Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department Computer Science and Engineering

Indian Institute of Technology, Madras

Module – 08

Lecture - 44

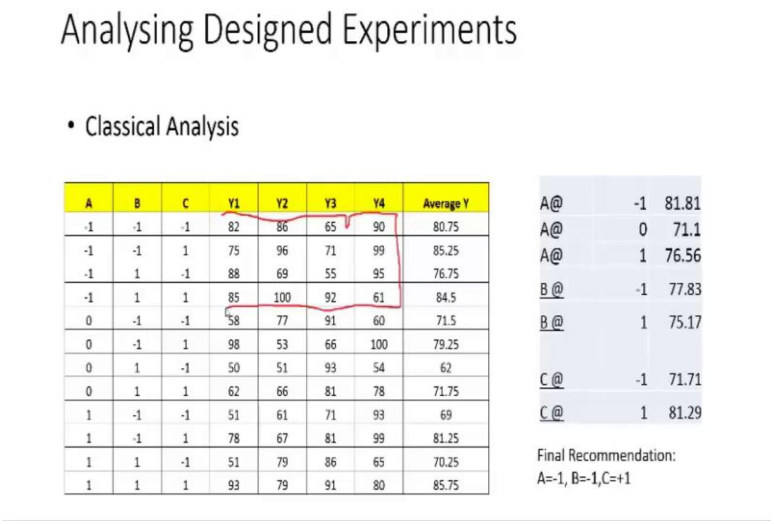
Introduction to Experimentation and Active Learning - 1

Hello and welcome to our second lecture in the series, where we talk about Experimentation and Active learning. In the first lecture we briefly spoke about, we motivated the idea we using active learning experimentation or re enforcement learning in a broader data analytic setting, where we said these are some approaches that are fairly relevant when you do not have data, when data does not exist or the data exist and it is not enough or you have only partial data.

In the first lecture, we also went into design of experiments and spoke about, how perhaps an approach, where you sequentially change one variable at a time need not be the best way to conduct an experiment. And we also spoke about something called orthogonal arrays and specifically, there we spoke about a form of experimentation called the full factorial design, where it is essentially a complete enumeration of a discrete input space, where even if you have continuous input variables you break them into 1, 2 or may be sometimes 3 discrete points.

And you essentially do a complete enumeration, which means that every variable is set to every possible value it can be set to with relation to every other variable being set to all their possibility. So, all the possible points in the input space are essentially looked at and that was essentially looking at the design meaning, what points in the input space to you choose to experiment. In today's lecture we briefly take the just full factorial, which is a very basic design and talk about some approaches that are used in analyzing such an approach, such a design.

(Refer Slide Time: 02:07)



So, jumping into the subject in this slide, what we have is a full factorial design associated with three input variables A, B and C. And the idea here is that what we have is a full factorial design, because A can take on three values minus 1, 0, plus 1 and B can take on two values and C can take on two values just minus 1 plus 1 and that gives you 12 combinations totally and we have taken, the output variable is Y. Now, I call them Y 1, Y 2, Y 3, Y 4, only because they are Y 1, Y 2, Y 3 and Y 4 are replicates, so it is a same core variable.

But, essentially you conduct the experiment at, the settings for instance minus 1, minus 1, minus 1. You conduct that experiment 4 times; again it can be a parallel effort or a sequential effort either way. What we mean by that is, when we say we conduct the experiment 4 times, you might have one experimental unit and you separately conduct the experiment 4 times on it or you could have 4 experimental units and you might, you choose to parallely try the same setting of minus 1, minus 1, minus 1 on those four different experimental units.

But, essentially this is almost a setting, where you conduct 48 separate experiments and in design of experiments, language that just called 48 trails or 48 runs and these are essentially your results, these are your outputs. And the convenience of this is sometimes you can look at this raw data or you can look at the average Y and this is nothing but, the average for instance for the average 80.75 comes from taking the average of these four

numbers. So, each row is average and it is represented and that is the, those are the results that we haven, this is essentially what we look to analyze.

So, what is one way that you can take these results and arrive it a conclusion of what values A, B and C should be set to, because that is kind of the goal. The goal is to figure out, what is said A, B and C 2, I mean accurate one of the goals can be on to what values to set A, B and C; such that you get best Y and here we are going to treat best as the highest value. So, how should I set value A, B and C, so I get the best Y?

