Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

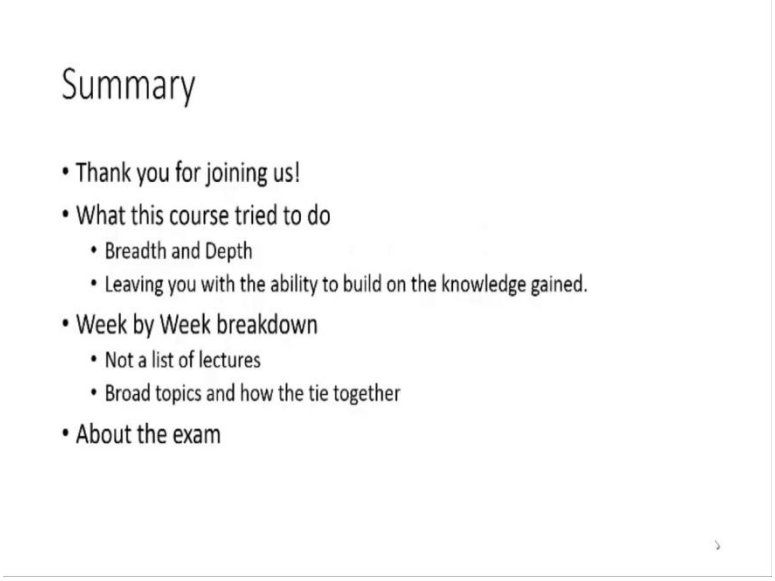
Module – 09

Lecture – 47

Summary + Insight into the Final Exam

Hello and welcome to our last lecture for the course Introduction to Data Analytics. This is Prof. Nandan and in this lecture I just briefly give you a Summary of all the topics that we have covered during this course, perhaps tie some topics together. And in the last few minutes, I would try to give you some insights into, what we will be looking at during the final exam.

(Refer Slide Time: 00:42)



So, first of thank you for joining us, I speak for both Prof. Ravindran and myself in saying that this has been a very valuable and enriching experience for us as teachers. You know one thing to understand this that, this whole online learning, teaching model is relatively new in the Indian context and so, for us as teachers as well as you as learners. But, the hope is that you know we have been able to reach wider and hopefully, more people have been able to kind of get access to our content and are able to apply some of

the things that they have learnt in this course, either in their academic life's or in their jobs, their organizations or may be even their personal life, but in any case we have be happy to have undertaken this end hour and you know, from the bottom of our hearts thank you for joining us.

What this course in many ways try to do is... So, this course was not a course strictly in machine learning, it was not a course strictly in inferential statistics, it was not a course in probability one over one alone, you know. So, in many ways we try to give you the broad brushstroke associated with the entire landscape of data analytics. So, in some sense hopefully you are now familiar with a lot more terms and lot more techniques and concepts than if you taken a far narrow course. So, the hope is that we covered the breadth of data analytics and it is never possible of course, to cover the full breadth, but to a reasonable extent, the fundamentals and basics of most concepts.

We have also try to go deep in certain areas and it is primarily been more in the conceptual front to give you insight into the house in the wise and the what is of the different techniques, why do some techniques work in some places, why do not they work, what are things to watch out for and so on and so forth. And so we also hope that is given you well rounded enough knowledge to build upon further.

So, you know, but this highly changing landscape of different software packages that come on different days, we have try to make this little a less tutorial like, we given you some derivations, but this is not then fully about the theoretical math of the course. Because, the idea is that, that kind of depth is something that depending on your place of application and depending on your more specific area views of data analytics. You should hopefully be able to build upon given the knowledge that you gain in this course.

In this lecture, I am going to go week by week and just talk about, you know what we hope to cover there, but this is not as you will see shortly, not just a list of each lecture and talking about it. So, to some extent the topics themselves are beautifully aligned with list of lectures, but please do not think that each bullet point as I go from one week to the next is a comprehensive list of lectures, it is more the topics in the concepts. So, I will just talk to you little bit about the broad topics and how they tie together and in the last few minutes as I mentioned, I will give you some insight into the example.

(Refer Slide Time: 04:19)

So, in week 1 we really started off with the course overview and the hope was both Professor Ravindran and I were able to give these, you know long overview lectures. We had an intro lecture and then I gave these, fairly I would say long lectures that just try to cover the entire landscape in one to two hours. And so hopefully that gave you some insight into what the overall course was about. The rest of the week was spent in descriptive statistics, we talked about graphical approaches, there are different ways of representing data.

