# **Mathematica 11 Machine Learning**

### **Welcome**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Curt] Hi, I'm Curt Frye. Welcome to Mathematica 11 Machine Learning. In this course, I'll show you how to get started with Machine Learning in Mathematica. I'll start by showing you how to split your data into training and test sets, and to import those files into Mathematica. Next, I'll show you how to manage your data, specifically how to standardize your data to avoid bias measurements, interpolate data to identify missing values, and to group or sort your data in meaningful ways. Then, I'll show you how to perform linear regression, analyze time series data, and identify any known sequences or generating functions behind your data. Finally, I'll demonstrate how to separate data into classes, identify clusters, and assess your model's performance. I'm sure that you'll find that your time with Mathematica 11 Machine Learning will be time well spent. Dive right in.

### **What you should know**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Before I get started with the main content of Mathematica 11 Machine Learning, I wanted to give a quick overview of what you should know to get the most out of this course. First, it'll help if you have basic knowledge of Mathematica 11 or earlier. The program, as well as the Wolfram Language, has some idiosyncrasies that you should know about. Second, it would help, but it's not necessary, for you to have experience doing data analysis. Machine Learning is an extension of basic statistical analysis, so the more you know of that and how to analyze data, the better off you'll be. Third, one thing you should know is that when you open any of the notebooks, you should open the Evaluate menu and click Evaluate Notebook. That way, all of the data and variable assignments that I have created inside the script will be evaluated and actually assigned to that data. I mention that step specifically in every movie, but I wanted to mention it now because it is very important. Finally, you should have a willingness to experiment and explore. Mathematica 11 is an extremely powerful program and the new Machine Learning capabilities are also very powerful and flexible. So, I hope that you take everything you learn from this course and use it as a base for further exploration. With all that in place, let's go ahead with the Mathematica 11 Machine Learning content.

### **Exercise files**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] This course comes with a set of exercise files that you can use to follow along with the movies. I have downloaded the exercise files onto my computer and I have them in a folder on my desktop called Exercise Files. To find a file for a particular movie, open the Exercise Files folder and then double-click the chapter folder that you want, such as chapter two. From there, you can work with the files or just follow along.

### **Overview of machine learning tasks**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] As a discipline, machine learning covers a lot of ground. Mathematica 11 includes many new machine learning capabilities, and in this movie, I'll give you an overview of the tasks and the information they capture. Before you get started analyzing data, you need to prepare your data, and frankly, that's the best done outside of Mathematica. I'll show you how to prepare data in Excel, but if you have the ability to do it in R or another program or programming language, such as using the SQL database language, that's the best way to do it. There are two major types of machine learning, and those are supervised learning, where you tell the computer the answer, and see how well it does. And then, unsupervised learning, where you just let it go, and it can make its own conclusions, and then you take a look at the results later. An example of supervised learning is where you take, say, two-thirds of your data, and separate it out and create a model. But you already know the answers. Well, unsupervised learning is more keen to neural networks, and other techniques. To be more specific about supervised learning, and again, that's where you tell the machine the answer, so that it can calibrate its decision rules. You have linear regression and data interpolation. Linear regression is a technique of finding the best straight line that describes a dataset. That's with the minimum error. You can use that for future forecasts. A related technique is data interpolation, where you try to fill in missing values from a dataset. You can also find it generating functions and sequences. This is useful if you think there might be an underlying relationship in a dataset, or if you need to identify a data sequence. And you can also identify clusters of data. For example, you might have a set of customers who work at a certain industry, and like to stay in your hotel. If you're able to identify that cluster, then you can market more effectively. Machine learning is an extremely broad and extremely deep topic, but by introducing the techniques that I've described here, I hope to give you a background for your own investigations.

### **Separate training data from test data**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] One of the traps of machine learning is over-fitting, where your algorithm does too good a job of classifying your data. While it sounds like a good idea to get fit you can, you're only guaranteed a good fit on your present data. Future values might not follow exactly the same patterns, which would throw off your predictions. You can lower the risk of over-fitting by separating your data into training and test sets. In this movie I'll show you how to split your data randomly, so you get a good training set to work with. My sample file is the HousePriceData workbook, it's an Excel file that you can find in the chapter one folder of your exercise files collection. You might wonder why I'm working in Excel instead of Mathematica. Well, you can do this type of thing in Mathematica, but frankly, it's not the best tool. You have to do some fairly intricate work with commands. So I think you can do a better job in Excel or R or in a database program using SQL to split your tables. I'll show you how to do it in Excel, so you get the general pattern in mind. My data set consists of two columns and I have the number of SquareFeet of a house that was sold and the Price. And let's say I want to find a relationship between SquareFeet and Price and to do that I want to separate my data into a training set of about 70% of the rows and a test set with the remainder, so about 30%. Here's how you do that in Excel. First, you click to the right of the first row of data, not the header, but the actual data row. In this case I clicked in cell C2, then type an =, and type in the function rand, R-A-N-D, followed by an ( and ) parentheses, and then press Enter. This displays a random decimal number between zero and one in cell C2. Now I can fill this formula down by clicking on the cell C2, and you can see the value as a formula as apposed to an actual number. I can fill the formula down to the bottom of the column by moving my mouse pointer over the bottom right corner of cell C2. And I'll know I'm in the right place when my mouse pointer changes from a large white cross to a black crosshair. Now I double-click and Excel fills in the data. And if I scroll down you can see that it filled in the formula actually. So it goes down to cell C31. Now with the cell still selected I can replace the formula with the actual value. To do that I will copy the contents of the cells, so I'll press Control + C, and then I'm going to press a sequence of keys, you don't press these keys at the same time, you press one and then you release and then go to the next one. I'll press the Alt key, which you can see displays letters and numbers on the ribbon. I want to work on the Home tab, so I will press H, I want to work with the Paste button, so I will paste V, and I want to paste values, so I'll press V again. And when I do you see that the active cell, which is C2, contains a decimal that starts with 4100, the next one 4467, and so on. Now I can filter my data based on these random values. I want to create a training set of about 70% of these rows, so I will click in cell C2, and then go to the Data tab, and click the Filter button, that displays the Filter arrows. Then I will click the Filter arrow for Column C, which is my random values, point to Number Filters, and then click Greater Than. I want the value to be greater than 0.7, so I'll type that in the box, and click OK. And there I see my data. So this will actually be my test set, because I was picking about 30% of the values. I'll create a new worksheet for the values to go to, so I'll click the new sheet button, I'll double-click the sheet tab, and I'll change the name to TestSet, and Enter, then go back to Sheet1. And I'll copy the visible rows of just the square footage and price data, not the headers, then I'll press Control + C to copy the visible rows, go to the TestSet worksheet, and press Control + V, and there's my data. Now I'll create my training worksheet, so I'll click the new sheet button again, and double-click the sheet tab. Going to call it TrainingSet, and Enter. And I'll go back to Sheet1 and I'll Clear the Filter, but the Filter arrows still remain. I'll click the down arrow for Column C, point to Number Filter, and then click Less Than Or Equal To. And again, I want to filter to display only those values that are less than or equal to 0.7, so I'll type that, and click OK, and there is the other part of the data set. I'll select the visible values, but again, only the square footage and the price, press Control + C, go to TrainingSet, make sure I'm in A1, and press Control + V to paste. The most versatile file format for exporting and importing data among these types of programs is the comma separated value file. So I'll show you how to export the TrainingSet and TestSet worksheets to files of that type. I'm on the TrainingSet worksheet, so I'll go down and right-click the sheet tab, and from the shortcut menu that appears I will click Move or Copy. Now I can move my selected sheets. I'll click the To book down arrow and click New Book. I'll Create a copy, and then click OK. Now I have a new workbook called Book1, so I'll press F12 to open the Save As dialog box and I'll save it in the Exercise Files folder for Chapter01. Now I need to do two things, I need to change the type of the workbook, and I always do this first before I change the name, so I don't forget and hit Save. So I'll click the Save as type control and I'll click CSV, Comma delimited, and then instead of Book1 I'm working with my TrainingSet, so I'll type TrainingSet, and click Save. Now I'm done with this file, so I can click the close button. Now I'm back in my HousePriceData workbook, so I will right-click TestSet, Move or Copy, and I'll go a little faster this time. I want to go to a New Book, I want to Create a copy, I'll click OK. Now I'm in the new workbook called Book2, I'll press F12 to open the Save As dialog. I'm in Chapter01, first thing I'll do is change the file type to a CSV, and then I'll call it TestSet, and click Save. Now I can close my file and I have my data exported, ready to go, and import into Mathematica.