One approach is called the classical analysis. The idea behind classical analysis is to take each variable individually, each input variable individually and ask the question, at what setting am I getting the best results. So, A is set to minus 1, A is set to 0 and A is set to plus 1. So, if I am ask the question, what is my Y on average when A is minus 1, what is my Y on average when A is equal to 0, what is my Y on average when A is equal to plus 1 and I look at these three numbers and I will choose the value of A, where my average Y is the best. I will similarly do that for to B and C and I will come up with a recommendation on that basis.

So, what is that look like for this data? It is a fairly straight forward calculation, when A is set to minus 1 you essentially get 81.81. So, it just means you can think of it in many ways, you can think of it as the average of these data points. Right here, what I circled and the average of these data points or you can think it as the average of these data points, because ultimately each row comes from an average from that respective row and the sizes are equal. But, you can think of it either way, either way this is the average Y when A is set to minus 1 and that is 81.8.

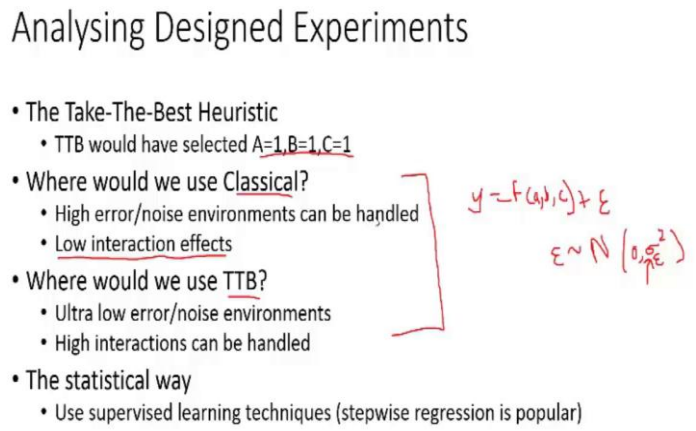
And similarly you get an average for 0, you get an average for plus 1 and you would basically say, I like A at minus 1 it is giving me the best result. Similarly you do it for B at minus 1 and B at plus 1. Now, when you do it for B at minus 1 and B at plus 1, how do you take the average? It is the same principle. So, you would for instance at B at minus 1 you would be interested in taking the average of these points, B is minus 1 at these points, so it would ultimately be the average of these points.

So, essentially that average would be B at minus 1 and that is 77.83. So, you do the same process and it is clear that A at minus 1 is the best, B at minus 1 is good, because B at minus 1 is greater than B at plus 1 and C at plus 1 is good, because that is better than C at

minus 1, so that is essentially classical analysis. And it is a fairly heuristic approach and it is like a first cart approach. Now that is one way of going about the analysis.

Another approach is to take the best. The take the best essentially says, I am going to look at that treatment combination, which gave me the average highest average Y. So, would that be at here? So, it looks like 85.75 is the highest average Y and that setting is I believe A 1, 1, 1 and, so we would essentially go with that recommendation.

(Refer Slide Time: 07:59)



So, take the best would have selected A is equal to 1, B is equal to 1and C is equal to plus 1. So, here are two fairly contrasting approaches and the question you need to ask yourself is, where would you want to use classical analysis and where would you want to use take the best. And to answer that question we need to answer that question in terms of, if you want to take a one factor at a time approach, what are two reasons that experiments that particular approach fails or for that matter, what is the two major difficulties with design of experiments.

The two major difficulties are the following. One is that, you could have some interactive effects between the input variables, which means the effect that n input variable will have on the output variable really depends on how some other input variable is set. So, that is called an interactive effect. The other reason is that sometimes you get fooled, because yes, why is some function of A, B and C from top of that, there is also some noise.

So, that noise would have given you results, which you are erroneously interpreting and you are essentially over fitting and you are getting fooled. Now, the question is, under what circumstances would classical analysis work and under what circumstance would take the best work and the quick answer is, a classical analysis essentially works really well in an environment of high error or noise. So, when this, when the noise, which is so we said Y is equal to F of a, b and c, but it is also got this noise component and when this noise component, although this there could be no bias to this noise.

So, this noise could be something like it is normally distributed with mean 0 and standard deviation, some standard deviation. Then, if sigma e is very high and, so it is a very noisy environment, then classical analysis would work quite well because it is averaging a lot of data points. Now, where it move to work well is when there is lots of interactions, so you need low interaction effects for classical analysis to work, because it just taking the average at a and average at b.