And in this particular case with graphical approaches, you really talking about different ways of visualizing data. An important concept that was also covered in this lecture was, that data it comes in different types and we spoke about continuous data, categorically data, we spoke about actually we broke it down sportivate versus qualitative and then quantitative broke down is continuous versus discrete and then, categorical broke down as ordinal and nominal, but that is an important concept that we kept on saying at different points of the course.

We moved on from visualization and we moved on to kind of individual statistics, statistics that represent or matrix that represent, you know central tendency. So, those were the mean, median mode and then they were these measures of dispersions. So, thinks like standard deviation, mean absolute deviation and a lot of the time was not I mean... So, some amount of time was spent on explaining how they work, but some

amount on where they work and how they work and why some are suited in some situations and others in other situations.

(Refer Slide Time: 06:11)

We then moved on to probability distributions and inferential statistics. So, we spent a good amount of time, you know introducing the concept of random variables as different from, you know regular variables or constants that we used to. So, what is it mean when we talk about random variable and how do you represent random variables, through you know probability density functions, cumulative density functions, probability mass functions and about how probability distributions can again be discrete or continuous and you know, how they get treated differently.

We also spent some time in talking about a set of specific probability distributions, namely we spoke about the uniform, the continuous version, the discrete version, we spoke about the binominal, the ((Refer Time: 07:02)), the geometric, the exponential as well as the normal distribution, finally. In the latter half of this week we moved onto inferential statistics, we spoke about this core idea where some times we have a phenomena that is generating data and what you have is essentially a sample of that, what you have therefore, in terms of data is a sample.

The data is a sample partly perhaps, because you were only able to access a part of the data, the overall data or sometimes it is a sample, because there is a broader phenomena that is creating this data and we just take a subset of that. And the core of this inferential

statistics has been about how you do not just want to make statements about the data that you have at hand, but more about the underline process that is creating this data and that was our departure from descriptive statistics and that is what we try to motivate in those lectures and we also spoke about this idea of single sample tests.

So, tests where you have a sample and the sample is coming from this broader universe and how do you compare these two specific numbers or tests hypothesis for parameters associated with the population.

(Refer Slide Time: 08:29)

Within got deep into inferential statistics, so we went into two sample tests, the idea that you might have two sets of data that represent two different populations or two different phenomena and how can you compare them. You can compare their means, you can compare their standard deviations, you can compare their proportions and so on and then, we got into a lot of other concepts. So, we spoke about type 1 and type 2 errors, the errors associated with saying that the null hypothesis is incorrect or you rejecting the null hypothesis, when actually the null hypothesis was not incorrect.

And the other error, the type 2 error which is the idea that you fail to reject the null hypothesis when you really should have, you also extended that to confidence intervals another important concept in inferential statistics and we spoke about ANOVA, analysis of variance and the tests of independence. So, with the ANOVA we spoke about how you are able to frame more complex hypothesis, like the mean of population A is equal

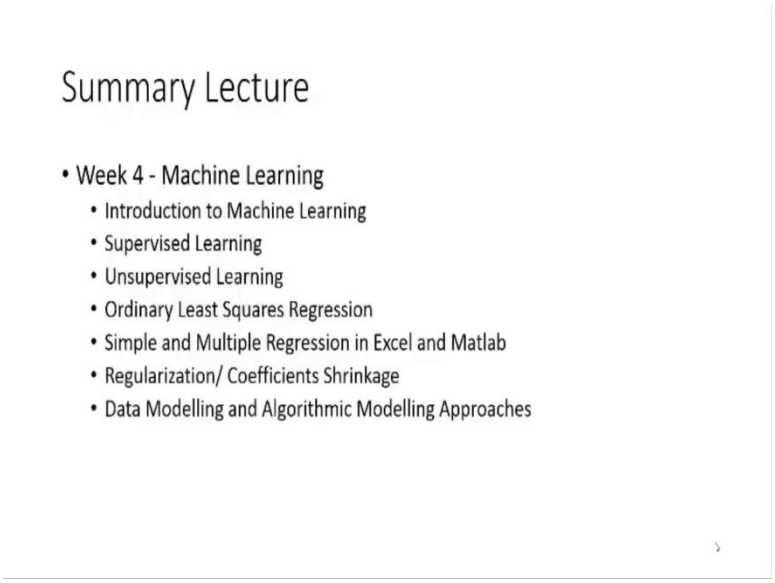
to the mean of population B is equal to the mean of population C and so on.

