly a distance measure because it does not

follow symmetry, it does not follow triangle inequality. However it in some sense quantifies the difference between the two that is applicable in terms of communication in source coding basically shows the amount of or the extra amount of loss incurred by making a mistake in encoding a distribution p by another distribution q or representing one distribution p by a distribution q. So there is a physical meaning associated with it which is more applicable to communication systems or source coding. Here however we can use this for a variety of purposes where we can quantify the distance or the difference between two distributions without making any assumptions about the possible distributions or the underlying distributions.

So that is given empirically or observed p and observed q if we can quantify D_{kl} of the two distributions we can quantify how different they are and make conclusions based on that. So this does not require us to assume that okay p must be uniform q or q must be uniform or Gaussian as we will be doing in many other cases in computational neuroscience we will see that later on. So this allows us assumption free way of quantifying differences between two distributions. So but in another way that this D_{kl} is important to us is that this also provides a way of defining mutual information between two random variables x and y. So what is the definition of $D_{kl}(p, q)$ it is simply that it is summation $p(x)$ or an x is sum over all possible x's that is the support of the two distributions $log(p(x)/q(x))$.

So as you can see this is not symmetric and so it is not truly a distance although we will call it kl distance. So this definition of kl distance appears in the definition of I_{xy} also that is we have I_{xy} equals we have said that it is $h(x)$ minus $h(x)$ given y. So by our definitions of $h(x)$ and $h(x)$ given y we can show that it is also given in your handout that this is equal to basically sum over or sum over either all the x's and all the y's $p(x, y)$ this is the joint distribution of p of x and y $log(p(x, y)/(p(x)p(y))$. Now if you look at this this is very similar expression as kl distance. So in this case we are looking at two distributions now this let us what was p of x before is nothing but the joint distribution $p(x, y)$ and what was q of x before is nothing but the product of the marginal $p(x)$ and $p(y)$.

So this mutual information then turns out to be the d_{kl} of the joint distribution $p(x, y)$ and the product of the marginals $p(x)$ into $p(y)$. So now the interesting way of looking at mutual information is how different is the joint distribution from the product of the marginals $p(x)$ and $p(y)$. So remember that when two random variables x and y are independent then p the joint distribution $p(x, y)$ is the same as the product of the marginals $p(x)$ into $p(y)$ and in that case when x and y are independent as we expect the mutual information I_{xy} in that case is 0 and the more different the joint distribution is from the product of the marginals that is the

mutual information keeps on increasing. In other words the mutual information is the distance between the joint distribution and what the joint distribution would have been had the two random variables been independent of each other. So had they been independent of each other then the joint distribution $p(x, y)$ should have been $p(x) * p(y)$.

However the observed joint distribution $p(x, y)$ is whatever we are looking at and so the distance between them or the kl distance between them is the quantification of dependence between x and y that is the distance from independence. So now if we go back to our idea of stimulus and response then we have the stimulus and response all we need in order to compute this mutual information is just one thing which is the joint distribution of p s and r. What is p s and r? $p(s, r)$ is the joint distribution this is simply the output of experiment where we use different stimuli depending on the question we want to answer and make observations about the response depending on the way we want to define the response measure and by varying the stimuli and repeating the stimuli multiple number of times we get an estimate of this joint distribution. So if as we discussed in the earlier class let us say our stimulus takes on values s_1 up to s_n and we observe the responses are as either r_1 up to r_m these are the possible responses that we observe and the stimuli are s_1 up to s_n. Let us write this down in the matrix form here s_1 up to s_n.

So each row here is like this depicting its stimulus and we already have the marginals let us say we give the stimuli equal number of times or present the stimuli equal number of times so it is 1 by n in each case or maybe it is naturally equally probable. So these are all 1 by n so this here is our marginal probability $p(s)$ that is 1 by n each time. So now in the experiment what are we doing we present the stimulus s_1 let us say capital k times and in each of those k times we observe some a few times r_1 a few times r_2 and so on up to r_m. So based on these values we get an estimate of probability of response equal to r_1 given s equals s_1 probability of r equals r_2 given s equals s_2 let us say our k is 20 and our r_1 is our the number of times we observe r_1 for stimulus 1 is 2 then this turns out to be 1 by 10 and let us say we observe this r_2 0 times so this becomes 0 let us say we observe r_m 5 times then we have probability of r equals r_m given s equals I am sorry this is s_1 s_1 is equal to 1 by 4 and so on. So we have the probabilities r_1 r_2 up to r_m given the stimulus is 1 which is this particular row here and based on what we have we have 1 by 10 and multiplied by probability of s equals s_1 which is 1 by n which we already have there this is 0 and this is 1 by 4n that is we have now multiplied this by probability of s equals s_1 in order to get the joint probability.

