Machine Learning with Python Notes

Module 1 - Machine Learning

Learning Objectives

In this lesson you will learn about:

- Machine Learning applications
- Python libraries for Machine Learning
- Supervised vs Unsupervised Learning

Intro to Machine Learning

Well, imagine that you've obtained a dataset containing characteristics of thousands of human cell samples extracted from patients who were believed to be at risk of developing cancer.
Analysis of the original data showed that many of the characteristics differed significantly between benign and malignant samples.
You can use the values of these cell characteristics in samples from other patients to give an early indication of whether a new sample might be benign or malignant.
You should clean your data, select a proper algorithm for building a prediction model, and train your model to understand patterns of benign or malignant cells within the data.
Once the model has been trained by going through data iteratively, it can be used to predict your new or unknown cell with a rather high accuracy.
This is machine learning!
It is the way that a machine learning model can do a doctor's task or at least help that doctor make the process faster.Now, let me give a formal definition of machine learning.

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed."
Let me explain what I mean when I say "without being explicitly programmed."
Assume that you have a dataset of images of animals such as cats and dogs, and you want to have software or an application that can recognize and differentiate them.
The first thing that you have to do here is interpret the images as a set of feature sets.
For example, does the image show the animal's eyes?
If so, what is their size?
Does it have ears?
What about a tail?
How many legs?
Does it have wings?
Prior to machine learning, each image would be transformed to a vector of features.
Then, traditionally, we had to write down some rules or methods in order to get computers to be intelligent and detect the animals.
But, it was a failure.
Why?
Well, as you can guess, it needed a lot of rules, highly dependent on the current dataset, and not generalized enough to detect out-of-sample cases.
This is when machine learning entered the scene.

Using machine learning allows us to build a model that looks at all the feature sets, and their corresponding type of animals, and learn it learns the pattern of each animal.
It is a model built by machine learning algorithms.
It detects without explicitly being programmed to do so.
In essence, machine learning follows the same process that a 4-year-old child uses to learn, understand, and differentiate animals.
So, machine learning algorithms, inspired by the human learning process, iteratively learn from data, and allow computers to find hidden insights.
These models help us in a variety of tasks, such as object recognition, summarization, recommendation, and so on.
Machine Learning impacts society in a very influential way.
Here are some real-life examples.
First, how do you think Netflix and Amazon recommend videos, movies, and TV shows to its users?
They use Machine Learning to produce suggestions that you might enjoy!
This is similar to how your friends might recommend a television show to you, based on their knowledge of the types of shows you like to watch.
How do you think banks make a decision when approving a loan application?
They use machine learning to predict the probability of default for each applicant, and then approve
or refuse the loan application based on that probability.
Telecommunication companies use their customers' demographic data to segment them, or predict
if they will unsubscribe from their company the next month.
There are many other applications of machine learning that we see every day in our daily life, such as chatbots, logging into our phones or even computer games using face recognition.
Each of these use different machine learning techniques and algorithms.
So, let's quickly examine a few of the more popular techniques.
The Regression/Estimation technique is used for predicting a continuous value, for example, predicting things like the price of a house based on its characteristics, or to estimate the Co2 emission from a car's engine.
A Classification technique is used for Predicting the class or category of a case, for example, if a cell is benign or malignant, or whether or not a customer will churn.
Clustering groups of similar cases, for example, can find similar patients, or can be used for customer segmentation in the banking field.
Association technique is used for finding items or events that often co-occur, for example, grocery items that are usually bought together by a particular customer.
Anomaly detection is used to discover abnormal and unusual cases, for example, it is used for credit card fraud detection.
Sequence mining is used for predicting the next event, for instance, the click-stream in websites.
Dimension reduction is used to reduce the size of data.
And finally, recommendation systems; this associates people's preferences with others who have similar tastes, and recommends new items to them, such as books or movies.
We will cover some of these techniques in the next videos.
By this point, I'm quite sure this question has crossed your mind, "What is the difference between these buzzwords that we keep hearing these days, such as Artificial intelligence (or AI), Machine Learning and Deep Learning?"
Well, let me explain what is different between them.
In brief, AI tries to make computers intelligent in order to mimic the cognitive functions of humans.

So, Artificial Intelligence is a general field with a broad scope including: Computer Vision, Language Processing, Creativity, and Summarization.
Machine Learning is the branch of AI that covers the statistical part of artificial intelligence.
It teaches the computer to solve problems by looking at hundreds or thousands of examples, learning from them, and then using that experience to solve the same problem in new situations.
And Deep Learning is a very special field of Machine Learning where computers can actually learn and make intelligent decisions on their own.
Deep learning involves a deeper level of automation in comparison with most machine learning algorithms.
Now that we've completed the introduction to Machine Learning, subsequent videos will focus on reviewing two main components: First, you'll be learning about the purpose of Machine Learning and where it can be applied in the real world; and
Second, you'll get a general overview of Machine Learning topics, such as supervised vs unsupervised learning, model evaluation and various Machine Learning algorithms.
So now that you have a sense with what's in store on this journey, let's continue our exploration of Machine Learning!
Thanks for watching!

Python for Machine Learning

Hello, and welcome!
In this video, we'll talk about how to use python for machine learning.
So let's get started.
Python is a popular and powerful general-purpose programming language that recently emerged
as the preferred language among data scientists.
You can write your machine learning algorithm using python, and it works very well.
However, there are a lot of modules and libraries already implemented in python that can make your life much easier.
We try to introduce the Python packages in this course and use it in the labs to give you better hands-on experience.
The first package is Numpy, which is a math library to work with n-dimensional arrays in Python.
It enables you to do computation efficiently and effectively.
It is better than regular python because of it's amazing capabilities.
For example, for working with arrays, dictionaries, functions, datatypes, and working with images,
you need to know Numpy.
SciPy is a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics and much more.
SciPy is a good library for scientific and high-performance computation.
Matplotlib is a very popular plotting package that provides 2D plotting as well as 3D plotting.
Basic Knowledge about these 3 packages, which are built on top of python, is a good asset for data scientists who want to work with real world problems.
If you are not familiar with these packages, I recommend that you take the "Data Analysis with Python" course first.
This course covers most of the useful topics in these packages.
Pandas library, is a very high-level python library that provides high-performance, easy

to use data structures.
It has many functions for data importing, manipulation and analysis.
In particular, it offers data structures and operations for manipulating numerical tables and time series.