So, with ANOVA test of independence, really for the first time we were able to look at more than two states, more than essentially at two samples if you can think of that way, more than two states of a particular variable. So, upon till the one sample and two sample tests at most at two sample test, you had this variable that could take on two possible states and we would compare them.

But, for now you have, you still looking at categorical variables with the ANOVA, especially in terms of the input variable, but it is taking on more than two states and with the test of independence, both the two variables that you are interested in whether you want to think of the them as input and output variables or just two independent variables, two variables could take on two or more states and you are interested in seeing if these two variables are independent of each other or not.

We also use this week to come up with the short introduction to regression. What is regression all about, what is it trying to achieve and what kinds of problems statements do we address in regression. A part of the focus is been about talking about how typically the input and output variables associated with the regression or continuous quantitative variables and how this possessive regression gets used for prediction as well as interpretation of the relationship between the input and output variables.

(Refer Slide Time: 11:18)



We then choose to introduce machine learning and the whole idea of mixing up the regression classes, the machine learning process is very carefully thought through. Because, machine learning being relatively new feel then regression, but also one where a huge subset of it which is supervise learning really focuses on many problems that regression at one point try do and perhaps does it much better, but to kind of not think of them is isolated topics, but think of regression very much as one of the tools in the tool box of machine learning irrespective of you know which when they came in to prominence was importance.

So, many of the concepts in machine learning just is equally applies to regression. So, we introduce the idea of supervised and unsupervised learning, where the supervised learning we said there is this notion of an explicit output variable that you are looking to predict often the way it is formally defined is that it is labeled data, but it is un supervised learning you have unlabeled data or basically there is no one single output variable that your trying to you know predict or to classify.

We then were able to go back to the regression and talk about using the ordinary least squares regression, we saw derivation associated with how we get to that. And we also spoke about how we can do regression and excel in mat lab, what are the various concept associated with it we spoke about idea using back words and forewords step wise regression, best subsets regression and so on.

We then went on to the concept that is very important with a regression called regularization and coefficients shrinkage, but it is in some sense extendable to other methods. But, we focused it more on the regression front and the idea with the regularization was how do you focus on models that do not windup over fitting the data, because you can always introduce enough complexity into the model make the model, so complex that it winds up doing a very good job of representing the data at hand, it really does not do very good job of predicting new data, because all the complexity and all the use of many variables more than those required they want getting used to kind of just over fit the data to just do an exact replica of the data.

Highly related to that is also the concepts of co efficiency shrinkage, especially when you have the couple of variables that are highly co related to each other. A couple of input variables that are highly co related to each other and we call this the problem of

multicolinearity and how there are methods of co efficient shrinkage like rig regression or laso regression which attempts to not allow coefficients to balloon in typically opposite directions in an effort to kind of better represent the data.

We also give you some inside in to the whole data modeling versus algorithmic modeling ideas and the core concept that we were trying to install there is we introduce the new approach call the k nearest neighbors. And contrasting the k nearest neighbors with the regression based approaches, really spoke about how which some approaches which we are going to call the algorithmic modeling approaches, you do not have to have an explicit mathematical model.

And in fact, it was better represented by a set of algorithmic instructions rather than a mathematical model that connects the input to the output or at least the mathematical model is not constricting the you know potential relationship that the input and output variables could have, because an algorithmic representation could be more flexible in terms of representing this relationship and the K-NN versus regression hopefully capture that dichotomy.

(Refer Slide Time: 15:57)

In our next week we focused on supervised learning, which is as we just previously discussed this area where you have an explicit output variable and you really trying to predict that output variable. And as you can see in this set and this is commonly used nomenclature of having regression and classification techniques, here the word

regression unfortunately while it is being used really means problems, where the output variable is continuous and quantitative rather than discrete or categorical.

And for those where the output variables diacritic or categorical you had these classification techniques. And in that light we spoke about logistic regression and support vector machines and decision trees. Now, all of these techniques are definitely work as classification techniques, while classification and regression trees can do both regression as well as classification, with logistic regression we introduce the subject we also thought you how to train logistic regression classifier.

