Data Science for Engineers

Prof. Raghunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 30
Solving Data Analysis Problems – A Guided Thought Process

In the previous lecture, we looked at data analytic problem types, the types of data analytics problems that one typically solves and we also looked at the various techniques that have a being used. Though not a comprehensive list gives you a variety of techniques that people are looking at to use to solve data analysis problems and then we also asked questions as to why there are so many techniques and I described one way to think about the techniques and the problems that are being solved.

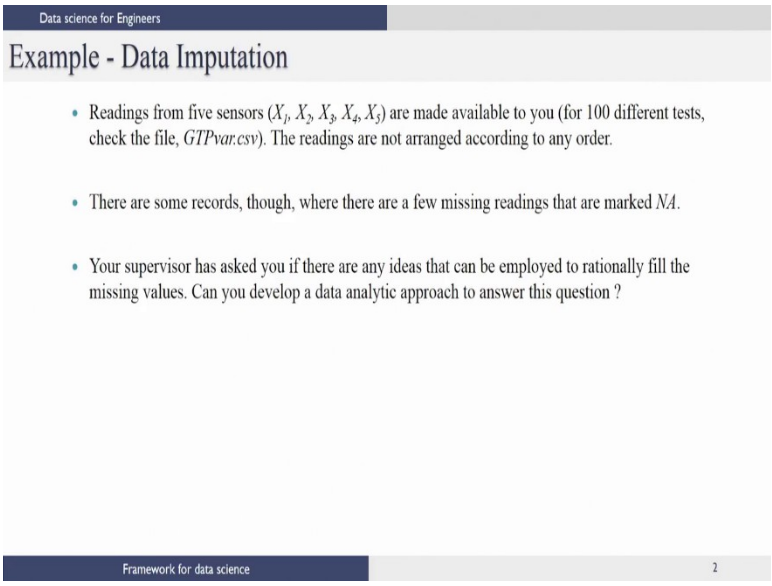
Now, in many cases as far as this data analysis problems are concerned typically you start with a very not well defined problem I would say. In a typical industrial scenario now-a-days there is a feeling that there is lots of data that is around and everyone seems to suggest that you should be able to use this kind of big data to derive some value to your organization. So, the question then is how do I do it, so typically people start by saying lots of data what is it I can do with this data. You know you might simply say I want improve performance are I want to minimize maintenance problems and so on.

So, you could start talking about a class of problems which could be either performance related or improving the operations, doing things on time and so on. So, typically you start with a loose set of words a vague definition of a problem and the data that you have. Now the question really is to then drive your thought process towards something that is codable, something that you can process the data with to derive value to do any problem that you are solving and so on.

While this is to some extent currently a little bit of unstructured process, good data scientists are able to come up with a solution ow that makes sense and that is relevant for the problem that needs to be solved. So, what we are going to attempt before we show you the techniques is to give an example of such a thought process so, that if possible you could use this kind of thought process for problems that you need to solve.

I just want to mention that this is not set in stone. We are trying to give structure to your thought process when you solve data analysis problems. So, to do this we are actually going to take a very simple example and then illustrate how you should think about solving data science problems and at the end of it we will come up with a flow chart that might be useful. So, the problem that we are going to deal with is what is called the data imputation problem.

(Refer Slide Time: 04:05)



Data science for Engineers

Example - Data Imputation

- Readings from five sensors (X_1, X_2, X_3, X_4, X_5) are made available to you (for 100 different tests, check the file, *GTPvar.csv*). The readings are not arranged according to any order.
- There are some records, though, where there are a few missing readings that are marked *NA*.
- Your supervisor has asked you if there are any ideas that can be employed to rationally fill the missing values. Can you develop a data analytic approach to answer this question ?

Framework for data science

2

.

This is a very standard problem that we see in many engineering applications and not only engineering applications in other fields also. The problem is the following. Remember we talked about the matrix as having values for different attributes at different samples. Now, in many cases I might have samples where I have values for all the attributes and there might be samples there I might not have values for some the attributes and so on. This is a very typical scenario when you look at large data that one deals with in real life.