scikit-learn is a collection of algorithms and tools for machine learning, which is our focus here, and which you'll learn to use with in this course.
As we'll be using scikit-learn quite a bit, in the labs, let me explain more about it and show you why it is so popular among data scientists.
Scikit-learn is a free machine learning library for the Python programming language.
It has most of the classification, regression and clustering algorithms, and it's designed to work with the Python numerical and scientific libraries, NumPy and SciPy.
Also, it includes very good documentation.
On top of that, implementing machine learning models with scikit learn is really easy, with a few lines of python code.
Most of the tasks that need to be done in a machine learning pipeline are implemented already in scikit learn, including, pre-processing of data, feature selection, feature extraction, train/test splitting, defining the algorithms, fitting models, tuning parameters, prediction, evaluation, and exporting the model.
Let me show you an example of how scikit learn looks like when you use this library.
You don't have to understand the code for now, but just see how easily you can build a model with just a few lines of code.
Basically, Machine learning algorithms benefit from standardization of the data set.
If there are some outliers, or different scales fields in your data set, you have to fix them.
The preprocessing package of scikit learn provides several common utility functions and transformer classes to change raw feature vectors into a suitable form of vector for modeling.
You have to split your dataset into train and test sets to train your model, and then test the model's accuracy separately.
Scikit learn can split arrays or matrices into random train and test subsets for you, in one line of code.
Then, you can setup your algorithm.
For example, you can build a classifier using a support vector classification algorithm.
We call our estimator instance clf, and initialize its parameters.
Now, you can train your model with the train set.
By passing our training set to the fit method, the clf model learns to classify unknown cases.
Then, we can use our test set to run predictions.
And, the result tells us what the class of each unknown value is.
Also, you can use different metrics to evaluate your model accuracy, for example, using a confusion matrix to show the results.
And finally, you save your model.
You may find all or some of these machine learning terms confusing, but don't worry, we will talk about all of these topics in the following videos.
The most important point to remember is that the entire process of a Machine Learning task can be done simply in a few lines of code, using scikit learn.
Please notice that, though it is possible, it would not be that easy if you want to do all of this using Numpy or Scipy packages.
And of course, it needs much more coding if you use pure python programming to implement all of these tasks.
Thanks for watching.

Supervised vs Unsupervised

Hello, and welcome!
In this video, we'll introduce supervised algorithms versus unsupervised algorithms.
So let's get started.
An easy way to begin grasping the concept of supervised learning is by looking directly at the words that make it up.
Supervise means to observe and direct the execution of a task, project, or activity.
Obviously, we aren't going to be supervising a person…
Instead, we'll be supervising a machine learning model that might be able to produce classification regions like we see here.
So, how do we supervise a machine learning model?
We do this by "teaching" the model.
That is, we load the model with knowledge so that we can have it predict future instances.
But … this leads to the next question, which is, "How exactly do we teach a model?"
We teach the model by training it with some data from a labeled dataset.
It's important to note that the data is labeled.
And what does a labeled dataset look like?
Well, it can look something like this.
This example is taken from the cancer dataset.
As you can see, we have some historical data for patients, and we already know the class of each row.
Let's start by introducing some components of this table.
The names up here, which are called Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, and so on, are called Attributes.
The columns are called Features, which include the data.
If you plot this data, and look at a single data point on a plot, it'll have all of these attributes.
That would make a row on this chart, also referred to as an observation.
Looking directly at the value of the data, you can have two kinds.
The first is numerical.
When dealing with machine learning, the most commonly used data is numeric.
The second is categorical… that is, it's non-numeric, because it contains characters rather than numbers.
In this case, it's categorical because this dataset is made for Classification.

There are two types of Supervised Learning techniques.
They are: classification and regression.
Classification is the process of predicting a discrete class label or category.
Regression is the process of predicting a continuous value as opposed to predicting a categorical value in Classification.
Look at this dataset.
It is related to Co2 emissions of different cars.
It includes Engine size, Cylinders, Fuel Consumption and Co2 emission of various models of automobiles.
Given this dataset, you can use regression to predict the Co2 emission of a new car by using other fields, such as Engine size or number of Cylinders.
Since we know the meaning of supervised learning, what do you think unsupervised learning means?
Yes!
Unsupervised Learning is exactly as it sounds.

We do not supervise the model, but we let the model work on its own to discover information that may not be visible to the human eye.
It means, The Unsupervised algorithm trains on the dataset, and draws conclusions on UNLABELED
data.
Generally speaking, unsupervised learning has more difficult algorithms than supervised learning, since we know little to no information about the data, or the outcomes that are to be expected.
Dimension reduction, Density estimation, Market basket analysis and Clustering are the most widely used unsupervised machine learning techniques.
Dimensionality Reduction and/or feature selection play a large role in this by reducing redundant features to make the classification easier.
Market basket analysis is a modelling technique based upon the theory that if you buy a certain group of items, you're more likely to buy another group of items.
Density estimation is a very simple concept that is mostly used to explore the data to find some structure within it.
And finally, clustering.
Clustering is considered to be one of the most popular unsupervised machine learning techniques used for grouping data points or objects that are somehow similar.
Cluster analysis has many applications in different domains, whether it be a bank's desire to segment its customers based on certain characteristics, or helping an individual to organize and group his/her favourite types of music!
Generally speaking, though, Clustering is used mostly for: Discovering structure, Summarization,
and Anomaly detection.
So, to recap, the biggest difference between Supervised and Unsupervised Learning is that supervised learning deals with labeled data while Unsupervised Learning deals with unlabeled data.
In supervised learning, we have machine learning algorithms for Classification and Regression.
In unsupervised learning, we have methods such as clustering.
In comparison to supervised learning, unsupervised learning has fewer models and fewer evaluation
methods that can be used to ensure that the outcome of the model is accurate.
As such, unsupervised learning creates a less controllable environment, as the machine is creating outcomes for us.
Thanks for watching!

Quiz

Review Question 1

1/1 point (graded)

Machine Learning uses algorithms that can learn from data without relying on explicitly programmed methods.

Answer: True

Review Question 2

1/1 point (graded)

Which are the two types of supervised learning techniques?

Answer: Classification and Regression

Review Question 3

1/1 point (graded)

Which of the following statements best describes the Python scikit library?

Answer: A collection of algorithms and tools for machine learning.

Evaluation Metrics in Regression

Hello, and welcome!In this video, we'll be covering accuracy metrics for model evaluation.So, let's get started.Evaluation metrics are used to explain the performance of a model.Let's talk more about the model evaluation metrics that are used for regression.As mentioned, basically, we can compare the actual values and predicted values to calculate the accuracy of a regression model.Evaluation metrics provide a key role in the development of a model, as it provides insight to areas that require improvement.We'll be reviewing a number of model evaluation metrics, including:Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).But, before we get into defining these, we need to define what an error actually is.In the context of regression, the error of the model is the difference between the data points and the trend line generated by the algorithm.Since there are multiple data points, an error can be determined in multiple ways.Mean absolute error is the mean of the absolute value of the errors.This is the easiest of the metrics to understand, since it's just the average error.Mean Squared Error (MSE) is the mean of the squared error.It's more popular than Mean absolute error because the focus is geared more towards large errors.This is due to the squared term exponentially increasing larger errors in comparison to smaller ones.Root Mean Squared Error (RMSE) is the square root of the mean squared error.This is one of the most popular of the evaluation metrics because Root Mean Squared Error is interpretable in the same units as the response vector (or 'y' units) making it easy to relate its information.Relative Absolute Error (RAE), also known as Residual sum of square, where y-bar is a mean value of y, takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.Relative Squared Error (RSE) is very similar to "Relative absolute error ", but is widely adopted by the data science community, as it is used for calculating R-squared.R-squared is not error, per se, but is a popular metric for the accuracy of your model.It represents how close the data values are to the fitted regression line.The higher the R-squared, the better the model fits your data.Each of these metrics can be used for quantifying of your prediction.The choice of

metric completely depends on the type of model, your data type, and domain of knowledge.Unfortunately, further review is out of scope of this course.Thanks for watching!