### **Import data from a file**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Narrator] After you've separated your data into training and test sets, you can import the CSV files into Mathematica. In this movie I'll show you how to import your data from those comma separated value files. That's the most common format that's used for data export and import. I have opened a blank Mathematica Notebook and I will bring in my data using a series of commands. I'll start by typing in the name of the variable to which I want to assign the list of data that I'm bringing in; I'll call it alist = I want that variable to hold the contents of a CSV file that I'm importing. So I will use the import keyword and then I need to provide arguments within a pair of square brackets. I need to give Mathematica the specific address of the file and in this case that means that I need to tell it how to get to it in my file system. And there's a trick that you can do in Windows that I will show you now. I already have File Explorer open in another window so I'll press Alt Tab to move to that. And here I have my data. I want to bring in my training set. So I'll click "TrainingSet." And that's Trainingset.csv. And you see here on the address bar it shows the hierarchy of my file storage system. What I need, though, is the actual path. So if I click this folder here: I get C colon users Curt Desktop Exercise Files Chapter One and it's highlighted. So I'll press Control C to copy the address. I'll tab to move back to Mathematica. This is a string so I will type double quotes and press Control V to paste in a text string. I don't want to use escape characters so I will click " No" and there it is. Now you notice that the file name is missing. So I'll type a back slash and then the name of the file and you can see here that I can just click. So I have Trainingset.csv. And I can see that Mathematica did update my path...good. That's working correctly. Then I'll type a second double quote to close and a right square bracket to close the argument list for import and then I'll press Shift Enter. And when I do, I see my data has been brought in; all the pairs are ready and I can analyze them at my leisure.

## **Question 1 of 3**

Creating a model that describes a specific data set very well at the expense of future performance is an example of \_\_\_\_\_.

* regression
* standardization
* overfitting  
  Correct
* normalization

## **Question 2 of 3**

You should split training data from test data using a \_\_\_\_\_ factor.

* congruent
* time-based
* random  
  Correct
* linear

## **Question 3 of 3**

A commonly used file format for data import is the \_\_\_\_\_ format.

* standardized
* rescaled
* CSV  
  Correct
* binary

### **Standardize (normalize) or rescale data**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Machine learning data sets will occasionally contain a column of values that are much larger or much smaller than the others. The classic example is house prices, where the value of the house is much larger than the number of square feet, bedrooms or bathrooms. In this movie, I will show you how to standardize, also called normalize, your data so one larger or smaller set of values won't throw off your result. My sample file is the standardize notebook, and you can find it in the Chapter 2 folder of your exercise files collection. I just opened this notebook, so I need to evaluate it to assign the values to my variable A list. So I will open the evaluation menu, and click evaluate notebook. Great. Now I can perform some calculations on the data as I have it. So let's say I have to find the mean, or the arithmetic average, of the variables, or the values, in A list. I can do that by typing mean followed by left square bracket and the variable name A list followed by right square bracket and Shift Enter. And I get an average of 2051. I can also check for the variance, which is the sum of the squared errors. In other words, however much anyone of the individual values varies from the mean, 2051, we square that value to get the variance. And I can calculate that using the variance keyword. Left square bracket and A list again, right square bracket and Shift Enter. I get the expression 1,422,494 divided by 9. If I wanted to see that as a numerical value, I could type N followed by left square bracket and then variance A list right square bracket to close out variance right square bracket to close out N Shift Enter and there's the value. If I want to standardize the list, I am dividing the list so that it has a mean of zero and a standard deviation of one. To do that I will type standardize actually I need to assign this to a variable. So I will backspace over standardize and I will call this new list equals and then the standardize keyword, followed by left square bracket and then A list, which is my original data, followed by a right square bracket and Shift Enter. You'll note that the original values are being divided by the square of the total. And any value tha is less than the average, 2051, is negative. And any value that's greater is positive. Let me show you what new list looks like as a number series. So I'll type N, new list, and I have new list in square brackets. Shift Enter and there are the values. If I want to see the mean of new list I can use the mean keyword again, and new list again, in square brackets, Shift Enter, and I get a mean of zero. And the standard deviation, which is the square root of the variance, we can calculate using standard deviation then left square bracket and new list right square bracket, Shift Enter, and we get a standard deviation of one. So if you ever have a data set where you're concerned that perhaps one of the columns might be throwing off your results, because of the magnitude of the values in that column, you can use standardized to bring the results more in line.