Similarly with classification regression trees also known as cart or more broadly cart becomes algorithms the decision trees are just a way of talking about such trees. Support vector machines was also introduce along with some derivation on how the classification through the support vector machine happens and the core concept there being that you have been really creating this hyper plane and you are trying to maximize the distance between different classes to the hyper plane, which is actually splitting the input space and thereby enabling you to build a classifier.

And we also spoke about the case, where you have some kind of violations, where some data points of one class over spill to the other side of the hyper plane and the case where there are non-linearly separable data and the use of kernel transformations. Now, we also spoke about this concept of the bias variance dichotomy and the bias variance dichotomy talks about the concept of over fitting versus under fitting, where the idea is that there is the sweet spot in terms of the model complexity for a given data set, for a given underlying relationship between input and output variables, there is a sweet spot in terms of the model complexity.

And when you have too little model complexity, you are essentially building a regressor or classifier that is bound to have a high amount of buyers, which means that irrespective of how many times you repeat the process with different data sets, you are always going to be chronically under representing or over representing certain reasons in the input space in your process of prediction. Now, when you move to the other side, which is you have a model that is too complex, you over shocked your sweet spot for this given data set I mean for this given underlying relationship.

You are essentially likely to create a model with high amount of variance and low

amount of buyers, which just means that each time you do this task, you are going to get a completely different model on average, if you are to do this millions of times, the average of volume models might not have much buyers, see you are not chronically under representing or over representing some time we just do not know what we are going to get each time you do it.

And so at a given exercise where you have a given data set and you are choosing to build a model it could stood be a really bad model. So, this is known as the bias variance dichotomy and it that really segued into the idea of how do you therefore, evaluate models and therefore, select models and that like we spoke about some of the matrix, some of the matrix that penalize model complexity beyond the point. Because, you can use that penalization and conjunction with trying to do the best job you can fitting to the data to come up with that balance.

But, we also spoke about some more computationally intensive, but more effective I would say techniques like cross validation, which do not just do a blanket penalization based on complexity. But, instead allow you to hide some data where build the model on other data and then try to actually do the job of predicting on this new data set and essentially do this kind of out of sample validation and there by make decisions on both the complexity and the type of model.

A small note on the use of a linear discriminate analysis and the quadratic discriminate analysis, you can also think of them as methods those were essentially this is the last bullet point that I am talking about we made a careful decision not actually include those lectures. If you look at this week this was a very packed week with the lot of content and we more ours in lectures and so while this is included in the syllabus we made a conscious decision not to give out lectures in this topic.

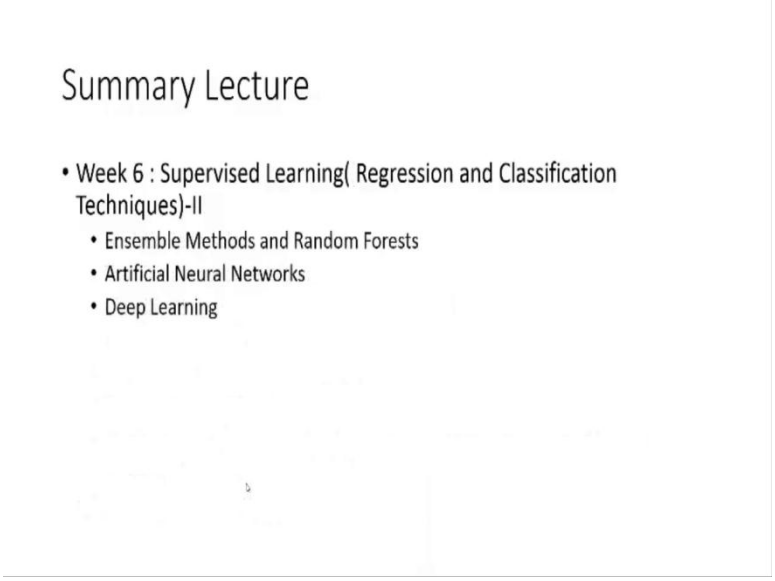
And this partly become be actually we had discussed a lot very important classification techniques, relatively I would also same more popular once. But, it just give a some insight the way linear and quadratic discriminate analysis works is that it tries to create a distribution, a multivariate distribution essentially in the input space associated with each class and come up with the decision boundary based on the probabilities associated with this multivariate distribution of each class for a given input point.

So, we do not wait actually ask the question of each multivariate distribution we instead

create these multivariate distributions for each class, because this is primarily use of classification and it is a multivariate distribution in the input space and we basically come up with decision boundaries and decision boundaries are linear or quadratic, but essentially you come up with these decision boundaries based of the distribution you built. So, right of the bad tomorrow and some on comes and says here is input can you classify this for me, you can see on which side of the boundary this falls.