Module 2 - Regression

Learning Objectives

In this lesson you will learn about:

- Regression Algorithms
- Model Evaluation
- Model Evaluation: Overfitting & Underfitting
- Understanding Different Evaluation Models
- Simple Linear Regression

https://courses.cognitiveclass.ai/courses/course-v1:BDU+ML0101EN+v4/courseware/bd64ccdf56ad4ea1afe870e26d583038/6d061cb07d0c4659ac479f72e041c5fc/?child=first

Intro to Regression

Hello, and welcome!
In this video, we'll be giving a brief introduction to regression.
So let's get started.
Look at this dataset.
It's related to Co2 emissions from different cars.
It includes Engine size, number of Cylinders, Fuel Consumption and Co2 emission from various automobile models.
The question is, "Given this dataset, can we predict the Co2 emission of a car using other fields, such as EngineSize or Cylinders?"
Let's assume we have some historical data from different cars, and assume that a car, such as in row 9, has not been manufactured yet, but we're interested in estimating its approximate Co2 emission, after production.
Is it possible?
We can use regression methods to predict a continuous value, such as CO2 Emission, using some other variables.
Indeed, regression is the process of predicting a continuous value.
In regression there are two types of variables: a dependent variable and one or more independent
variables.
The dependent variable can be seen as the "state", "target" or "final goal" we study and try to predict, and the independent variables, also known as explanatory variables, can be

seen as the "causes" of those "states".
The independent variables are shown conventionally by x; and the dependent variable is notated
by y.
A regression model relates y, or the dependent variable, to a function of x, i.e., the independent variables.
The key point in the regression is that our dependent value should be continuous, and cannot be a discreet value.
However, the independent variable or variables can be measured on either a categorical or continuous measurement scale.
So, what we want to do here is to use the historical data of some cars, using one or more of their features, and from that data, make a model.
We use regression to build such a regression/estimation model.
Then the model is used to predict the expected Co2 emission for a new or unknown car.
Basically there are 2 types of regression models: simple regression and multiple regression.
Simple regression is when one independent variable is used to estimate a dependent variable.
It can be either linear on non-linear.
For example, predicting Co2emission using the variable of EngineSize.
Linearity of regression is based on the nature of relationship between independent and dependent
variables.
When more than one independent variable is present, the process is called multiple linear regression.
For example, predicting Co2emission using EngineSize and the number of Cylinders in any given car.
Again, depending on the relation between dependent and independent variables, it can be either
linear or non-linear regression.
Let’s examine some sample applications of regression.
Essentially, we use regression when we want to estimate a continuous value.
For instance, one of the applications of regression analysis could be in the area of sales forecasting.
You can try to predict a salesperson's total yearly sales from independent variables such as age, education, and years of experience.
It can also be used in the field of psychology, for example, to determine individual satisfaction based on demographic and psychological factors.
We can use regression analysis to predict the price of a house in an area, based on its size, number of bedrooms, and so on.
We can even use it to predict employment income for independent variables, such as hours of work, education, occupation, sex, age, years of experience, and so on.
Indeed, you can find many examples of the usefulness of regression analysis in these and many other fields or domains, such as finance, healthcare, retail, and more.
We have many regression algorithms.
Each of them has its own importance and a specific condition to which their application

is best suited.
And while we've covered just a few of them in this course, it gives you enough base knowledge for you to explore different regression techniques.
Thanks for watching!

Simple Linear Regression

Hello, and welcome! In this video, we'll be covering linear regression.
You don't need to know any linear algebra to understand topics in linear regression.
This high-level introduction will give you enough background information on linear regression to be able to use it effectively on your own problems.
So, let's get started.
Let's take a look at this dataset. It's related to the Co2 emission of different cars. It includes Engine size, Cylinders, Fuel Consumption and Co2 emissions for various car models. The question is: Given this dataset, can we predict the Co2 emission of a car, using another field, such as Engine size?
Quite simply, yes! We can use linear regression to predict a continuous value such as Co2 Emission, by using other variables.
Linear regression is the approximation of a linear model used to describe the relationship between two or more variables. In simple linear regression, there are two variables: a dependent variable and an independent variable.
The key point in the linear regression is that our dependent value should be continuous and cannot be a discreet value. However, the independent variable(s) can be measured on either a categorical or continuous measurement scale.
There are two types of linear regression models. They are: simple regression and multiple regression.
Simple linear regression is when one independent variable is used to estimate a dependent variable. For example, predicting Co2 emission using the EngineSize variable. When more than one independent variable is present, the process is called multiple linear regression.
For example, predicting Co2 emission using EngineSize and Cylinders of cars.
Our focus in this video is on simple linear regression.
Now, let's see how linear regression works. OK, so let's look at our dataset again.
To understand linear regression, we can plot our variables here.
We show Engine size as an independent variable, and Emission as the target value that we would like to predict. A scatterplot clearly shows the relation between variables where changes in one variable "explain" or possibly "cause" changes in the other variable.
Also, it indicates that these variables are linearly related.
With linear regression you can fit a line through the data.
For instance, as the EngineSize increases, so do the emissions.

With linear regression, you can model the relationship of these variables.
A good model can be used to predict what the approximate emission of each car is.
How do we use this line for prediction now? Let us assume, for a moment, that the line is a good fit of data. We can use it to predict the emission of an unknown car. For example, for a sample car, with engine size 2.4, you can find the emission is 214.
Now, let's talk about what this fitting line actually is.
We're going to predict the target value, y.
In our case, using the independent variable, "Engine Size," represented by x1.
The fit line is shown traditionally as a polynomial. In a simple regression problem (a single x), the form of the model would be θ0 +θ1 x1. In this equation, yˆ is the dependent variable or the predicted value, and x1 is the independent variable; θ0 and θ1 are the parameters of the line that we must adjust. θ1 is known as the "slope" or "gradient" of the fitting line and θ0 is known as the "intercept."
θ0 and θ1 are also called the coefficients of the linear equation.
You can interpret this equation as yˆ being a function of x1, or yˆ being dependent of x1.
Now the questions are: "How would you draw a line through the points?" And, "How do you determine which line 'fits best'?"
Linear regression estimates the coefficients of the line.
This means we must calculate θ0 and θ1 to find the best line to 'fit' the data.
This line would best estimate the emission of the unknown data points.
Let's see how we can find this line, or to be more precise, how we can adjust the parameters to make the line the best fit for the data.
For a moment, let's assume we've already found the best fit line for our data.
Now, let's go through all the points and check how well they align with this line.
Best fit, here, means that if we have, for instance, a car with engine size x1=5.4, and actual Co2=250, its Co2 should be predicted very close to the actual value, which is y=250, based on historical data.
But, if we use the fit line, or better to say, using our polynomial with known parameters to predict the Co2 emission, it will return yˆ=340.
Now, if you compare the actual value of the emission of the car with what we predicted using our model, you will find out that we have a 90-unit error.
This means our prediction line is not accurate. This error is also called the residual error.
So, we can say the error is the distance from the data point to the fitted regression line.
The mean of all residual errors shows how poorly the line fits with the whole dataset.
Mathematically, it can be shown by the equation, mean squared error, shown as (MSE).
Our objective is to find a line where the mean of all these errors is minimized.
In other words, the mean error of the prediction using the fit line should be minimized.
Let's re-word it more technically. The objective of linear regression is to minimize this MSE equation, and to minimize it, we should find the best parameters, θ0 and θ1.
Now, the question is, how to find θ0 and θ1 in such a way that it minimizes this error?
How can we find such a perfect line? Or, said another way, how should we find the best parameters for our line? Should we move the line a lot randomly and