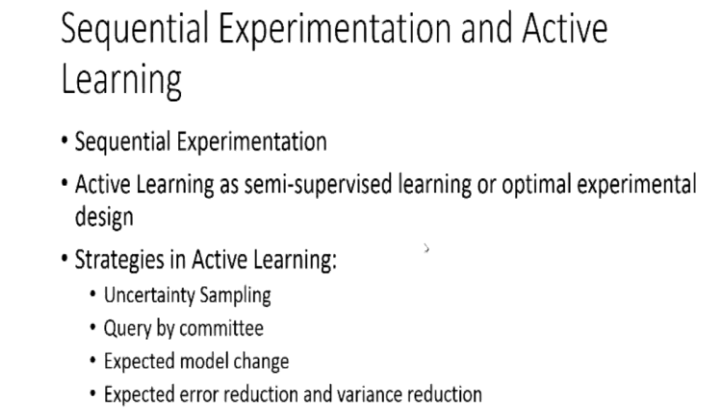
In contrast, take the best would work only when there is very low error or noise. It would not work well when sigma e square is high, but because it just takes the best combination, it is almost like it does not care about the interaction. So, high interactions work very well in it is favor and, so you would you take the best there. Now, the reason we talked about these two heuristics is to really motivate the statistical way of doing things, which is this dichotomy that you see between high noise versus low noise and fitting to any shape you want versus not being able to fit any shape you want has a lot to do with something you seen before, which is the bias variance dichotomy.

And this bias variance dichotomy is what you seeing in these two extreme approaches. The statistical way can sometimes balance that out and, so the truth is almost any supervised learning algorithm could be fair game, the only context is that the data set is fairly small. But, any data set, any supervised learning algorithm that is not require a very large data set that requires a very small data set, but they still give you meaningful results could be a fair game.

And in that regard step wise regression is usually popular in analyzing designed experiments. You start with the basic model that Y is equal to linear function of all the inputs a, b and c and then, you also try to incorporate two way interactions in the form of saying a times b, a times c, b times c and you could even have a three way interaction a

times b times c. And, because your variables are coded to minus 1 and plus 1, just multiplying a and b as an input can have a meaningful interpretation.

(Refer Slide Time: 12:28)



Now, one thing that can be said about this process of design of experiments is the way we have done it with full factorials and the way we explained it makes it essentially a one short approach. It basically means you decided even before you see any results the entire set of points that you want you look at the input space. So, if you look at this you already decided, so each treatment combination is a point in the input space.

We already decided the all the points and how many times you want to look query each of these points even before you look at even one result. So, it ideal in some sense for a parallel deployment, but if you could do the sequentially is there a better way of doing it, which is can you react to the data your seeing to say in want to query this point more, because I am less certain about this point or variable verses saying o I am very confident about something else.

So, I do not want to waste my resources in conducting experiments in a particular place and that is primarily the motivation of sequential experimentation this idea that you can you can conduct experiment sequentially and that gives you an opportunity to react to the data. Now, in many ways while sequential experimentation has been an effort from the statistic community active learning essentially is this same core idea and it is been

motivated more often the machine learning community and the context in which, the problems are applied to do sometimes differ as the result.

Active learning is often seen as a semi supervised learning approach and understandably it is also called is optimal experimental design. Because, active learning is a process where the system sequentially chooses to query the design space and therefore, be able to build knowledge, on what we see and how it can be better understood.

The key different especially in a in the typical context of application except while design of experiments often focuses a lot on the sole idea of starting with 0 data often active learning will start with this notion that there is some amount of data and it is not enough and you might a want to sequentially query the system to improve upon your understanding of the system itself.

And, so in many ways it is kind of seen as semi supervised learning, because there is an abundance of unsupervised learning data just means there is an abundance of states of the input space and you sequentially get you choose points in the input state for, which you really want to get answers. And therefore, get outputs for and by doing that the whole idea is that you can get away with much less data on the output space and you can still come up with predictions that are meaningful.

So, what are some prominence strategies in the whole active learning frame work is that in a sense all of these strategies at we going to talk about now, rely on one thing. It realize on you know evaluating the, how informative different points on the input space can be if you query them and you got an answer, what you mean by queried is you choose a particular point in the input space and say can you please give me an output for that.