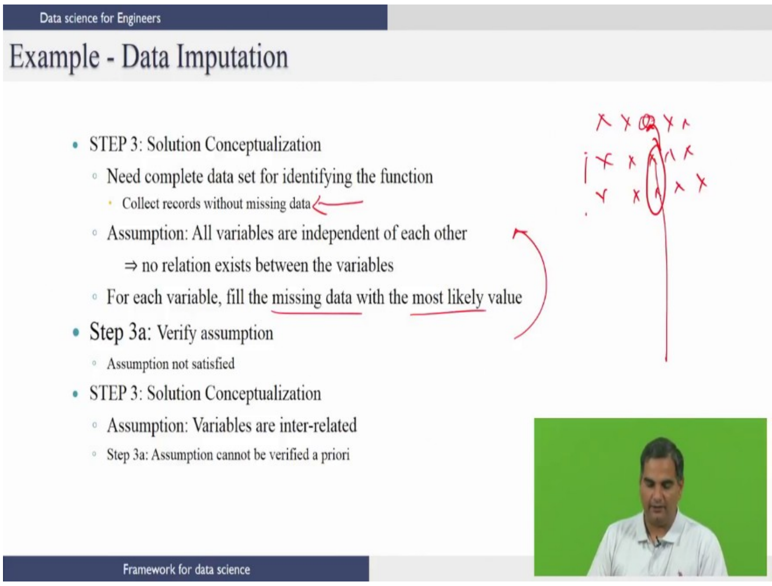
Now, the question is when I have data that is missing in some samples some attribute values that are missing in some samples, is there some way to fill in that data. So that I make the whole dataset complete and ready for further analysis. So this problem is called the data imputation problem.

So, here we pose an example data imputation problem and then we use this problem to kind of explain what this flowchart or the guided thought processes. What we are going to do is since we have not taught any machine learning techniques as yet we are going to use techniques that we have learnt in linear algebra and statistics to illustrate how we come up with this solution and for a more complicated problem, you would also bring in the other machine learning techniques that you would learn in this course and then use that in this guided thought process also to solve problems that are of interest.

So, let me read the problem. So, it says reading from five sensors, x one takes five, are made available to you for 100 different tests and in the website we also have this file of data which you can use to go through this process on your own the readings are not arranged according to any order. So, basically what this means is you are not in for any time sequence or any other sequence from this data samples.

Now if all the data were available then you could process this data for other activities. However, in this case there are some records which has data that is missing these are marked as not available n a. Now let us assume your supervisor has asked you if you can come up with any ideas that can be employed to rationally fill the missing values. So, the question is can you develop a data analytic approach to answer this question. So, this is a very typical question that we deal with in real problems and there are many solutions depending on the situation that one looks at. Here we gives one example to illustrate the idea of the framework.

(Refer Slide Time: 07:32)



Data science for Engineers

Example - Data Imputation

- STEP 3: Solution Conceptualization
  - Need complete data set for identifying the function
    - Collect records without missing data
  - Assumption: All variables are independent of each other
    - ⇒ no relation exists between the variables
  - For each variable, fill the missing data with the most likely value
- Step 3a: Verify assumption
  - Assumption not satisfied
- STEP 3: Solution Conceptualization
  - Assumption: Variables are inter-related
  - Step 3a: Assumption cannot be verified a priori

Framework for data science

So, I am going to call the first step in any data analysis problem solving as defining the problem. You define the problem in as broad a way as possible. So that it is very very understandable to everyone and this part of problem definition and the second step which is what I want to call as problem characterization are very important steps and in fact, if these are done properly then the solution, uncovering the solution process, becomes easier.

So, in this case the problem definition is actually fill in missing data records very simple. I have some data which is missing I simply need to ll in these missing data records. Now we drill down a little more and we go on to step 2 which is what type of problem is this. Now in this simple case it will turn out to be one type of problem that I am going to talk about, but in more complicated cases it might be a combination of these basic problem types. So, here at least for this example the problem characterization goes like this. So, we see that given part of the information which is the data that is available fill the missing information.