calculate the MSE value every time, and choose the minimum one?
Not really! Actually, we have two options here:
Option 1 - We can use a mathematic approach. Or, Option 2 - We can use an optimization approach.
Let's see how we can easily use a mathematic formula to find the θ0 and θ1.
As mentioned before, θ0 and θ1, in the simple linear regression, are the coefficients of the fit line. We can use a simple equation to estimate these coefficients. That is, given that it's a simple linear regression, with only 2 parameters, and knowing that θ0 and θ1 are the intercept and slope of the line, we can estimate them directly from our data.
It requires that we calculate the mean of the independent and dependent or target columns, from the dataset. Notice that all of the data must be available to traverse and calculate the parameters. It can be shown that the intercept and slope can be calculated using these equations. We can start off by estimating the value for θ1.
This is how you can find the slope of a line based on the data. $\bar{x}$ is the average value for the engine size in our dataset. Please consider that we have 9 rows here, row 0 to 8. First, we calculate the average of x1 and average of y. Then we plug it into the slope equation, to find θ1. The xi and yi in the equation refer to the fact that we need to repeat these calculations across all values in our dataset and i refers to the i'th value of x or y.
Applying all values, we find θ1=39; it is our second parameter.
It is used to calculate the first parameter, which is the intercept of the line.
Now, we can plug θ1 into the line equation to find θ0.
It is easily calculated that θ0=125.74. So, these are the two parameters for the line, where θ0 is also called the bias coefficient and θ1 is the coefficient for the Co2 Emission column. As a side note, you really don't need to remember the formula for calculating these parameters, as most of the libraries used for machine learning in Python, R, and Scala can easily find these parameters for you. But it's always good to understand how it works.
Now, we can write down the polynomial of the line.
So, we know how to find the best fit for our data, and its equation.
Now the question is: "How can we use it to predict the emission of a new car based on its engine size?"
After we found the parameters of the linear equation, making predictions is as simple as solving the equation for a specific set of inputs.
Imagine we are predicting Co2 Emission(y) from EngineSize(x) for the Automobile in record number 9. Our linear regression model representation for this problem would be: $\hat{y} = \theta 0 + \theta 1\ x1$.
Or if we map it to our dataset, it would be Co2Emission = θ0 + θ1 EngineSize.
As we saw, we can find θ0, θ1 using the equations that we just talked about.
Once found, we can plug in the equation of the linear model.

For example, let's use θ0=125 and θ1=39. So, we can rewrite the linear model as $Co2Emission=125+39EngineSize$.
Now, let's plug in the 9th row of our dataset and calculate the Co2 Emission for a car with an EngineSize of 2.4. So Co2Emission = 125 + 39 × 2.4.
Therefore, we can predict that the Co2 Emission for this specific car would be 218.6.
Let's talk a bit about why Linear Regression is so useful.
Quite simply, it is the most basic regression to use and understand.
In fact, one reason why Linear Regression is so useful is that it's fast!
It also doesn't require tuning of parameters. So, something like tuning the K parameter in K-Nearest Neighbors or the learning rate in Neural Networks isn't something to worry about. Linear Regression is also easy to understand
and highly interpretable.
Thanks for watching this video.

Model Evaluation in Regression Models

Hello, and welcome!In this video, we'll be covering model evaluation. So, let's get started.The goal of regression is to build a model to accurately predict an unknown case.To this end, we have to perform regression evaluation after building the model.In this video, we'll introduce and discuss two types of evaluation approaches that can be used to achieve this goal.These approaches are: train and test on the same dataset, and train/test split.We'll talk about what each of these are, as well as the pros and cons of using each of these models.Also, we'll introduce some metrics for accuracy of regression models.Let's look at the first approach.When considering evaluation models, we clearly want to choose the one that will give us the most accurate results.So, the question is, how we can calculate the accuracy of our model?In other words, how much can we trust this model for prediction of an unknown sample,using a given a dataset and having built a model such as linear regression.One of the solutions is to select a portion of our dataset for testing.For instance, assume that we have 10 records in our dataset.We use the entire dataset for training, and we build a model using this training set.Now, we select a small portion of the dataset, such as row numbers 6 to 9, but without the labels.This set, is called a test set, which has the labels, but the labels are not used for prediction, and is used only as ground truth.The labels are called "Actual values" of the test set.Now, we pass the feature set of the testing portion to our built model, and predict the target values.Finally, we compare the predicted values by our model with the actual values in the testset.This indicates how accurate our model actually is.There are different metrics to report the accuracy of the model, but most of them work generally, based on the similarity of the predicted and actual values.Let's look at one of the simplest metrics to calculate the accuracy of our regressionmodel.As mentioned, we just compare the actual values, y, with the predicted values, which is noted as yˆfor the testing set.The error of the model is calculated as the average difference between the predicted and actual values for all the rows.We can write this error as an equation.So, the first evaluation approach we just talked about is the simplest one: train and test

on the SAME dataset.Essentially, the name of this approach says it all … you train the model on the entire dataset, then you test it using a portion of the same dataset.In a general sense, when you test with a dataset in which you know the target value for each data point, you're able to obtain a percentage of accurate predictions for the model.This evaluation approach would most likely have a high "training accuracy" and slow "out-of-sample accuracy", since the model knows all of the testing data points from the training.What is training accuracy and out-of-sample accuracy?We said that training and testing on the same dataset produces a high training accuracy,but what exactly is "training accuracy?"Training accuracy is the percentage of correct predictions that the model makes when using the test dataset.However, a high training accuracy isn't necessarily a good thing.For instance, having a high training accuracy may result in an 'over-fit' of the data.This means that the model is overly trained to the dataset, which may capture noise and produce a non-generalized model.Out-of-sample accuracy is the percentage of correct predictions that the model makes ondata that the model has NOT been trained on.Doing a "train and test" on the same dataset will most likely have low out-of-sample accuracy due to the likelihood of being over-fit.It's important that our models have high, out-of-sample accuracy, because the purpose of our model is, of course, to make correct predictions on unknown data.So, how can we improve out-of-sample accuracy?One way is to use another evaluation approach called "Train/Test Split."In this approach, we select a portion of our dataset for training, for example, rows 0 to 5.And the rest is used for testing, for example, rows 6 to 9.The model is built on the training set.Then, the test feature set is passed to the model for prediction.And finally, the predicted values for the test set are compared with the actual values of the testing set.This second evaluation approach, is called "Train/Test Split."Train/Test Split involves splitting the dataset into training and testing sets, respectively,which are mutually exclusive, after which, you train with the training set and test with the testing set.This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is NOT part of the dataset that has been used to train the data.It is more realistic for real world problems.This means that we know the outcome of each data point in this dataset, making it great to test with!And since this data has not been used to train the model, the model has no knowledge of the outcome of these data points.So, in essence, it's truly out-of-sample testing.However, please ensure that you train your model with the testing set afterwards, as you don't want to lose potentially valuable data.The issue with train/test split is that it's highly dependent on the datasets on which the data was trained and tested.The variation of this causes train/test split to have a better out-of-sample prediction than training and testing on the same dataset, but it still has some problems due to thisdependency.Another evaluation model, called "K-Fold Cross-Validation" resolves most of these issues.How do you fix a high variation that results from a dependency?Well, you average it.Let me explain the basic concept of "k-fold cross-validation" to see how we can solve this problem.The entire dataset is represented by the points in the image at the top left.If we have k=4 folds, then we split up this dataset as shown here.In the first fold, for example, we use the first 25 percent of the dataset for testing,and the rest for training.The model is built using the training set, and is evaluated using the test set.Then, in the next round (or in the second fold), the second 25 percent of the dataset is used for testing and the rest for training the model.Again the accuracy of the model is calculated.We continue for all folds.Finally the result of all 4 evaluations are averaged.That is, the accuracy of each fold is then averaged, keeping in mind that each fold is distinct, where no training data in