And more often than not the typical approach is to use the multivariate Gaussian or the normal distribution and an important distinction between the linear case and the quadratic case has to do with the assumption of equal variance. So, if you make the assumption of equal variance is essentially using the linear discriminate analysis, if you do not assume equality of variance using quadratic discriminate analysis.

(Refer Slide Time: 23:18)



So, in the next week we really took on some more advanced topics in supervised learning, we spoke about the use of one ensemble methods and most specifically we elastrator that with random forests. The idea behind ensemble methods and random forest were really to do with the idea that you can sometimes have a deterministic algorithm on a given data set. But, sometimes it really helps to inject some form of stochasticity, either in the algorithm itself or in the data and thereby creating multiple versions of this algorithm.

More often the knot is in the form of versions of this same algorithm, but sometimes you

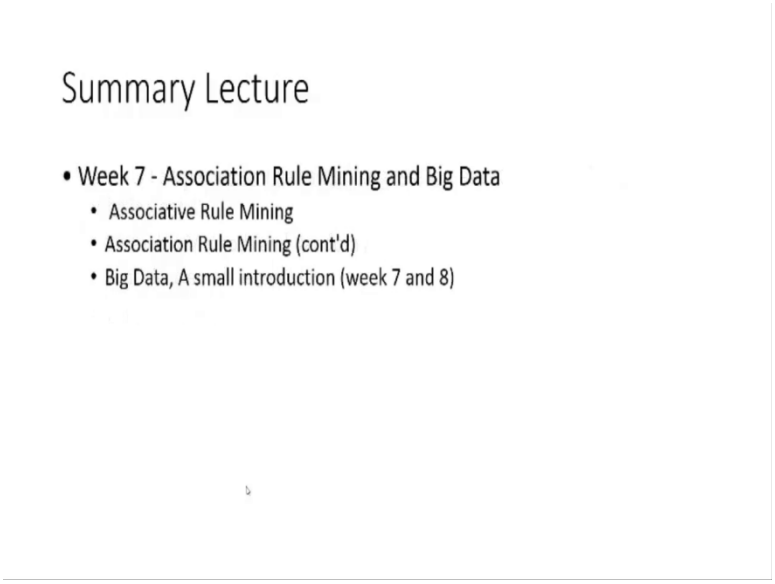
can also have completely different algorithms that are trying to do the same job. But, the idea is that you create multiple versions and then you aggregate their predictions, whether the predictions or registrations or a classifications you try to kind of aggregate them to come up with one final prediction. And this sometimes helps especially this helps in these method, where fairly like greedy algorithm has gotten stuck into some form of sub optimal.

Because, as you would have realized by now a lot of the machine learning techniques relaying on optimizing some function or the other whether it is support vector machines, whether it is you know ordinary least square regression or whether it is classification and regression trees here at some point you know trying to optimized some function and there by improve your ability to predict, but as with any optimization, the most common thing that is gets discussed is getting stuck and sometimes this kind of ensemble approach, where you just forcefully introduce some noise into the whole process helps you break out of that and overall prediction kind of improves.

We spoke about how random forest is one way of doing that with classification and regression trees. We, then address the topic of artificial neural networks and extremely popular and becoming even more popular approach to machine learning and predictive analytics. We introduce the topic in terms of how artificial neural networks or constructed and what is their motivation or inspiration, we spoke about how we can use the back propagation algorithm to construct the artificial neural networks.

And we finally, took it to towards deep learning in reality it is essentially just this neural network with many, many, many layers and these really new and innovative techniques of actually being able to do the computation and the math to create such elaborate networks, which was previously not really feasible more than anything.

(Refer Slide Time: 26:32)



We move onto week 7, where we talk about association rule mining and big data with association rule mining we spoke about this in the first context of it being there would two topics in unsupervised learning and one of them was association rule mining, the other being clustering analysis. We introduce the idea of association rule mining and spoke about it is fairly, you know most common use case which is the market basket analysis.

The idea that you can create this table, where the rows represent various instances or transactions that have happened and the columns, really represent features and what you have is essentially a binary table of, how frequently these attributes or columns have been seen with each instance. And with association rule mining we really trying to make a statement not about the rows themselves, you using the rows, you using the instances, but to make statements of the form that typically when column a occurs column b also occurs in the market basket context market which in visions super market, we are really saying things like the rows here or people making actual transactions of purchases in the columns of the different products that they choose.