### **Replace values near zero with zero**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] When you work with large data collections, every so often you will end up with a value that is extremely close to zero. For example, the variable B list contains two data points that are very close to zero. The first is eight times ten to the minus seventh and the second is seven times ten to the minus 45th. So we have seven zeros to the right of the decimal point followed by an eight and then 44 zeros to the right of the decimal point followed by a seven. Both of those values are pretty close to zero and in some cases I will want to just round them to zero to make my calculations easier. Before I start working with the data, I need to evaluate the notebooks so that I actually assign my lists data to the B list variable so I'll go the evaluation menu and click evaluate notebook. There we go. And now I can assign my data to a variable so I will call it B list 2 equal and I'll use the threshold keyword. T-H-R-E-S-H-O-L-D I always try to add a second h before the o but it's not correct. It's spelled the way you see it there and I will type a left square bracket and use B list which is my original list name right square bracket to close and shift enter. There are my values. So you see that now I have 14.8 which is correct, same as before. Eight times ten to the minus seventh which is the same as before. Minus 9.2 and 0. So that means that the threshold keyword rounded one of the values to zero but not the other. The reason that threshold rounded one of those values but not the other is that eight times ten to the minus seventh isn't below the threshold where it gets rounded to zero. In this case, the threshold if you don't set any options is one times ten to the minus tenth. So nine zeros to the right of the decimal point followed by a one. If you want you can set your own threshold and also how Mathematica should evaluate it so if I reassign to the same variable, B list 2 equals threshold followed by a left square bracket, then the name of the list that I'm working with and that is B list which I didn't change before then a comma and now I need to add a list of arguments. All lists start with a left curly bracket and then in double quotes I'll type the keyword hard H-A-R-D and double quote again then the comma. Hard indicates that Mathematica should use this as a strict cut-off. In other words, if it's equal to ten or one times ten to the minus tenth then you should round to zero. It is a hard limit. And let's set the value to ten to the minus fifth. So I'll do ten and then the caret which indicates exponentiation, then minus five. I'll close my lists with a right curly bracket and close the arguments for threshold with a right square bracket. Now when I press shift enter to evaluate I get 14.80 minus 9.2 and 0. Contemporary computers are powerful enough that most times you don't have to worry about the extra calculation overhead that comes with values that are very close to zero. However, if you're working with a huge data set then you might find that rounding the values to zero if they're close to zero anyway will improve your performance.

### **Interpolate data to enter missing values**

Selecting transcript lines in this section will navigate to timestamp in the video

- Missing values can create problems when you analyze data. If you find that your data set has a few missing values, you can interpolate the data to estimate the values you want. In this movie, I will describe interpolation and how to do it in Mathematica 11. My sample file is the interpolate notebook, and you can find it in the chapter two folder of the exercise files collection. This notebook contains a list of values. 14, 19, 27, 31, 28, 35, and 40. And one thing I should note is that there's also an implied index. So, the first item, which is 14, is in position one. 19 is in position two, and so on. What I want to do, is to input a value, and use interpolation to return what that value might be. So, for example, I might get a decimal value of 3.5, and I want to see what the 3.5th element of the sequence would be. The first thing I need to do is to evaluate the notebook, so that I can assign my variable or values to the dlist variable, so I'll click the Evaluation menu header, and click Evaluate Notebook. There we go. Now, I can assign the interpolation keywords output to a function name. So, I will just call the function fn, which is what I usually use for function. Equal, and the keyword is interpolation. There we go, interpolation, followed by a left square bracket, and the name of my variable, which is dlist. Type right square bracket, and Shift, Enter to evaluate, and I get the interpolating function. Couple of things to note, and the first is that we have scalar output, which means that it is one to one. It's not a vector, so we don't have two components to it. And the second is the domain from the values one to seven. That means that interpolation will work for any value between one and seven. Interpolation doesn't work for eight, 8.5, that sort of thing. Now, I can use my function name, fn, and input a value to see what the result would be. So, if I want to see what the value of position 3.5 would be, I can type fn, left square bracket, and then 3.5, right square bracket, and Shift, Enter, and I get the value of 29.6875. That is the function's estimation of what the value of position 3.5 would be. You're probably wondering what would happen if I enter a value that is greater than seven. In other words, outside of the domain, so let's see what happens. I'll type fn, left square bracket, and then say, 9.2. And a right square bracket, and Shift, Enter. I get a note saying that the interpolating function lies outside the range, so it's using extrapolation, and I get a value of minus 15.176, which honestly doesn't make any sense. So, you can see that interpolation works very well for numbers within the domain of values you already have, but it doesn't work at all for numbers outside of it.

### **Count values by adherence or non-adherence to a rule**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Businesses expect to perform within limits. Whether those limits are statistical bounds in a Six Sigma framework or daily store earning targets, you want to know how consistent your processes are. In this movie I'll show you how to count values by adherence or non-adherence to a rule. My sample file is the CountValues notebook. And you can find it in the chapter two folder of the exercise files collection. Before I get started, I need to evaluate the notebook to assign the list of values that I have to my variable a list. So I will open the Evaluation menu and click Evaluate Notebook. There we go. If you want to count the number of occurrences based on a rule, you need to use the CountsBy keyword. So I'll go ahead and type that in. CountsBy, the C and the B are capitalized. Then a left square bracket. And now you need two bits of information. The first is the list of data and that is a list, which I have above, then a coma, and the rule. There are number of different rules that you can use. For example you can check to see if numbers are even. The keyword for that is Even and then a capital Q, followed by a right square bracket and Shift, Enter. And I see that there are two even numbers and four odd numbers. So I have 14 and 40 and the rest are odd. If you want to find the converse, that is the number of odd numbers, you can do CountsBy, left square bracket, a list, comas, OddQ, right square bracket and Shift, Enter. And you get the expected result, False, two, and True, four. And you can also check for prime numbers. So CountsBy, left square bracket, a list, coma, PrimeQ, right square bracket and Shift, Enter. And I see there are two prime numbers. Those would be 19 and 31. You can also check to determine whether values are divisible by another value. To do that you use the divisible keyword. So I'll type D-i-v-i-s-I-b-l-e, followed by a left square bracket, and the variable name is a list, coma. And I want to see if my values are divisible by five. So I'll type five, right square bracket, and Shift, Enter. And I get a series of Falses and Trues. If I want to see and output with the number of Falses and the number of Trues then I can use the Total keyword. So I'll type Total, left square bracket and then repeat my last command. So that's Divisible, left square bracket, a list, coma, five, right square bracket to close divisible, another right square bracket to close total, Shift, Enter. And I see that I have three False and three True. These examples demonstrate how you can determine which values within your data set meet a rule based on evenness, primality, oddness, or if they're divisible by another number.