one fold is used in another.K-fold cross-validation, in its simplest form, performs multiple train/test splits using the same dataset where each split is different.Then, the result is averaged to produce a more consistent out-of-sample accuracy.We wanted to show you an evaluation model that addressed some of the issues we've described in the previous approaches.However, going in-depth with the K-fold cross-validation model is out of the scope for this course.

Multiple Linear Regression

Hello, and welcome! In this video, we'll be covering multiplelinear regression.As you know there are two types of linear regression models: simple regression and multiple regression. Simple linear regression is when one independent variable is used to estimate a dependent variable. For example, predicting Co2 emission using the variable of EngineSize. In reality, there are multiple variables that predict the Co2 emission. When multiple independent variables are present,the process is called "multiple linear regression." For example, predicting Co2 emission usingEngineSize and the number of Cylinders in the car's engine.Our focus in this video is on multiple linear regression.The good thing is that multiple linear regression is the extension of the simple linear regression model. So, I suggest you go through the Simple Linear Regression video first, if you haven't watched it already.Before we dive into a sample dataset and see how multiple linear regression works, I want to tell you what kind of problems it can solve; when we should use it; and, specifically,what kind of questions we can answer using it.Basically, there are two applications for multiple linear regression.First, it can be used when we would like to identify the strength of the effect that the independent variables have on a dependent variable.For example, does revision time, test anxiety, lecture attendance, and gender, have any effect on exam performance of students? Second, it can be used to predict the impact of changes. That is, to understand how the dependent variable changes when we change the independent variables. For example, if we were reviewing a person's health data, a multiple linear regression can tell you how much that person's blood pressure goes up (or down) for every unit increase (or decrease) in a patient's body mass index (BMI), holding other factors constant.As is the case with simple linear regression, multiple linear regression is a method of predicting a continuous variable. It uses multiple variables, called independent variables, or predictors, that best predict the value of the target variable, which is also called the dependent variable. In multiple linear regression, the targetvalue, y, is a linear combination of independent variables, x.For example, you can predict how much Co2 a car might emit due to independent variables,such as the car's Engine Size, Number of Cylinders and Fuel Consumption.Multiple linear regression is very useful because you can examine which variables are significant predictors of the outcome variable. Also, you can find out how each feature impacts the outcome variable. And again, as is the case in simple linear regression, if you manage to build such a regression model, you can use it to predict the emission amount of an unknown case, such as record number 9.Generally, the model is of the form: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ and so on, up to ... $+\theta_n x_n$.Mathematically, we can show it as a vector form as well.This means, it can be shown as a dot product of 2 vectors: the parameters vector and the feature set vector.Generally, we can show the equation for a multi-dimensional space as $\theta^T x$, where θis an n-by-one vector of unknown parameters in a multi-dimensional space, and x is the vector of the feature sets, as θ

is a vector of coefficients, and is supposed to be multiplied by x. Conventionally, it is shown as transpose θ.θ is also called the parameters, or, weight vector of the regression equation … both these terms can be used interchangeably. And x is the feature set, which represents a car. For example x1 for engine size, or x2 for cylinders, and so on.The first element of the feature set would be set to 1, because it turns them θ_0 into the intercept or bias parameter when the vector is multiplied by the parameter vector.Please notice that θ^T x in a one dimensional space, is the equation of a line.It is what we use in simple linear regression. In higher dimensions, when we have more than one input (or x), the line is called a plane or a hyper-plane.And this is what we use for multiple linear regression.So, the whole idea is to find the best fit hyper-plane for our data.To this end, and as is the case in linear regression, we should estimate the values for θ vector that best predict the value of the target field in each row.To achieve this goal, we have to minimize the error of the prediction.Now, the question is, "How do we find the optimized parameters?"To find the optimized parameters for our model, we should first understand what the optimized parameters are. Then we will find a way to optimize the parameters.In short, optimized parameters are the ones which lead to a model with the fewest errors.Let's assume, for a moment, that we have already found the parameter vector of our model. It means we already know the values of θvector. Now, we can use the model, and the featureset of the first row of our dataset to predict the Co2 emission for the first car, correct?If we plug the feature set values into the model equation, we find yˆ.Let's say, for example, it returns 140 as the predicted value for this specific row.What is the actual value? y=196. How different is the predicted value from the actual value of 196? Well, we can calculate it quite simply, as196-140, which of course = 56. This is the error of our model, only for one row, or one car, in our case. As is the case in linear regression, we can say the error here is the distance from the data point to the fitted regression model.The mean of all residual errors shows how bad the model is representing the dataset.It is called the mean squared error, or MSE.Mathematically, MSE can be shown by an equation. While this is not the only way to expose the error of a multiple linear regression model, it is one the most popular ways to do so.The best model for our dataset is the one with minimum error for all prediction values.So, the objective of multiple linear regression is to minimize the MSE equation.To minimize it, we should find the best parameters θ, but how?Okay, "How do we find the parameter or coefficients for multiple linear regression?"There are many ways to estimate the value of these coefficients.However, the most common methods are the ordinary least squares and optimization approach.Ordinary least squares tries to estimate the values of the coefficients by minimizing the "Mean Square Error." This approach uses the data as a matrix and uses linear algebra operations to estimate the optimal values for the theta.The problem with this technique is the time complexity of calculating matrix operations,as it can take a very long time to finish. When the number of rows in your dataset isless 10,000 you can think of this technique as an option, however, for greater values,you should try other faster approaches.The second option is to use an optimization algorithm to find the best parameters.That is, you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.For example, you can use Gradient Descent, which starts optimization with random values for each coefficient. Then, calculates the errors, and tries to minimize it through wise changing of the coefficients in multiple iterations.Gradient descent is a proper approach if you have a large dataset.Please understand, however, that there are other approaches to estimate