So, one way to think about this is I need to get some knowledge about the missing information from somewhere and the only place that I can get this information is really from whatever I know about this data from whatever I have with me currently. So, I might say the idea is really to somehow relate the missing information or missing data to the known information or known data. So, that seems like a logical way to solve this problem.

The minute we get to this point right here then we understand that this is a function approximation problem where I am going to write this equation where I have x unknown and if somehow I can relate it to the data, known data or x that is known, then whenever I have something missing I could simply put it into this function and as solve for my variable. So, at the highest level it proceeds in a rather general fashion, but you will see you will have to add more and more ideas into solving these problems as we go along.

Now that we basically said it is a function approximation problem where I am going to relate the unknown data as a function of known data, we need to segregate the data record to completely known samples and samples that are unknown. Because what we are going to do is we are going to relate unknown samples to known samples. So, in some sense we have to get a functional form. So, we try and get complete datasets for identifying these function forms.

So, we might say in a solution conceptualization step, it is important to collect records of data which have no missing data. So, this is a complete set of information, so we could possibly derive something out of this complete set of data some information out of this data and then see whether we could use that information to fill in the missing data. At this point we start making assumptions about the data, remember in the previous lecture I said you could have a function approximation problem you could have many types of functional forms you can use and so on and if you think about statistics there are different types of distributions that you can assume the data follows. You could assume the data is completely deterministic, variables are completely stochastic, variables are some combination and so on.

So, there are many types of assumptions that you can make. So, since there is the first course on data analysis and this the first time we are introducing this framework we are going to keep things very simple. Let us take an assumption which is very commonly made in solving these types of problems, we might simply say these variables are let us say independent of each other; that means, what value a particular variable takes does not really affect the value of other variables.

So, that basically means that no relationship exists between these variables and if we make this assumption then the data filling activity could be for each variable you can fill the missing data with the most likely value. So, there are some fundamental assumptions that you are making when you say that you are going to fill the missing data with most likely value, but from a very simple conceptual layman terms this could just be the average of all the values for that particular variable.

So, if I have let us say let us say these are all known data and let us say this is missing data and I have something like this I have something like this and if I let us say want to just fill in this missing data we start and then get this set of data separately where nothing is missing and then basically say the best way to really fill this is to take an average of these two and put it here. So, somehow we are going to talk about this most likely value. So, we will see how we feel the most likely value. But basically what we are a essentially saying is when I want to fill this data all I am going to do is I am going to look at the values only in this column and I am not really going to look at values in the other columns because I have assumed that these variables are not related to each other.

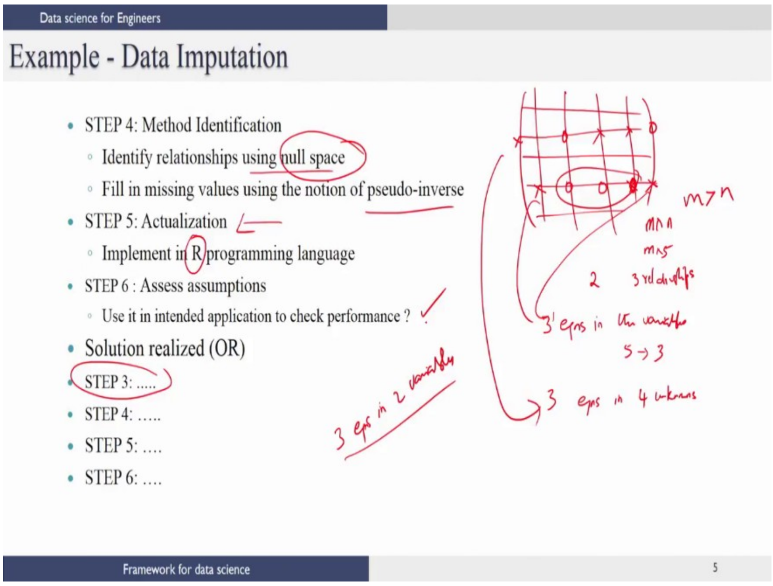
Now, in this case, this assumption can quite easily be verified right at this point from your statistics part of the lecture. You would have seen a quantity called correlation coefficient or quantity that measures the correlation between variables that you calculate. So, you could calculate and find out whether these variables are correlated or not and if they are not correlated then this assumption is fine, but if they are correlated then this assumption that no relation exists between the variables it is not a good one to make. So, I want to emphasize that some of the assumptions that that we make can be verified right at this stage based on known statistical ideas and other ideas in terms of linear algebra and so on.