### **Sort elements using a rule**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] In the previous movie, I showed you how to group elements by using a rule. In this movie, I'll show you how to sort your data using similar techniques. My sample file is the sort elements notebook. And you can find that in the chapter two folder of the exercise files collection. My variable b list contains a list of lists. I have triple values. 2500, zero, zero for the first one. Zero, 14, 23, 1908 for the second and so on. These values could represent sales within a store on a given day, arranged by department. Or they could represent any other type of value that you want to work with. And in this movie, I will show you different ways of working with those values. But before I start, I need to evaluate my notebook. That way I assign my list to the b list variable. I'll go to the Evaluation menu and click Evaluate Notebook. There we go. Let's say that I want to sort these values based on the contents of the triplets. To do that I will use the SortBy keyword. Follow that with a left square bracket. And now I need to identify the list of values and that is b list, followed by a coma. And now the rule that I want to use. And the three that I'll show are last, first, and total. To sort by the last value, I'll type the keyword Last. And that starts with a capital L. Then a right square bracket, and Shift, Enter. And I see that the last values go in ascending order. So I have zero, 11, 10, 1728 and 1908. I can do the same thing for the first values. Sort by, left square bracket b list, coma first. Then Shift, Enter. And I have zero, 302, six, 20, and 2500. I can also sort by the total. So rather than look at the value from department one or department three, I can find the total of all of them. So for that, I'll do a SortBy, left square bracket, b list, coma, total. Then a right square bracket and a Shift, Enter. And you can see the values there. When you import your business data into Mathematica, you can perform a lot of different sorting operations. The ones that I've shown here are very useful in a business and machine learning context.

## **Question 1 of 4**

The presence of comparatively large values in one field can hurt the performance of machine learning models.

* TRUE  
  Correct
* FALSE

## **Question 2 of 4**

You can use \_\_\_\_\_ to enter missing values without changing the overall pattern of the data.

* normalization
* regression
* standardization
* interpolation  
  Correct

## **Question 3 of 4**

Which of the following keywords is not a valid comparison for a CountsBy or GroupBy expression?

* IntegerQ  
  Correct
* EvenQ  
  Incorrect
* OddQ
* PrimeQ  
  Incorrect

## **Question 4 of 4**

You can sort items from a list of lists using the total of the values in each sub-list.

* TRUE  
  Correct
* FALSE

### **Find a fit using a linear model**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] When you analyze data, you will often have a set of inputs or independent variables that combine to produce an output or dependent variable. The latter values are called dependent variables because they depend on the inputs for their value. In this movie, I'll show you how to generate a linear model that predicts the results from a set of inputs with the least possible error. My sample file is the Linear.nb notebook. You can find it in the Chapter03 folder of the exercise files collection. The data that I'm going to assign to my clist variable are value pairs. The first value in each pair is the number of square feet for a house, and the first one it is 1,400, and the second is the sale price of the house... for the first pair that's 225,000. Before I start working with my data, I need to evaluate the notebook. So I'll go to the Evaluation menu and click Evaluate Notebook. Now my list has been assigned to the clist variable. I want to create a linear model and that will take the form of A times X plus B equals Y. So sum multiplier A will be multiplied by the first value, in this case 1,400, plus B, which is another constant value, will equal the output Y. The values won't be exact. That's because there is some variation in the relationship between square footage of a house and the sale price, but the linear value should give us a good indication or good prediction of what should occur based on a given square footage. I'll create my linear model. I'll assign it to a variable named lm, then equal sign. The function or keyword I'll use is LinearModelFit. There we go. Then a left square bracket. And I want to use my data variable, so that's clist, comma. And then x and x. So, what I'm doing is saying take a look at the input variable, which is the square footage of the house, and use it for the criteria. I'll type a right square bracket. Everything looks good. And Shift + Enter. And I get my model as an output. The model says that there is a base price of 4,776.17. In other words, if a house had zero square feet, then that's what you would expect to have the value of the land be, I guess. And then, the x value, which is the square footage of the house, is multiplied by 141.473. And when you add 4,776.17 to the value of the multiplication, then you get a predicted value for the house. So let's see how that works. I will type lm, which again is the variable that I assigned the model to. And then I'll give it a square footage to estimate. So I'll say 2,000. Right square bracket to close and Shift + Enter. So I have a predicted value of the house of 287,723. And based on the admittedly limited dataset that I have, that seems like a reasonable value. Now to check the model, let's put in an existing value and see what the model predicts. Let's go with a 1,900 square foot house. So I'll type in lm, left square bracket, 1,900. Right square bracket and Shift + Enter. And we get a value of 273,576, which is substantially higher than the actual value of 235,000. And there can be a lot of reasons for that error. The first and most likely is that we don't have enough data to make a really educated guess. And second, we also don't have any indication of the condition of the house or of the location. All we're looking at is a simple one-to-one relationship, square footage versus sale price. When you're working with that simple of a model, a substantial amount of error is not surprising.