the parameters of the multiple linear regression that you can explore on your own.After you find the best parameters for your model, you can go to the prediction phase.After we found the parameters of the linear equation, making predictions is as simple as solving the equation for a specific set of inputs.Imagine we are predicting Co2 emission (or y) from other variables for the automobile in record number 9. Our linear regression model representation for this problem would be: y̆=θ^T x. Once we find the parameters, we can plug them into the equation of the linear model. For example, let's use θ0 = 125, θ1 = 6.2,θ2 = 14, and so on. If we map it to our dataset, we can rewrite the linear model as "Co2 Emission=125 plus 6.2 multiplied by EngineSize plus 14 multiplied by Cylinder," and so on. As you can see, multiple linear regression estimates the relative importance of predictors. For example, it shows Cylinder has higher impact on Co2 emission amounts in comparison with EngineSize.Now, let's plug in the 9th row of our dataset and calculate the Co2 emission for a car with the EngineSize of 2.4. So Co2Emission=125 + 6.2 × 2.4 + 14 × 4… and so on. We can predict the Co2 emission for this specific car would be 214.1.Now let me address some concerns that you might already be having regarding multiplelinear regression. As you saw, you can use multiple independent variables to predict a target value in multiple linear regression.It sometimes results in a better model compared to using a simple linear regression, which uses only one independent variable to predict the dependent variable.Now, the question is, "How many independent variables should we use for the prediction?"Should we use all the fields in our dataset? Does adding independent variables to a multiple linear regression model always increase the accuracy of the model?Basically, adding too many independent variables without any theoretical justification may result in an over-fit model. An over-fit model is a real problem because it is too complicated for your data set and not general enough to be used for prediction.So, it is recommended to avoid using many variables for prediction.There are different ways to avoid overfitting a model in regression, however, that is outside the scope of this video.The next question is, "Should independent variables be continuous?"Basically, categorical independent variables can be incorporated into a regression model by converting them into numerical variables. For example, given a binary variable such as car type, the code dummies ``0" for "Manual" and 1 for "automatic" cars.As a last point, remember that "multiple linear regression" is a specific type of linear regression. So, there needs to be a linear relationship between the dependent variable and each of your independent variables.There are a number of ways to check for linear relationship.For example, you can use scatter plots, and then visually check for linearity.If the relationship displayed in your scatterplot is not linear, then, you need to use nonlinear regression.

Non-Linear Regression

Hello, and welcome!In this video, we'll be covering non-linear regression basics.So let's get started!These data points correspond to China's Gross Domestic Product (or GDP) from 1960 to 2014.The first column, is the years, and the second, is China's corresponding annual gross domestic income in US dollars for that year.This is what the data points look like.Now, we have a couple of interesting questions.First, "Can GDP be predicted based on time?"And second,

“Can we use a simple linear regression to model it?”Indeed, if the data shows a curvy trend, then linear regression will not produce very accurate results when compared to a non-linear regression -- simply because, as the name implies, linearregression presumes that the data is linear.The scatterplot shows that there seems to be a strong relationship between GDP and time,but the relationship is not linear.As you can see, the growth starts off slowly, then from 2005 onward, the growth is very significant.And finally, it decelerates slightly in the 2010s.It kind of looks like either a logistical or exponential function.So, it requires a special estimation method of the non-linear regression procedure.For example, if we assume that the model for these data points are exponential functions,such as yˆ= θ_0 + θ_1 〖θ_2〗^x, our job is to estimate the parameters of the model,i.e. θs, and use the fitted model to predict GDP for unknown or future cases.In fact, many different regressions exist that can be used to fit whatever the dataset looks like.You can see a quadratic and cubic regression lines here, and it can go on and on to infinitedegrees.In essence, we can call all of these "polynomial regression", where the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.With many types of regression to choose from, there’s a good chance that one will fit your dataset well.Remember, it’s important to pick a regression that fits the data the best.So, what is polynomial Regression?Polynomial regression fits a curved line to your data.A simple example of polynomial, with degree 3, is shown as yˆ= θ_0 + θ_1x + θ_2x^2+ θ_3x^3 or to the power of 3, where θs are parameters to be estimated that makes the model fit perfectly to the underlying data.Though the relationship between x and y is non-linear here, and polynomial regression can fit them, a polynomial regression model can still be expressed as linear regression.I know it's a bit confusing, but let’s look at an example.Given the 3rd degree polynomial equation, by defining x_1 = x and x_2 = x^2 or x to the power of 2 and so on,the model is converted to a simple linear regression with new variables, as yˆ= θ_0+θ_1x_1 + θ_2x_2 + θ_3x_3. This model is linear in the parameters to be estimated, right?Therefore, this polynomial regression is considered to be a special case of traditional multiplelinear regression.So, you can use the same mechanism as linear regression to solve such a problem.Therefore, polynomial regression models CAN fit using the model of least squares.Least squares is a method for estimating the unknown parameters in a linear regression model, by minimizing the sum of the squares of the differences between the observed dependent variable in the given dataset and those predicted by the linear function.So, what is “non-linear regression” exactly?First, non-linear regression is a method to model a nonlinear relationship between the dependent variable and a set of independent variables.Second, for a model to be considered non-linear, yˆmust be a nonlinear function of the parameters θ, not necessarily the features x.When it comes to nonlinear equation, it can be the shape of exponential, logarithmic,and logistic, or many other types.As you can see, in all of these equations, the change of yˆdepends on changes in the parameters θ, not necessarily on x only.That is, in non-linear regression, a model is non-linear by parameters.In contrast to linear regression, we cannot use the ordinary "least squares" method to fit the data in nonlinear regression, and in general, estimation of the parameters is not easy.Let me answer two important questions here: First, “How can I know if a problem is linear or nonlinear in an easy way?”To answer this question, we have to do two things:The first is to visually figure out if the relation is linear or nonlinear.It's best to plot bivariate plots of output variables with each input variable.Also, you can calculate the correlation coefficient between independent and dependent

variables,and if for all variables it is 0.7 or higher there is a linear tendency, and, thus, it's not appropriate to fit a nonlinear regression.The second thing we have to do is to use non-linear regression instead of linear regression when we cannot accurately model the relationship with linear parameters.The second important questions is, “How should I model my data, if it displays non-linear on a scatter plot?”Well, to address this, you have to use either a polynomial regression, use a non-linear regression model, or "transform" your data, which is not in scope for this course.Thanks for watching.

Building Decision Trees

Hello, and welcome!
In this video, we’ll be covering the process of building decision trees.
So let’s get started!
Consider the drug dataset again.
The question is, “How do we build the decision tree based on that dataset?”
Decision trees are built using recursive partitioning to classify the data.
Let’s say we have 14 patients in our dataset.
The algorithm chooses the most predictive feature to split the data on.
What is important in making a decision tree, is to determine “which attribute is the best, or more predictive, to split data based on the feature.”
Let’s say we pick “Cholesterol” as the first attribute to split data.
It will split our data into 2 branches.
As you can see, if the patient has high “Cholesterol,” we cannot say with high confidence that Drug B
might be suitable for him.
Also, if the Patient’s “Cholesterol” is normal, we still don’t have sufficient evidence or information to determine if either Drug A or Drug B is, in fact, suitable.
It is a sample of bad attribute selection for splitting data.
So, let’s try another attribute.
Again, we have our 14 cases.
This time, we pick the “sex” attribute of patients.
It will split our data into 2 branches, Male and Female.
As you can see, if the patient is Female, we can say Drug B might be suitable for her with high certainty.
But, if the patient is Male, we don’t have sufficient evidence or information to determine if Drug A or Drug B is suitable.
However, it is still a better choice in comparison with the “Cholesterol” attribute, because the result in the nodes are more pure.
It means, nodes that are either mostly Drug A or Drug B.
So, we can say the “Sex” attribute is more significant than “Cholesterol,” or in other words, it’s more predictive than the other attributes.
Indeed, “predictiveness” is based on decrease in “impurity” of nodes.