So this way we can fill up this entire matrix that is obtain the joint probabilities ps equals si and r equals rj so this is the ijth element in this so this is the jth this

is the ith this is the p_{ij} let us say. So now we have with the different i's and j's given the experiment we can fill up this matrix and we have the by summing these along the columns we have the probabilities marginal probabilities of r that is $p(r)$ this is r equals r_1 up to probability of r equals r_m. So all the components required to compute mutual information is present that is we have the joint values that is sum over sum over we have s equals s_1 to s_n and r which is we can write this sum as basically not s_1 to s_n we can write it as i equals 1 to n which is basically it will sum from s_1 to s_n and this is j equals 1 to n probability of s equals s_i, comma r equals r_j $log(probability(s = s_i, r = r_j)/probability(s = s_i) * probability(r = r_j))$ and by summing them over the all possible combinations of stimulus and responses we can get mutual information between the stimulus and response. There are number of issues practical issues that will come about when we actually do the computation because of associated bias in measurements in the estimation of entropy and mutual information. So you will you can see from the KL distance approach that and also the conditional I mean entropy minus conditional entropy approach that mutual information is greater than 0 that is it is always I mean it is non-negative it can be 0 when the two random variables are independent.

So there being non-negative even entropy is also non-negative because the lower limit the lowest possible uncertainty is 0 and in the axioms also it is the measures have to be positive in that sense and you can see that negative logarithm pi log pi is will turn out to be also always positive or non-negative and so because of that even when there is no dependence we will get some spurious dependence present from the noise itself because no matter what it is always positive. That leads to many issues in the estimation and so there are ways that in I mean if you are going to be actually implementing these you can take resort to other ways other means of removing that bias which is de-biasing information theoretic estimates. There are number of ways that have been used and will provide references for those however for our purposes we will take this as one way to compute mutual information and now we will take this forward in order to see how mutual information and KL distances can be applied in terms of understanding neuronal coding. So if we go into the summary what we have shown is if we can if we think of a random variable X or let us say the stimulus and response then the connection between them is quantified by this or dependence between them is quantified by this mutual information of $I(S, R)$ or $I(X, Y)$. Now if we think of this circle or representing the uncertainty in the random variable X or S and let us say this circle represents the uncertainty in R or Y let us say so this circle represents $H(S)$ and this circle represents $H(R)$.

So the part that is remaining outside of R in S is the uncertainty remaining in S given we know Y. So we know Y so that means this shaded part is $H(S)$ given R we know R or Y this is $H(S)$ given R. Similarly we can think of the vertical shaded region here is the region of R uncertainty of R that is present if we know S that is $H(S)$ goes to 0 then the vertical shaded region provides us $H(R)$ given S that is the entropy remaining in the response given the stimulus. So the intersecting region which is the horizontal shaded region is what we call the mutual information $I(S, R)$ which is nothing but $H(S)$ minus $H(S)$ given R or $H(R)$ minus $H(R)$ given S. So in other words in this case the intersection in this Venn diagram form the intersection is providing us the dependence or representing the dependence between the two random variables.

So another important idea that is required in order to go forward is the idea of the data processing inequality which simply says that let us say we have I will not go into the technical details of this we let us say we have a chain of random variables that is from X we get to know Y from Y we get to know Z and it is such that if we know Y then X and Z become independent that is what we mean by a mark of chain here that is X and Z are conditionally independent that is given Y. If we have a situation like this so let us say we have a stimulus and let us say from there we have a response now given from the response we have an estimate of the stimulus let us say another random variable S' which is something which is also a random variable and given R S and S' are independent because S' is only dependent on R the S in S does not influence it in any way once we know R. In such a scenario what we can we can show or what has been shown is that the mutual information between X and Y is greater than equal to the mutual information between X and Z that is by processing the response or a random variable to a new random variable Z we actually lose information that is the this is what we call the data processing inequality. So what this tells us is that if we use some data to know about a particular random variable let us say we take R to know about S and then we process R further to a new random variable then the amount of information that is going to be there in the new random variable about the first random variable is going to be less than what was there between the previous random variable and the original random variable that is R and S. So that means there is a lower bound in terms of if we if let us say we somehow estimate $I(X, Z)$ without knowing Y then $I(X, Z)$ is a lower bound of $I(X, Y)$.

So this will come up when we look at discrimination based decoding with the of stimulus and response. So with these all these ideas of information theory we will now later go to applications of them and we will first consider the case of applying information theory in order to derive possible models between stimulus

and response in a particular way that is what the methodology is maximally informative dimensions. So remember we motivated the idea of going into information theory from the fact that we can only model stimulus response relationships only up to a certain degree or order. We definitely can use linear models which we did using the spike triggered average which can be extended to higher order models but obviously there is a limitation as to how far we can go and so we took the help of information theory or we said that we will take the help of information theory in order to build or understand what models can be between can be possible between stimulus and response. So given this background of information theory in our two lectures now in the next lecture we will start off with the applications of the ideas first with the idea of maximally informative dimensions. Thank you.