### **Find a time series that fits given data**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Most business data has a time component to it. Whether it's the date of an order or quarterly earnings. In this movie I'll show you how to calculate a time series that best fits the data you have collected. My sample file is the time series notebook and you can find it in the chapter three folder of the exercise files collection. In this data I have a series of values, they are going to be assigned to the T list variable. So I have in position 117, in position 224, and so on. Each of those values represents a measurement at a particular time period. And for the purposes of creating the time series, it doesn't matter what the interval is. It could be every hour, every day, every week, it doesn't matter. Your calculations will be relevant to any time at that increment, hour, day, week, and so on. Before I start my analysis, I need to evaluate the notebook to actually assign the variables to T list. So I'll go to the evaluation menu and click evaluate notebook. There we go. Now I can define a time series model. I'll start by typing in my variable tm then an equal sign. The keyword I want to use is time series model fit, so time series model fit. Then a left square bracket. And I need to input my data and that is in the variable T list. Type a right square bracket to close out the argument list and shift enter and I get my time series model. Now if I want to get a forecast for a particular value from a time series I can do that. I'll start by showing you a forecast for a value that occurs within the time series. There are ten items on my list. So let's say that I wanted to find a prediction for value 6.2 To calculate that, I would type tm, left square bracket and then 6.2. Then a right square bracket and shift enter and I get the value of 40. Well if I count in the list, value 1, 2, 3, 4, 5, 6 and then anything greater than six would be assigned to position seven and that is 40 and that is the prediction that I get. But now let's say that I look for a value that is outside of my existing values that in other words we're extending the time series and making a prediction. I have ten values so let's say that I want to look for the value at position 12. To calculate that, I would type tm, left squre bracket, number 12, right square bracket, shift enter, and I get a point prediction of 127.126. So if the time series continues on as expected then that would be the value at position 12. I can also calculate what's called a confidence interval. Using statistical techniques, you can calculate a range of values so that you are 95 percent certain that the value in position 12 will fall in that range. For that I will type tm, left square bracket, and now I need to type in the keyword prediction limits in double quotes. So double quote, prediction limits, double quote, and this syntax is a little awkward but believe me it works. So I'll type a right square bracket to close out prediction limits then another left square bracket for the next argument and that is 12. That's the value that I want to predict. So I'll type a right square bracket to close out the second argument and when I press shift enter, I get the range and I see here that the range, again the 95 percent confidence interval, runs from 76.0911 to 178.162. That is a huge range and there are a couple of reasons for that. The first is that there are only ten values in this series. So if you are looking at those values then you want to be 95 percent certain, you have to have a wide range. And the second reason is that the future's uncertain so if you want to have a 95 percent level of confidence and be right 19 times out of 20, you have to have a wide range.

### **Find a formula that represents a data set**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Narrator] Businesses follow patterns. Whether they occur within a day, a month, or as the seasons go by. Scientific data also tends to follow patterns. In this movie, I will show you how to have Mathematica identify a formula that generates the values in a data set. My sample file is the Find Formula notebook, and you can find that in the chapter three folder of your exercise files collection. I have a specifically list of value pairs, and the idea is that I have the input to a function and then the output. So one generates seven, two gives you thirteen, three, twenty-three and so on. And again, I want to have Mathematica identify the formula that generates those values given that input. Before I start I need to evaluate the notebook, to assign my series to the e-list variable. So I'll click the evaluation menu, and click evaluate notebook. There we go. Now I can use the find formula keyword. Both effs are capitalized. So I'll type Find Formula, then a left square bracket, and my data which is an e-list, and a right square bracket. Mathematica understands that it's looking for pairs of value, an input and an output, and when I press shift enter, it thinks for a moment, and then identifies the formula that generates the values. In this case it did find a result, and if I look, it is five plus two times the input number, that's what the hashtag one means, or the number sign one means, raised to the second power. So two ex squared plus five. And if we look at the values, we have two times one squared, which is two, plus five is seven. So the first item is correct. For the second one, we have two times two squared, which is eight, plus five is thirteen, and so on. So we definitely have a formula that generates the values in the series. You might be wondering what this is good for, except maybe some logic problems, or identification problems you might find in a puzzle magazine. Well finding a formula for a data set is actually useful in a business context, if you are doing fraud detection. If you suspect the data that you're looking at might not be legitimate then one easy way to test it, is to use Find Formula to see if it was generated using some sort of a straight forward conversion from the input to the output. It's not likely that Mathematica will be able to find a formula, but if it does, it's an easy indication that you should look further.

### **Find a function that generates a given sequence of values**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Voiceover] In the previous movie, I showed you how to identify a formula that generates values in a data set. In this movie, I'll show you how to do the same thing but this time for known sequences of numbers. My sample file is the find sequence notebook and you can find it in the chapter three folder of the exercise files collection. I have three lists or sequences of data and you can read them here. Before I start my analysis, I need to evaluate the notebook to assign the list to the variables, so I will go to the evaluation menu and click evaluate notebook. There we go. And now I can try to identify a sequence function. To do that, I use the find sequence function keyword, so I'll type find sequence function followed by a left square bracket and then the name of the variable that contains the list I want to examine. I'll start with H list, which has the values one one two three five eight, 13, 21, 34, 55. It's probably familiar to some of you. So, I'll type H list, then a right square bracket and shift enter and I see that it is the Fibonacci sequence. If I want more information on the Fibonacci sequence, I can click the documentation link that appears in the results bar. Now, let's take a look at the second list that I have that is an I list. So again I'll type find sequence function, left square bracket I list, and right square bracket to close, shift enter and I see that is the Lucas series. It's like Fibonnaci where the two values are added together to form the next value and you continue the sequence. This time you have one plus three equals four, three plus four equals seven, seven and 11, 18, 11, 18, 29, and so on. Now finally, I have a sequence that is simply the odd positive integers. It starts at one, three, five, seven, nine, 11, 13. If I use find sequence function on that, which is assigned to J list so I'll do find sequence function, left square bracket J list, right square bracket and shift enter. Then, find sequence function works exactly like find formula, which I demonstrated in the previous movie. Here I have minus one plus two times the value, two times the number. So, two times one minus one is one, two times two minus one is three, and so on. The Fibonnaci and Lucas sequences show up a lot when you're examining data. They show up in nature and on occasion, they will show up in business if you're looking at natural growth patterns. If you suspect there might be an underlying sequence to data that you're looking at, then find the sequence function, especially in combination with find formula is a great way to get insights into your data.

## **Question 1 of 3**

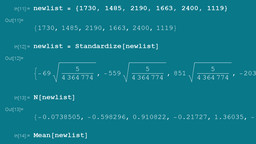
Finding a line that describes a data set with the least error is done through linear \_\_\_\_\_.

* interpolation
* regression  
  Correct
* normalization
* standardization

## **Question 2 of 3**

Estimating the value at a specific time in a time series is an example of a \_\_\_\_\_ forecast.

* specific  
  Incorrect
* standardized  
  Incorrect
* time/date  
  Incorrect
* point



Replay

Review this video

Find a time series that fits given data

4m 34s

## **Question 3 of 3**

Some known progressions of numbers, such as the Fibonacci Sequence, can be identified in Mathematica by using the \_\_\_\_\_ keyword.