We're looking for the best feature to decrease the "impurity" of patients in the leaves, after splitting them up based on that feature.
So, the "Sex" feature is a good candidate in the following case, because it almost found the pure patients.
Let's go one step further.
For the Male patient branch, we again test other attributes to split the subtree.
We test "Cholesterol" again here.
As you can see, it results in even more pure leaves.
So, we can easily make a decision here.
For example, if a patient is "Male", and his "Cholesterol" is "High", we can certainly prescribe Drug A, but if it is "Normal", we can prescribe Drug B with high confidence.
As you might notice, the choice of attribute to split data is very important, and it is all about "purity" of the leaves after the split.
A node in the tree is considered "pure" if, in 100% of the cases, the nodes fall into a specific category of the target field.
In fact, the method uses recursive partitioning to split the training records into segments by minimizing the "impurity" at each step.
"Impurity" of nodes is calculated by "Entropy" of data in the node.
So, what is "Entropy"?
Entropy is the amount of information disorder, or the amount of randomness in the data.
The entropy in the node depends on how much random data is in that node and is calculated for each node.
In decision trees, we're looking for trees that have the smallest entropy in their nodes.
The entropy is used to calculate the homogeneity of the samples in that node.
If the samples are completely homogeneous the entropy is zero and if the samples are equally divided, it has an entropy of one.
This means, if all the data in a node are either Drug A or Drug B, then the entropy is zero, but if half of the data are Drug A and other half are B, then the entropy is one.
You can easily calculate the entropy of a node using the frequency table of the attribute through the Entropy formula, where P is for the proportion or ratio of a category, such as Drug A or B. Please remember, though, that you don't have to calculate these, as it's easily calculated by the libraries or packages that you use.
As an example, let's calculate the entropy of the dataset before splitting it.
We have 9 occurrences of Drug B and 5 of Drug A.
You can embed these numbers into the Entropy formula to calculate the impurity of the target attribute before splitting it.
In this case, it is 0.94.
So, what is entropy after splitting?
Now we can test different attributes to find the one with the most "predictiveness," which results in two more pure branches.
Let's first select the "Cholesterol" of the patient and see how the data gets split, based on its values.

For example, when it is “normal,” we have 6 for Drug B, and 2 for Drug A.
We can calculate the Entropy of this node based on the distribution of drug A and B, which is 0.8 in this case.
But, when Cholesterol is “High,” the data is split into 3 for drug B and 3 for drug A.
Calculating its entropy, we can see it would be 1.0.
We should go through all the attributes and
calculate the “Entropy” after the split, and then chose the best attribute.
Ok, let’s try another field.
Let’s choose the Sex attribute for the next check.
As you can see, when we use the Sex attribute to split the data, when its value is “Female,” we have 3 patients that responded to Drug B, and 4 patients that responded to Drug A.
The entropy for this node is 0.98 which is not very promising.
However, on other side of the branch, when the value of the Sex attribute is Male, the result is more pure with 6 for Drug B and only 1 for Drug A.
The entropy for this group is 0.59.
Now, the question is, between the Cholesterol and Sex attributes, which one is a better choice?
Which one is better as the first attribute to divide the dataset into 2 branches?
Or, in other words, which attribute results in more pure nodes for our drugs?
Or, in which tree, do we have less entropy after splitting rather than before splitting?
The “Sex” attribute with entropy of 0.98 and 0.59, or the “Cholesterol” attribute with entropy of 0.81 and 1.0 in its branches?
The answer is, “The tree with the higher information gain after splitting."
So, what is information gain?
Information gain is the information that can increase the level of certainty after splitting.
It is the entropy of a tree before the split minus the weighted entropy after the split by an attribute.
We can think of information gain and entropy as opposites.
As entropy, or the amount of randomness, decreases, the information gain, or amount of certainty,
increases, and vice-versa.
So, constructing a decision tree is all about finding attributes that return the highest information gain.
Let’s see how “information gain” is calculated for the Sex attribute.
As mentioned, the information gain is the entropy of the tree before the split, minus the weighted entropy after the split.
The entropy of the tree before the split is 0.94.
The portion of Female patients is 7 out of 14, and its entropy is 0.985.
Also, the portion of men is 7 out of 14, and the entropy of the Male node is 0.592.
The result of a square bracket here is the weighted entropy after the split.
So, the information gain of the tree if we use the “Sex” attribute to split the dataset is 0.151.
As you can see, we will consider the entropy over the distribution of samples falling under each leaf node, and we’ll take a weighted average of that entropy – weighted by the

proportion of samples falling under that leaf.
We can calculate the information gain of the tree if we use “Cholesterol” as well.
It is 0.48.
Now, the question is, “Which attribute is more suitable?”
Well, as mentioned, the tree with the higher information gain after splitting.
This means the “Sex” attribute.
So, we select the “Sex” attribute as the first splitter.
Now, what is the next attribute after branching by the “Sex” attribute?
Well, as you can guess, we should repeat the process for each branch, and test each of the other attributes to continue to reach the most pure leaves.
This is the way that you build a decision tree!
Thanks for watching!

Evaluation Metrics in Classification

Hello, and welcome! In this video, we’ll be covering evaluation
metrics for classifiers. So let’s get started.
Evaluation metrics explain the performance of a model.
Let’s talk more about the model evaluation metrics that are used for classification.
Imagine that we have an historical dataset which shows the customer churn for a telecommunication
company. We have trained the model, and now we want
to calculate its accuracy using the test set. We pass the test set to our model, and we find the predicted labels. Now the question is, “How accurate is this
model?” Basically, we compare the actual values in
the test set with the values predicted by the model, to calculate the accuracy of the model. Evaluation metrics provide a key role in the
development of a model, as they provide insight to areas that might require improvement.
There are different model evaluation metrics but we just talk about three of them here, specifically: Jaccard index, F1-score, and Log Loss.
Let’s first look at one of the simplest accuracy measurements, the Jaccard index -- also known as the Jaccard similarity coefficient. Let’s say y shows the true labels of the churn dataset. And yˆ shows the predicted values by our
classifier. Then we can define Jaccard as the size of
the intersection divided by the size of the union of two label sets.
For example, for a test set of size 10, with 8 correct predictions, or 8 intersections, the accuracy by the Jaccard index would be 0.66.
If the entire set of predicted labels for a sample strictly matches with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.
Another way of looking at accuracy of classifiers is to look at a confusion matrix.
For example, let’s assume that our test set has only 40 rows.
This matrix shows the corrected and wrong predictions, in comparison with the actual