And that, now goes into your data set, where you can apply some kind of supervised learning approach to make sense of, what you have. And the strategies there are involved with active learning can be broadly classified in these four and you know this these is an area of you know where there is a active research and, so there they might be some strategies that also follow fall that are mu that might fall out it these have been historically the more popular once.

So, let us take look at the first one which, is uncertainty sampling, now this is perhaps some more simplest and also therefore, very fairly common frame work. And in this frame work essentially the active learner chooses to query instances, which it is least certain about, how to classify or how to predict. So, if it is a classification problem it says I am going to ask questions about I am going to choose points on which, I want answers on which, I want you to give me an output and I am going to choose points, where I am least certain currently given the information I already have I am least certain about if it is a classification problem 0 and 1.

I am least certain about whether to classify it as 0 or 1 in those instances I really want you to I want to conduct my experiment in that point, where I am least certain. Now, if you can also think of this as regression problem, where at a certain point in the input space you might have a prediction, but that prediction is not cast and stone some kind confidence bound. So, some kind of bounds around that predication is some amount of uncertainty associated with particular predicted value and you might choose to pick the point where your uncertainty is the highest.

Another approach also fairly popular one is this idea of query by committee, let me just mark, where we are we finished this and the second is query by committee. Query by committee approach essentially involves having committee of different models, which are all trained on the current data set and when we say data here we are talking about the data set for which, we have the outputs you have a data set with the outputs in those. So, you have a some kind of supervised learner, which is capable of interpreting the data, but you might have a committee of a models, which are all trained on the current data set, but they might represent competing hypothesis.

Now, each of these members is now, are allowed to vote and you basically choose to take you choose to get a data point from the input space, where the committee members have the most disagreement. So, think of it you know one way to one example that I always kind about liked about this is it really maps on to lot of ensemble methods something that we saw earlier in the score. So, if you had a set of methods set of supervised learning approaches all trying to predict the same thing.

So, think of it as a random forest, where we have multiple trees trying to predict same thing. So, for a given input vector each tree does not going to come up with the exact

same class predication or it might not come up with exact same a predication even if it is the output is continues variable. So, this method simply says let us go with let us go get an answer for a data point, where the trees disagree most about what the classify it or what where were these highest variance in the predicated value of the models.

The next approach is expected model change the idea behind expected model change is to see if we knew the label of a particular data point. Then, which label of the input space if we knew would contribute to the greatest change in the current model we have based on the current data we have of the inputs and outputs. So, we have some data on the inputs and outputs and you basically extrapolate to this to the broader question of asking question saying you could get data of you could get an output for any point in the input space, which point would essentially lead to you making the largest change in the model.

And the idea is to kind of query that that point very specifically the last two sets, which is the expected error reduction and various reduction use the following approaches the approach is to basically say with error reduction the idea is they some deviation of the predication verses actual. So, you can think of it as essentially the residuals in a regression case and you know in other in every other case other it is bias or variances for whatever reason you are unable to predict you are unable to predict the exact value at a particular location the question.

We need to ask our self is and answer to which, point in the input design space could lead to the largest expected error reduction. In that sense its it is fairly close to a uncertainty sampling, but it is not just uncertainty sampling is the where it really differs from uncertainty sampling is uncertainly sampling ask this question about each data point in the input design space with respect to that data points.

So, I go to one data in the design space and ask the question saying, how much uncertainty there in that point. In terms my predication at that point, where as expected error reduction does not talk about a single point it talks about, how if I got and answer to a question at a particular points. So, I go to a particular point to the input design space and I query the oracle and I get an output.

Now, that output if I now, refit my supervised learning with this extra data point there is going to be overall reduction in my residuals across the boat, because this new data point could change the entire regression line fit. Now, this new regression line fit will create

new residuals and, so I am looking at the overall reduction in error between the data points and the fitted model. And, so I choose to query that data points in the input space and get an output for that data point, which lead to overall reduction in error.

Now, the variance reduction approach is a deviation from that it is a deviation in that it is you do not look at the error between the fitted model and the data points, which you just look to see reduce the overall variance in the output space. Now, that is done probably because it is much easier to do this, but what the core idea here is that you have some variance associated with the outputs. And you can still continues to reduce generalization error indirectly by minimizing output variance and that also can sometimes you do that because it is mathematically a easier you can kind of get close forms solution an and that is essentially the core idea.

So, I hope this gives you some feel for experimentation and the whole idea of active learning and at least some strategies that we could use an experimentation active of learning.

Thank you.