So, if you could verify this assumption and then if there was no relationship, then you say the most likely value then you have to define what the most likely value means. There are many ways of actually defining the most likely value the simplest is what I said you could take the average or you could take the median value you could take a mode and so on. So, there are many ways of actually defining what the most likely value is.

So, let us assume in this case that this assumption is not satisfied then you go back to the drawing board and then say I have to change my assumption. So, this is where what I explained in the last lecture is important. If this assumption is not satisfied this does not mean there is any problem with the correlation calculation that you have done. That is a good calculation do to do anyway. The only problem is that the assumption that these are not related to each other is not a good assumption to make. So, we go and modify that assumption.

So, we could make the assumption that these variables are interrelated, now at this point just at this point I might not be able to verify this though in this case one could strictly make an argument that that you could verify it here itself, but in more complicated cases there are examples where the assumptions cannot be validated right at this point and you go back to this assumption only after you have used some machine learning technique to solve the problem and then come back and validate any case let us say we cannot validate this assumption a priori.

(Refer Slide Time: 17:28)



Data science for Engineers

Example - Data Imputation

- STEP 4: Method Identification
  - Identify relationships using null space
  - Fill in missing values using the notion of pseudo-inverse
- STEP 5: Actualization
  - Implement in R programming language
- STEP 6 : Assess assumptions
  - Use it in intended application to check performance ?
- Solution realized (OR)
- STEP 3: .....
- STEP 4: .....
- STEP 5: ....
- STEP 6: ....

Framework for data science

5

.

So, if we assume that there are relationships between these variables, then there is this question of actually coming up with a method that will answer our question as to what are the relationships among these variables. So, this is what we call as method identification. So, in this case remember if I have a matrix of data this we have discussed several times let us say I have m samples in n variables.

So, we said we have samples in rows and variables in columns, we said if there are a lot more samples than variables which is the case here because we have 100 samples and out of those we have picked out samples that are complete in all respect. So, let us assume there are a certain number of samples which are complete we have only 5 variables. So, the number of records where information is complete is going to be lot more than the 5 variables.

So, basically m is greater than n and then if you want to use things that we have learned before to come back and say I want to solve this problem using things that I have learned, you would quickly realize that we can use the notion of rank and null space to solve this problem and why is it that we can use the notion of rank a null space to solve the problem. We said if you want to identify how many of these variables are actually independent the quantity that you should compute is the rank of the matrix which is what we saw in the linear algebra class.

So, if it turns out that if the rank of this matrix m by n matrix, let us say whatever is the number of records that are complete times 5, because we have 5 variables in this problem, if the rank of this matrix turns out to be 2, then we automatically know that there are 3 relationships. And we also know how to identify these relationships. We can use the notion of null space to identify these relationships.

So, let us assume that we have identified these relationships. That means I basically have A_3 equations in the 5 variables that are there in this problem. Now we are ready to solve the actual problem. For example, if there is a record where there are let us say 3 variables that are missing. So, let us take an exam-ple here. So, let us say I have this I do not have this, I do not have this, I have this, I have this, sorry 3 variables are missing, so, I do not have this and I have this and this.

Then if you want to fill this data what you do is basically take these two known values and substitute these two values into the 3 equations. So, those are now known so these 3 equations will go from 5 unknowns to 3 unknowns. So, you have now 3 equations in 3 variables then you can basically solve this problem and fill in this data. If for example, there are 4 variables that are missing in a record then that is one case that we have discussed already in the linear algebra framework. We have let us say 4 variables missing then the 1 variable that we know the value for, we can substitute into these 3 equations and then we will end up with 3 equations in 4 unknowns.

So, this is a case where I have a lot more variables than equations. So, there are infinite number of solutions. So, one possible solution that you could use is to use the pseudo inverse and then find a solution to this problem. Similarly if I have for example, 2 variables that are missing, but 3 are available then when I put these 3 values into the 3 equations I will end up with the system of equations where I have 3 equations in 2 variables.