* FindSequenceFunction  
  Correct
* LinearModelFit
* FindFormula
* LogisticSigmoid

### **Calculate the logistic sigmoid function for a data set**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Much data analysis, and a lot of everyday analysis, comes down to a simple question, is something A or B? Is it one thing or another? In civil trials, the legal system often asks whether something is more likely than not under the preponderance of evidence standard. Machine learning embodies this concept in the sigmoid function, or logistics sigmoid, where algorithms analyze your data to identify the pattern that makes it more likely for something to belong to class one or class two. In this movie I'll show you how to use Mathematica to display the LogisticSidmoid function for a data range. My sample file is the Logistic Sigmoid notebook, and you can find it in the chapter four folder of your exercise files collection. I'll start by evaluating the notebook. So I'll go to the Evaluation menu and click Evaluate Notebook. And there I have my logistic sigmoid plot that I created earlier. If you examine the plot, there are a couple things to notice. The first is that at the zero point on the horizontal access the function curve crosses at 0.5. So any output of the LogisticSigmoid function that is greater than 0.5 would be one thing, would be class two, call it. And anything that the LogisticSigmoid function generates a value of less than 0.5, that would be class one. Anything that's exactly on the barrier is indeterminate. However, given the nature of the statistics, it is very unlikely that you would get a value of 0.5. Let me show you the formula that is used to calculate the LogisticSigmoid function. So I'll type LogisticSigmoid, followed by a left square bracket and an X, then a right square bracket, then two forward slashes, and I'll use the function expand keyword. So FunctionExpand, and a shift enter. Doing so displays the formula that is used for the LogisticSigmoid function. The LogisticSigmoid function is one divided by one plus E, raised to the minus X. And X is the value in a data range. So if we look at the data range that was used to generate the LogisticSigmoid function here, we see that we have minus 10 to 10. So looking at minus 10 you can see that the value is very close to zero. So we have one divided by one plus, and then E to the minus X. That means if we have minus 10, that is a negative of a negative and therefore a positive. E is Euler's constant, which is about 2.718. So we have 2.718 raised to the 10th power, and three to the 10th power is a very large number. One divided by that would be very close to zero. And that's why you see the curve tailing away the way it does. If you have a positive number, such as 10, then you would have one divided by one plus E to the minus 10. That's a very small number, very close to zero. So the output would be very close to one and that's why the LogisticSigmoid function is very close to one as you go past about 2 1/2. It's also possible to have an unbalanced LogisticSigmoid function. So let's say that I have a plot of the logistic sigmoid of x in square brackets, then a comma, and then I need to define the range. So that's a list. I need to have X, which is the name of the variable. Then the minimum, which I'll call minus seven, and the maximum, which I'll call 14. Then a right curly bracket and a right square bracket to close out all of my arguments. And shift enter, and there is the value. Or there is the plot. And you see at .05 it crosses the vertical axis. As we get closer to minus seven we see that the value's very close to zero, and we've extended the graph to the right to indicate how close the sigmoid function gets to one as the value increases beyond 10 all the way to 14. I know this discussion has been a bit abstract, but the LogisticSigmoid function is used in logistic regression and a lot of other machine learning techniques. Now that you know what it looks like, you can incorporate it in your own work.

### **Classify items using training data**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Voiceover] One of the most fundamental machine learning tasks is to distinguish between one or two options. In this movie, I'll show you how to develop a classification rule from training data. And then apply it to test data. My sample file is the classified notebook. And you can find it in the chapter four folder of your exercise files collection. I have a set of training data, where I have a series of assignments, or associations. Item number one is associated with category a. Number two with category b. And so on. To assign that list of associations to my variables, I need to go to the evaluation menu. And click evaluate notebook. Now my values have been assigned to the variables. My goal is to create a classifier function from my training dataset. I'll enter a variable to assign the model to. I'll call it crule for classification rule. Equal. Then I'll use the classify keyword. With capital c. Followed by a left square bracket. And my training data variable. Which is train1. Then right square bracket to close. And shift enter. And mathematica thinks for a moment. And comes back with a classifier function. Which in this case is a logistic regression. And it identified two classes. And yes, you can have three, four, five classes. It doesn't matter. There were only two in my dataset. A and b. Now I can use my test data to make guesses about which category they would go into. For that, I would type in crule, which again is the variable to which I assigned my model. And then test1. And that's the variable that has the list I want to check. And that list contains 1.6, 2.7, four and 5.6. And I'll press shift enter. And I see that the values are a a and b b. The rule assigned 1.6 to category a, which seems consistent. Same for 2.7. And then it assigned four and 5.6 to category b, which also is consistent with the data. Even though we have a very small training dataset, we were able to define a classification rule, and take our best guess at the values in the test data. And which category they fell into. In this case, the results look pretty good.

### **Predict values using training data**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] One of the most basic questions you can ask about a data set is this. Given and input x, what will the output y be? In this movie, I'll show you how to use Mathematica to create these prediction rules from a training data set. My sample file is the prediction notebook. And you can find it in the chapter four folder of the exercise files collection. Before I get started, I need to evaluate the notebook to assign values to my c list variable. So I will open the Evaluation menu and click Evaluate Notebook. There we go. If I want to create a prediction rule, then I can use a list in this format which has associations between values. So I'll type p rule, for prediction rule, equals, then the Predict keyword, followed by a left square bracket, and then my values. And those are stored in the variable c list. And Shift, Enter. After a moment, Mathematica comes back with a model, in this case a linear regression model with a predictor function. So that means that I can input my own values and get predictions out of them. If I examine the data, I see that I have a smaller value, 1400 which is associated with a larger value of 225,000. In this case the values represent the square footage of a house. And then the sales value of the house. This is the only data that I'm working with. So I don't have the number of bedrooms, the number of bathrooms, neighborhood, anything like that. Well let's say that given this data and this data only, I want to make a prediction about the possible sales price of the house based on these values. For that I can type in p rule, which is the variable to which I assigned my model. Then a left square bracket and the number of square feet of the house I want to use as an input. Let's say it's a fairly small house. 1,000 square feet. Right square bracket to close the argument. Shift, Enter. And I get a value of 146,250. And given that we had a sales price of 140,000 for a house of 980 square feet, that seems like a reasonable estimate. But let's see what happens if we go further away from the bottom of the sequence and look at a larger house. For that I'll do p rule, left square bracket. And let's say I go for a 2100 square foot house. I have values for 1900 and 2300 so let's average those two. Right square bracket to close. Shift, Enter. And I get 301,870. And that's a reasonable number based on 1900 square feet giving us 235,000 in value and 2300 giving us 350,000. Now let's see what happens by way of error checking if I enter in a value that's already in the sequence, what it gives us. So I'll do p rule, left square bracket. And let's go for the largest house in the data set, 2300. Right square bracket. And a Shift, Enter. And we get a predicted value of 330,165. So that's actually not a bad guess. We're off by about $20,000 off a 350,000. But that's a pretty good guess, especially given the small data set that we had to work with. For this type of model, if you are working from data that is consistent, that is from the same neighborhood, and at the same time, in the case of housing sales, if you have a lot of data to work with, you can generate pretty accurate models.