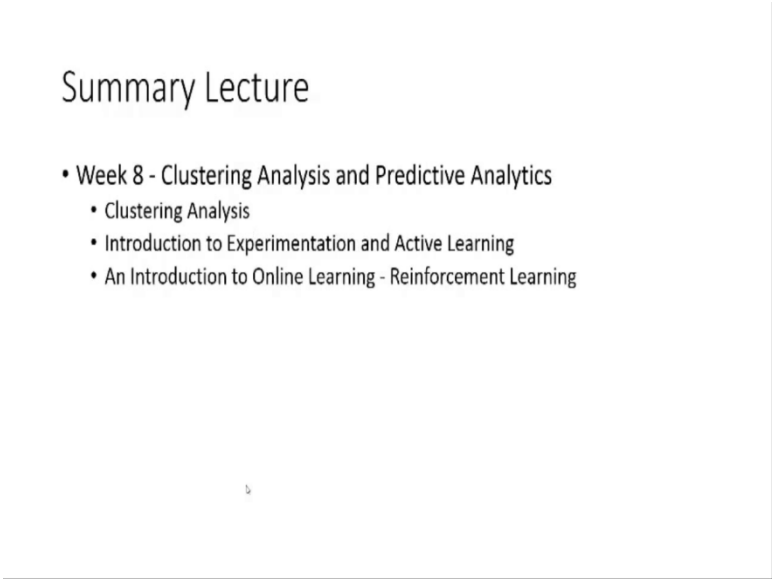
So, the end result being that you want to make statements like people who bought coffee also bought milk or more complex things like people who bought coffee and milk also bought sugar. So, association rule mining, associative rule mining is useful there and, but it really extents it extents to any contexts, where you can create such a rules and extents

to any context, where you have instances as rows, features as columns you have really binary data set in that sense.

We briefly spoke about different ways of measuring, how good a rule is and these are called interestingness measures and some of the other challenges that are associated with scaling, association rule mining. Finally, we spoke about a buzz word that we all been hearing about and this really spilt little over week 7 and week 8. The whole idea being that we further was big data, how is that different from the broader topic of just data analytics or data science in general what this means and why is it that we are seeing, what do we mean when we say big data, you know the does it just mean volume, we afford other v's associated with velocity variety, so on.

We also spoke about how some of the challenges, but the regular techniques take on a different form with big data, when you have a much larger data set. What does some of the computation or challenges that are associated with that or some methods more I am enable than other methods and how do you scale up such that your able to actually process huge amount of data and still get insights from it.

(Refer Slide Time: 29:46)



Finally in week 8, we presented clustering analysis one of the other unsupervised learning techniques and we spoke about how this is used primarily to segment data to take unique instances and group them together. So, here in some sense you have this data set which is represented by these instances or transactions or whatever it is that you want

to think of his rows and these rows share a common set of features in a based on the common set of features are they share and the different values they take up with the respect to these common set of features, you want to create cramps or groups or segments or clusters of your data points of your rows and the different ways of going about doing that.

We spoke about some core concepts associated with clustering, but we also introduce to concretes methods they came in the hierarchical forms of clustering and how they work and why they work and so on. Towards the end we spoke about you know some amount of prescriptive analytics and the idea here was that sometimes you do not have data and this lack of data, you know could mean that you do different things, a different forms of analytics or I should refresh that you sometimes you do not have enough data and actually small correction this should not be a predictable analytics should be prescriptive analytics.

But, my going back to my point you know you have you either do not have data at all or you have very limited data and in some sense you have some control over what data it is that your trying to get a hold of. So, in the case of a experimentation much of the focus is on figuring out, where in the input space do you want to generate data or another way of rephrasing that or another completely different way of looking at it is can you go about systematically and purposefully making changes to your input space by changes we just mean setting it to some value and then getting a data point getting an output associated with that.

Now, that has a lot of advantages, because you can get rid of problems multicolinearity and because you are very carefully choosing, where you are getting outputs you are able to analyze the data much better or with formal limited data and that is the same whole premise of active learning as well. Active learning starts more from a space of saying I might have some results already, but can I go, but sequentially queering the system and choosing one at a time, you know sequentially rather than a parallel deployment of 10, 20, 30 data points.