Now if the equations are all perfect, you could pick any 2 equations from this and then simply solve for the 2 variables. But even if there are minor errors in these equations and the equations are not really perfect in terms of just drop-ping one equation and solving with other 2 equations, you could still use the notion of pseudo inverse again to solve this problem where I have less variables and more equations.

So, this is a case where we said if all the equations are not consistent you might not be able to find solutions, but we know pseudo inverse is a concept that can be used to solve all types of these cases where you have the same number of equations and variables more equations than variables and less equations and variables and so on. So, this we saw in detail in the linear algebra part of the lecture. So, you notice how a general statement of a problem, fill in data, can be eshed out in a very systematic way and then you can use concepts that you know, right now since you know from this course at least only concepts from linear algebra and statistics.

We have used only those concepts to illustrate solution to this problem. Now this is a generic approach and once you learn more and more machine learning techniques you will be able to fill in more sophisticated techniques for things like this and then say I am going to use this technique because the assumptions that I have made are consistent with that particular technique and so on.

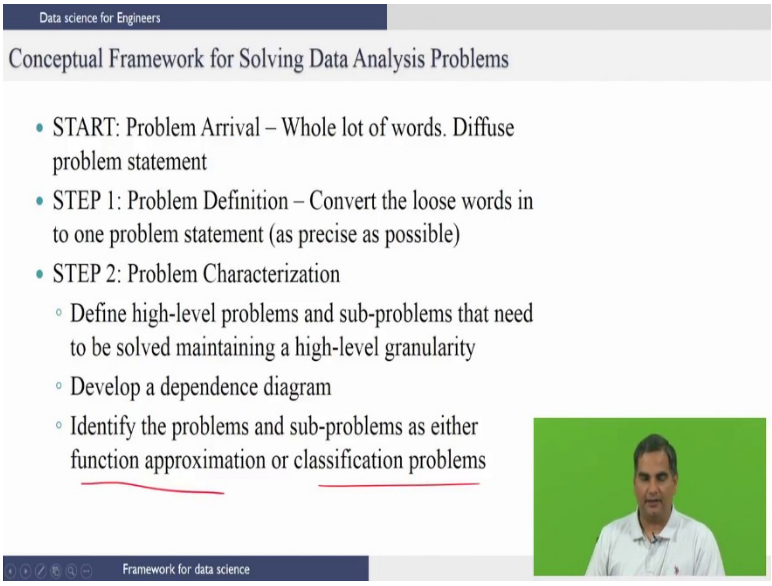
Now this is conceptually saying pseudo inverse and so on. In an actual data analytics problem solution, you have to actually what we call as actualize this solution which is basically implemented in some programming language of choice. You could do it in math lab, you could do it in scy lab, you could do it in r, you could do it in python and so on. So, in this case for this problem you might write an r code and then once you are done with this then you go back and assess the assumption.

So, maybe you could actually take the completely filled in data with your data imputation and use the data set for whatever the intended application is and then look at whether you are getting a performance that you are happy with. Now if you are not getting a performance that you are happy with then you basically do not blame the null space concept, but you say maybe the problem is that these relationships are not linear or if you assume that there is no error or noise in the data maybe that assumption is not valid.

So, we have to go back and look at those assumptions and then maybe you could say it is still a deterministic problem, but there are non-linear relationships. Then you have to figure out how to get those non-linear relationships or you could say there is a lot of noise. So, I might use some other idea to fill in the missing data and so on. So, you could say if the noise you could you could attribute a particular distribution for that and depending on the distribution you could use the correct technique and so on.

So, if the solution is realized at this point well and good. If not you go back again to making assumptions. So, the step 3 is where we keep going back where we keep refining our assumptions and what our assumptions can be verified right away; we verify whatever assumptions, we have to wait till the final result to verify, we verify, and then if things work out we are happy otherwise we keep this assumption validation cycle till we solve the problem to our satisfaction.