### **Measure predictor function performance**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] Defining predictor functions lets you estimate the output of a function given a specific input value. You can also measure how far the prediction is off for each test set value and plot the difference, which is called a residual. In this movie, I will show you how to measure a predictor function performance. My sample file is the predictor performance notebook. And you can find it in the chapter four folder of you exercise files collection. I have two sets of data. The first is my training set, assigned to the variable train1. And I have an association between square footage of the house and the sale price. I also have my test data assigned to the variable test1. It is a set of three square footage and sale price associations. What I want to do is create a model based on my training data and then use the test data to determine how close the model comes. The first thing I need to do is assign my values to the variables that I have. So I will go to the evaluation menu and click evaluate notebook. I'm going to create a prediction model. So I will assign it to a variable that I'll call pmodel equals, then I'll use the predict keyword, followed by a left square bracket and my training data, which is assigned to train1. I'll type a right square bracket to close out and shift-enter and after a moment, I get my model. Now I can see how accurate the model is and I can do it using a couple of comparisons. I'll start by typing the predictor measurements keyword followed by a left square bracket. Now I need to identify three things: the model that I'm using, which is in the variable pmodel, then a comma, then my test data set, which is test1, then a comma, and finally, how I want to display the results. This is a text argument, or a string, that goes in double quotes and let's take a look at the residual plot, followed by a right square bracket. Residual is another word for the error in a particular measurement. So if I said something was 20, and it turned out to be 18, then I would have a residual of two, which is the difference between 20 and 18. So let's see how our model, which is based on a very small data set, did. I'll press shift-enter and as you can see, we have one residual that's very close. It's within $20,000. And then we have another that is about 70,000 below the actual price. And another that is about 60,000 above the actual price. So what that tells us is that there is substantial error built into the model. Which isn't surprising. We only have four examples to build on. We can also do a comparison plot that looks at the values together. And for that I'll use predictor measurements again. Left square bracket, and then pmodel, which is my model, test1, the test data, a comma. And this time, instead of residual plot, I'll use comparison plot. Followed by a right square bracket and shift-enter. And I see a comparison between the perfect prediction line and the actual predictions. I see that I have my value that was much too high, which is here and it's for a house that actually sold for about $310,000. Also had a substantially low estimate for a house that sold for about 290. And then a not bad estimate for one that sold around 260. Those are two different ways to look at the results of a prediction model. You can look at the residuals, which are the errors, or you can compare the predictions to the actual value.

### **Identify data clusters**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] One interesting way to analyze data is to identify clusters of like elements. For example, you might be able to identify a group of customers who travel from Europe every year and stay in one of your company's hotels. Those customers likely have a lot in common and generate good revenue, so it's worthwhile to identify them and find out what keeps them coming back. In this movie, I'll show you how to identify groups of related data points through cluster analysis. I'll start off in PowerPoint to give you a visual representation of one way to do clustering and then switch to Mathematica. So the main question is how does clustering work? Well the idea is that you have sets of data and it's easiest to visualize on a two dimensional grid. So let's say that we have a bunch of values that are plotted in a scatter plot. Let's say that I want to divide them into two groups or two clusters. One method for doing that is to create a randomly placed centroid or medoid. Centroid referring to the center. Medoid referring to the median or middle. You analyze the data based on the centroids and you move the centroids to the middle of all of the points that are closest to it. So if you look at the left side of the screen, you'll see that the left red dot will probably move to the middle of the dots to its left, and the red dot on the right will move to the middle of the dots on its right. You start by calculating the distance to each centroid. For example, with these two dots here... First you calculate the vertical distance, then the horizontal, and that gives you the direct distance or the hypotenuse. This is called the Euclidean distance. And you can do the same thing for the other point. Once you have that measurement, you can calculate the hypotenuse and move the centroid to the center of the group. So here are two centroids. Here's the data, and after you have identified the group of points that are closest to it, that's its group, and you move to the middle of that group. So there's the one on the right and there's the one on the left. So what do you do now? Well, you do it more than once. And when you define random starting centroids, you might get different groups, based on where the process stops. Also, some sets of centroids don't end in valid solutions. If the centroids are very close to each other, or if one or more of them is at the far edge of the scatter plot, then again, you might not get a valid solution. That means that you need to use your domain knowledge and common sense to interpret the results. With that as background, I'll switch over to Mathematica and show you how to do cluster analysis in that program. My sample file here in Mathematica is the Clustering.nb notebook. You can find it in the Chapter04 folder of the exercise files collection. I have a variable called vpairs and I want to assign it pairs of data. So again, I'm using a scatter plot metaphor. The first value is the number of miles traveled by a particular guest, and the second is the amount of money that they spent at your hotel. And what I'm looking for is a relationship between these values. Now before I start, I will go ahead and evaluate my notebook to actually assign the values to vpairs. So I'll open the Evaluation menu and click Evaluate Notebook. There we go. Now I can start doing my analysis. I will assign the results of a cluster classification to a variable called cm. So cm equal, then I have ClusterClassify followed by a left square bracket, and my data is in the variable called vpairs. And I'll type a right square bracket to close, and Shift + Enter, and after a moment, I get my Classifierfunction that has two separate classes. Let's take a look at the data in a list plot. And that is the mathematical way of creating a scatter plot. So I'll type ListPlot, left square bracket, and then vpairs, and right square bracket, and Shift + Enter, and there's my plot. And it does look like, in fact, that I have two separate clusters... One at the top-right corner of the graph and another larger cluster at the bottom-left. However, it also looks like it might be the case that we have three clusters... One that's very close to the left-hand side, one that's sort of in the middle, and another that's on the right. So again, this is where your domain knowledge comes in. If you want, you can also gather individual points based on their clusters, and identifying them. So we have two clusters in this dataset. If I type GatherBy, and then a left square bracket. I can use my data, which is vpairs, comma, and then cm, which is my cluster model, followed by a right square bracket and Shift + Enter. Then I get my list, my separation of values. And you can see, if you follow the curly brackets, that 3700, 4200, and 3680 are part of one cluster, and 1200, 800, 1000, 120, 90, and 100 are members of the second cluster. As you can see, cluster analysis can give you a lot of insight into your data. If you want to separate your customers into bins based on common characteristics, then creating a cluster classification model is a great way to do it.