But, can you give me based on the results, based on the outputs that I see I am going to the come back to you and say can you please give me a data point at this point and get an output specifically for a given region and update your module or update your system

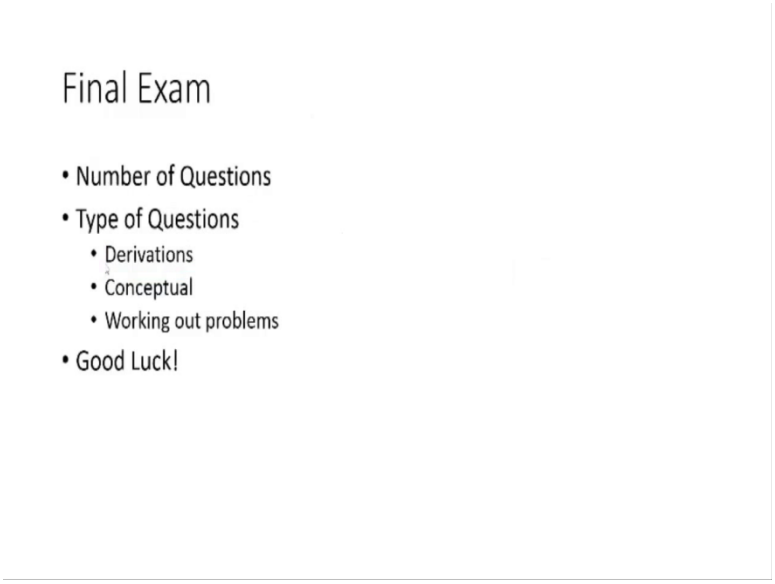
based on that. And again the whole premises that you should now technically need a lot fewer data points to learn about the system out of module the system than traditional system, where you only have observational data, you had no choice or right and saying, where you got your labels, where you got your output variables, which points in the input space for which points in the input space did you get your outputs.

This also links a lot with reinforcement learning, where again you have a say in which points of the input space you get an output, but here the output carries some value, the output is essentially also seen as a reward, you know seen as some form of online experimentation, where online here is not refer to the internet, but as much is online refers to this idea that whatever it is the output that you getting the sole purpose of you doing this querying of the system is not a just learn about the system, but it is to also to maximize some reward from the system.

So, you are essentially live you can think of it many context you can think of it a manufacturing process, you can think of it as a website or anyone who is choosing to experiment, but their experimenting with real system, a system that is actually creating products and services and tasks that are actually going to the end user. So, how do you go about experimenting on such a system, where at the same time as you wanted to learn about the system, you also want to provide your end user with the best service or the best product?

So, how do you go about experimenting to learn, but at the same time also to do as well as you can while your experiment, hope that gives you an overall picture of the course itself.

(Refer Slide Time: 35:13)



In terms of the final exam we are really looking at an exam, where you are given about 30 to 40 questions and they are all going to be multiple choice and that we have had some concern from students in terms of that types of questions and the one thing that I wanted to kind of emphasizes is that the questions themselves will not need you to do a derivation, you know partly because there multiple choice when I going to say here is the first five steps of the derivation, what is the next step kind of questions there you know derivations are important.

But, they are important in terms of from this courses perspective that you understand how their done and why their done, what is it that you are deriving. And so some sense lot of the questions themselves should be conceptual, we are not going to get like a fill in the blanks exercise, what is the next step of the derivation. And we are going to focus more on the conceptual part of the courses content and you know in some sense, the y is the what is where do some methods work, where do not they work, what will be the likely out coming if I do particular method or on a particular data set you know.

So, the very, very conceptual are in therefore, you know any amount of working out is going to be more back of that envelop. So, you can pretty much take pen and paper and you should be able to work out some set of and you know calculate us fine, but nothing you really do not need a compute of much of this and there are I would say fairly limited number of questions that required you actually put even pen and paper.

Most of them have to do with how deeply you understood the concept and how deeply you understood, how the methodology works or how a particular concept manifests self when you carry out a methodology. And so you know I do not think you should have too many worries about doing complex derivations or doing you know complex math they figure out the right number. Focus I would say more on understanding the concept associated with be it methodology, be it machine learning concept, be it inferential statistics versus descriptive statistics or be it derivation focus on the concept than in some sense the details.

And finally, good luck for your exams and we look forward to having this kind of engagement in interactions with students going forward.

Thank you.