(Refer Slide Time: 27:02)


Data science for Engineers

Conceptual Framework for Solving Data Analysis Problems

- START: Problem Arrival – Whole lot of words. Diffuse problem statement
- STEP 1: Problem Definition – Convert the loose words in to one problem statement (as precise as possible)
- STEP 2: Problem Characterization
  - Define high-level problems and sub-problems that need to be solved maintaining a high-level granularity
  - Develop a dependence diagram
  - Identify the problems and sub-problems as either function approximation or classification problems

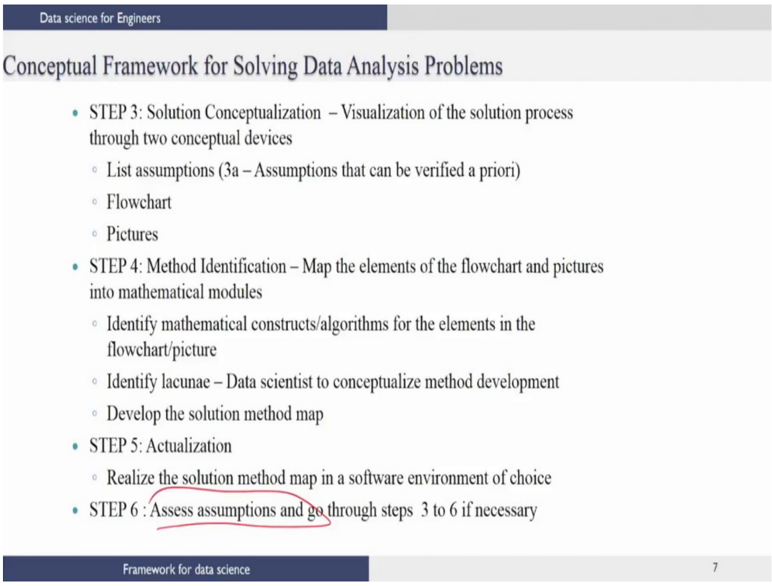
Framework for data science

.

So, in summary I would say the start of all of this is the first step is a problem arrival whole lot of words very diffuse problem statement. Step one is to convert this into one problem statement or set of problem statements as precise as possible and then to solve that problem you do what I would call as problem characterization. So, you break down this high level problem statement into sub problems and you kind of draw a ow process saying if I solve this sub problem then this result I am going to use in this sub problem and so on.

So, you can think of this like a flowchart that you are drawing with these sub problems and in general if possible you get to a granularity level where you are able to identify the class of problem that the sub problems belong to. In this case of this course we are calling these as function approximation or classification problem so you identify these problem sorry as either function approximation are classification problems.

(Refer Slide Time: 28:14)



Data science for Engineers

Conceptual Framework for Solving Data Analysis Problems

- STEP 3: Solution Conceptualization – Visualization of the solution process through two conceptual devices
  - List assumptions (3a – Assumptions that can be verified a priori)
  - Flowchart
  - Pictures
- STEP 4: Method Identification – Map the elements of the flowchart and pictures into mathematical modules
  - Identify mathematical constructs/algorithms for the elements in the flowchart/picture
  - Identify lacunae – Data scientist to conceptualize method development
  - Develop the solution method map
- STEP 5: Actualization
  - Realize the solution method map in a software environment of choice
- STEP 6 : Assess assumptions and go through steps 3 to 6 if necessary