### **Identify data clusters**

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] One interesting way to analyze data is to identify clusters of like elements. For example, you might be able to identify a group of customers who travel from Europe every year and stay in one of your company's hotels. Those customers likely have a lot in common and generate good revenue, so it's worthwhile to identify them and find out what keeps them coming back. In this movie, I'll show you how to identify groups of related data points through cluster analysis. I'll start off in PowerPoint to give you a visual representation of one way to do clustering and then switch to Mathematica. So the main question is how does clustering work? Well the idea is that you have sets of data and it's easiest to visualize on a two dimensional grid. So let's say that we have a bunch of values that are plotted in a scatter plot. Let's say that I want to divide them into two groups or two clusters. One method for doing that is to create a randomly placed centroid or medoid. Centroid referring to the center. Medoid referring to the median or middle. You analyze the data based on the centroids and you move the centroids to the middle of all of the points that are closest to it. So if you look at the left side of the screen, you'll see that the left red dot will probably move to the middle of the dots to its left, and the red dot on the right will move to the middle of the dots on its right. You start by calculating the distance to each centroid. For example, with these two dots here... First you calculate the vertical distance, then the horizontal, and that gives you the direct distance or the hypotenuse. This is called the Euclidean distance. And you can do the same thing for the other point. Once you have that measurement, you can calculate the hypotenuse and move the centroid to the center of the group. So here are two centroids. Here's the data, and after you have identified the group of points that are closest to it, that's its group, and you move to the middle of that group. So there's the one on the right and there's the one on the left. So what do you do now? Well, you do it more than once. And when you define random starting centroids, you might get different groups, based on where the process stops. Also, some sets of centroids don't end in valid solutions. If the centroids are very close to each other, or if one or more of them is at the far edge of the scatter plot, then again, you might not get a valid solution. That means that you need to use your domain knowledge and common sense to interpret the results. With that as background, I'll switch over to Mathematica and show you how to do cluster analysis in that program. My sample file here in Mathematica is the Clustering.nb notebook. You can find it in the Chapter04 folder of the exercise files collection. I have a variable called vpairs and I want to assign it pairs of data. So again, I'm using a scatter plot metaphor. The first value is the number of miles traveled by a particular guest, and the second is the amount of money that they spent at your hotel. And what I'm looking for is a relationship between these values. Now before I start, I will go ahead and evaluate my notebook to actually assign the values to vpairs. So I'll open the Evaluation menu and click Evaluate Notebook. There we go. Now I can start doing my analysis. I will assign the results of a cluster classification to a variable called cm. So cm equal, then I have ClusterClassify followed by a left square bracket, and my data is in the variable called vpairs. And I'll type a right square bracket to close, and Shift + Enter, and after a moment, I get my Classifierfunction that has two separate classes. Let's take a look at the data in a list plot. And that is the mathematical way of creating a scatter plot. So I'll type ListPlot, left square bracket, and then vpairs, and right square bracket, and Shift + Enter, and there's my plot. And it does look like, in fact, that I have two separate clusters... One at the top-right corner of the graph and another larger cluster at the bottom-left. However, it also looks like it might be the case that we have three clusters... One that's very close to the left-hand side, one that's sort of in the middle, and another that's on the right. So again, this is where your domain knowledge comes in. If you want, you can also gather individual points based on their clusters, and identifying them. So we have two clusters in this dataset. If I type GatherBy, and then a left square bracket. I can use my data, which is vpairs, comma, and then cm, which is my cluster model, followed by a right square bracket and Shift + Enter. Then I get my list, my separation of values. And you can see, if you follow the curly brackets, that 3700, 4200, and 3680 are part of one cluster, and 1200, 800, 1000, 120, 90, and 100 are members of the second cluster. As you can see, cluster analysis can give you a lot of insight into your data. If you want to separate your customers into bins based on common characteristics, then creating a cluster classification model is a great way to do it.

## **Question 1 of 4**

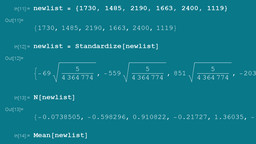
The \_\_\_\_\_ function estimates whether an entity is more likely to be a member of Class 1 or Class 2.

* time series
* interpolation
* logistic sigmoid  
  Correct
* linear regression

## **Question 2 of 4**

The percentage of times a classification rule correctly identifies a member of a class is the rule's error rate.

* TRUE  
  Incorrect
* FALSE



Replay

Review this video

Measure classifier function performance

3m 40s

## **Question 3 of 4**

The difference between a predicted value and the actual value is the \_\_\_\_\_.

* regression
* offset
* residual  
  Correct
* interpolation

## **Question 4 of 4**

Moving the focal points of clusters based on the average position of cluster members is an example of the \_\_\_\_\_ method.

* normalization
* centroid  
  Correct
* regression
* k-nearest neighbors

### **Next steps**

Selecting transcript lines in this section will navigate to timestamp in the video

- Thanks again for working through Mathematica 11 Machine Learning with me. I hope you've gotten a lot out of the course. Machine learning is a vast topic, so I wasn't able to get to everything, so I'd like to give you some more resources to work with before I go. First, make sure you go to the wolfram.com/mathematica site. That site is home to the Mathematica resources, and that includes a huge section on machine learning. If you're looking for a book to read, and you want a general introduction to machine learning, I highly recommend Real-World Machine Learning by Brink, Richards, and Fetherolf from Manning. That's a great book, and it has a little math, but it's mostly about high-level concepts, and the way different techniques work. If you have a stronger math background including statistics, then I recommend Introduction to Machine Learning Third Edition, by Ethem Alpaydin from MIT Press. This book goes into a lot more depth, and again, if you have an understanding of statistics, you can get a lot out of it. And finally, if you want more help on cleaning up your data in Microsoft Excel 2016, I highly recommend Dennis Taylor's LinkedIn Learning course "Excel 2016: Cleaning Up Your Data". He goes into a lot of depth on how to manage your data, and how to get it ready for analysis. Thanks again for working through Mathematic 11 Machine Learning, and I wish you the best of luck.