Framework for data science

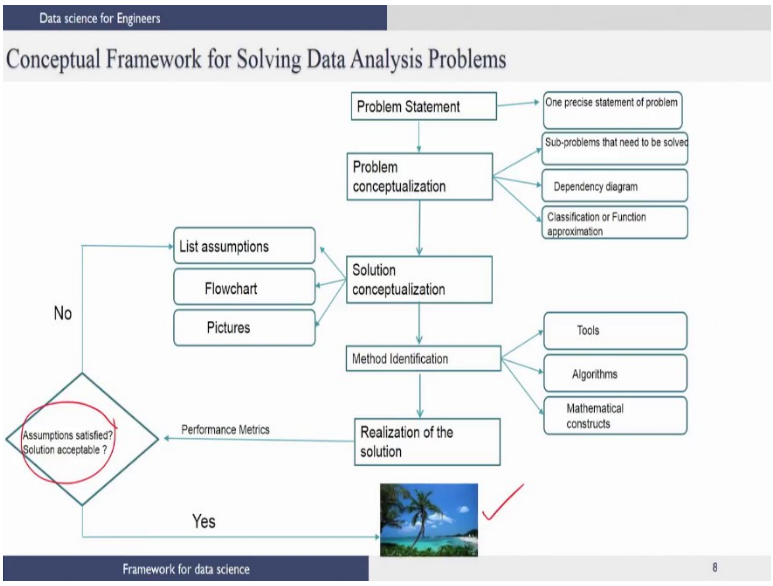
7

So, this is where you then look at solution conceptualization again you have to make assumptions here. So, you could make assumptions about distributions about linearity and non-linearity the type of non-linearity and so on and here if you if you kind of draw a flowchart and have some pictures in your head then it becomes easier to solve this problem. Then once you conceptualize the solution, then for each of these sub models or sub modules you have to identify a method and the identification of the method should be dictated by the assumptions that you have made.

So, just for classification there are so many techniques which one do you choose. So, we already addressed this in the last lecture you have to look at the assumptions and pick the right method for solution and if it turns out that for the kinds of assumptions that you have made that you do not like any method that is out there then you tweak the existing algorithms to a little bit and then find a method that is useful or that will work for your problem and then once you do this then you basically actualize the solution in some software environment of choice and you basically then get the solution and assess whether the assumptions are good, whether the solution satisfies your requirements and if it does you are done. If it does not you go back and relook at your assumptions and then see how you change or modify your assumptions so that you get a solution that you are comfortable with.

Now, again this step of assessing in the assumption which is basically through the solution that you get is a critical step. Here what we typically do is we partition the data in most cases to data that we use while we are going through this whole process and then data that has never been used when we were developing the solution and you basically test your algorithms or the flow process that they have come up with on data that has not been seen. So, that is called the test data and this is an important thing to remember and we will emphasize this more when we teach linear regression and classification. So, this is a critical component of assessing assumption. So, this we will see in more detail later.

(Refer Slide Time: 31:04)



Data science for Engineers

Conceptual Framework for Solving Data Analysis Problems

Problem Statement

One precise statement of problem

Problem conceptualization

Sub-problems that need to be solved

Dependency diagram

Classification or Function approximation

List assumptions

Flowchart

Pictures

Solution conceptualization

No

Method Identification

Tools

Algorithms

Mathematical constructs

Assumptions satisfied? Solution acceptable ?

Performance Metrics

Realization of the solution

Yes

✓

Framework for data science

8

So, the whole thing that we have described till now I have as a flowchart here where we start with the problem statement problem conceptualization, solution conceptualization, method identification and realization of solution and finally, when all are assumptions you think are satisfied, the solution is acceptable then you are home clear if not you go back and then redo this till you get a solution that is of value in terms of the problem that you are solving

So, with this the gentle introduction to data science part of the lecture is done. So, we looked at the types of problems, the techniques and why so many techniques and so on and we also provided a framework to guide your thought process and as I mentioned before this is a process that you can use to think about many different problems in a framework in the same way. So that the solution development process becomes easier for you as you go along. You might tweak this framework and then you will have some mental picture or mental framework that you use to solve data science problems which could be this which could be a tweaked version of this or something different, nonetheless the important thing to remember is that you should think about the problem in a consistent way whenever you solve a problem.

So, what I mean by this is if you use whatever framework it is for a particular type of problem you become aware that you are using such a mental framework when you are solving a problem then you can take the same framework to many different problems then that becomes your thought process for solving data science problems. You do not keep looking at books and say I have to do this, this and this. You have your unique scientific method for thinking about these problems and solving these problems.

So, with this we finished this part of the course and the next set of lectures would be on linear regression which is a type of a technique or a technique for solving function approximation problems and then after that, we will have a series of lectures on some clustering algorithms which can be used largely for classification problems, but can also be used in solving function approximation problems and we will then close the course with a case study and one practical problem description thank you and I hope to see you again when linear regression is started.

Thank you.