labels. Each confusion matrix row shows the Actual/True
labels in the test set, and the columns show the predicted labels by classifier.
Look at the first row. The first row is for customers whose actual
churn value in the test set is 1. As you can calculate, out of 40 customers,
the churn value of 15 of them is 1. And out of these 15, the classifier correctly
predicted 6 of them as 1, and 9 of them as 0.
This means that for 6 customers, the actual churn value was 1, in the test set, and the
classifier also correctly predicted those as 1.
However, while the actual label of 9 customers was 1, the classifier predicted those as 0,
which is not very good. We can consider this as an error of the model
for the first row. What about the customers with a churn value
0? Let's look at the second row.
It looks like there were 25 customers whose churn value was 0.
The classifier correctly predicted 24 of them as 0, and one of them wrongly predicted as
1. So, it has done a good job in predicting the
customers with a churn value of 0. A good thing about the confusion matrix is
that it shows the model's ability to correctly predict or separate the classes.
In the specific case of a binary classifier, such as this example, we can interpret these
numbers as the count of true positives, false negatives, true negatives, and false positives.
Based on the count of each section, we can calculate the precision and recall of each
label. Precision is a measure of the accuracy, provided
that a class label has been predicted. It is defined by: precision = True Positive
/ (True Positive + False Positive). And Recall is the true positive rate.
It is defined as: Recall = True Positive / (True Positive + False Negative).
So, we can calculate the precision and recall of each class.
Now we're in the position to calculate the F1 scores for each label, based on the precision
and recall of that label. The F1 score is the harmonic average of the
precision and recall, where an F1 score reaches its best value at 1 (which represents perfect
precision and recall) and its worst at 0. It is a good way to show that a classifier
has a good value for both recall and precision. It is defined using the F1-score equation.
For example, the F1-score for class 0 (i.e. churn=0), is 0.83, and the F1-score for class
1 (i.e. churn=1), is 0.55. And finally, we can tell the average accuracy
for this classifier is the average of the F1-score for both labels, which is 0.72 in
our case. Please notice that both Jaccard and F1-score
can be used for multi-class classifiers as well, which is out of scope for this course.
Now let's look at another accuracy metric for classifiers.
Sometimes, the output of a classifier is the probability of a class label, instead of the
label. For example, in logistic regression, the output
can be the probability of customer churn, i.e., yes (or equals to 1). This probability
is a value between 0 and 1. Logarithmic loss (also known as Log loss)
measures the performance of a classifier where the predicted output is a probability value
between 0 and 1. So, for example, predicting a probability
of 0.13 when the actual label is 1, would be bad and would result in a high log loss.
We can calculate the log loss for each row using the log loss equation, which measures

how far each prediction is, from the actual label.
Then, we calculate the average log loss across all rows of the test set.
It is obvious that more ideal classifiers have progressively smaller values of log loss.
So, the classifier with the lower log loss has better accuracy.
Thanks for watching!

Intro to Decision Trees

Hello, and welcome!
In this video, we're going to introduce and examine decision trees.
So let's get started
What exactly is a decision tree?
How do we use them to help us classify?
How can I grow my own decision tree?
These may be some of the questions that you have in mind from hearing the term, decision tree.
Hopefully, you'll soon be able to answer these questions, and many more, by watching this video!
Imagine that you're a medical researcher compiling data for a study.
You've already collected data about a set of patients, all of whom suffered from the same illness.
During their course of treatment, each patient responded to one of two medications; we'll call them Drug A and Drug B. Part of your job is to build a model to find out which drug might be appropriate for a future patient with the same illness.
The feature sets of this dataset are Age, Gender, Blood Pressure, and Cholesterol of our group of patients, and the target is the drug that each patient responded to.
It is a sample of binary classifiers, and you can use the training part of the dataset to build a decision tree, and then, use it to predict the class of an unknown patient … in essence, to come up with a decision on which drug to prescribe to a new patient.
Let's see how a decision tree is built for this dataset.
Decision trees are built by splitting the training set into distinct nodes, where one node contains all of, or most of, one category of the data.
If we look at the diagram here, we can see that it's a patient classifier.
So, as mentioned, we want to prescribe a drug to a new patient, but the decision to choose drug A or B, will be influenced by the patient's situation.
We start with the Age, which can be Young, Middle-aged, or Senior.
If the patient is Middle-aged, then we'll definitely go for Drug B.
On the other hand, if he is a Young or a Senior patient, we'll need more details to help us determine which drug to prescribe.
The additional decision variables can be things such as Cholesterol levels, Gender or Blood Pressure.
For example, if the patient is Female, then we will recommend Drug A, but if the patient is Male, then we'll go for Drug B. As you can see, decision trees are about testing

an attribute and branching the cases, based on the result of the test.
Each internal node corresponds to a test.
And each branch corresponds to a result of the test.
And each leaf node assigns a patient to a class.
Now the question is how can we build such a decision tree?
Here is the way that a decision tree is built.
A decision tree can be constructed by considering the attributes one by one.
First, choose an attribute from our dataset.
Calculate the significance of the attribute in the splitting of the data.
In the next video, we will explain how to calculate the significance of an attribute, to see if it's an effective attribute or not.
Next, split the data based on the value of the best attribute.
Then, go to each branch and repeat it for the rest of the attributes.
After building this tree, you can use it to predict the class of unknown cases or, in our case, the proper Drug for a new patient based on his/her characterestics.
This concludes this video.
Thanks for watching!

Intro to Logistic Regression

Hello, and welcome!
In this video we'll learn a machine learning method, called logistic regression, which is used for classification.
In examining this method, we'll specifically answer these three questions:
- What is logistic regression?
- What kind of problems can be solved by logistic regression?
- And, in which situations do we use logistic regression?
So, let's get started.
Logistic regression is a statistical and machine learning technique for classifying records of a dataset, based on the values of the input fields.
Let's say we have a telecommunication dataset that we'd would like to analyze, in order to understand which customers might leave us next month.
This is historical customer data where each row represents one customer.
Imagine that you're an analyst at this company and you have to find out who is leaving and why.
You'll use the dataset to build a model based on historical records and use it to predict the future churn within the customer group.
The data set includes information about: - Services that each customer has signed up for, - Customer account information,
- Demographic information about customers, like gender and age-range,
- And also Customers who've left the company within the last month.

The column is called Churn.
We can use logistic regression to build a model for predicting customer churn, using the given features.
In logistic regression, we use one or more independent variables such as tenure, age and income to predict an outcome, such as churn, which we call a dependent variable, representing whether or not customers will stop using the service.
Logistic regression is analogous to linear regression but tries to predict a categorical or discrete target field instead of a numeric one.
In linear regression, we might try to predict a continuous value of variables, such as the price of a house, blood pressure of patient, or fuel consumption of a car.
But, in logistic regression, we predict a variable which is binary, such as, Yes/No, TRUE/FALSE, successful or Not successful, pregnant/Not pregnant, and so on, all of which can all be coded as 0 or 1.
In logistic regression, dependent variables should be continuous; if categorical, they should be dummy or indicator-coded.
This means we have to transform them to some continuous value.
Please note that logistic regression can be used for both binary classification and multiclass classification, but for simplicity, in this video, we'll focus on binary classification.
Let's examine some applications of logistic regression before we explain how they work.
As mentioned, logistic regression is a type of classification algorithm, so it can be used in different situations, for example: - To predict the probability of a person having a heart attack within a specified time period, based on our knowledge of the person's age, sex, and body mass index.
- Or to predict the chance of mortality in an injured patient, or to predict whether a patient has a given disease, such as diabetes, based on observed characteristics of that patient, such as weight, height, blood pressure, and results of various blood tests, and so on.
- In a marketing context, we can use it to predict the likelihood of a customer purchasing a product or halting a subscription, as we've done in our churn example.
- We can also use logistic regression to predict the probability of failure of a given process, system, or product.
- We can even use it to predict the likelihood of a homeowner defaulting on a mortgage.
These are all good examples of problems that can be solved using logistic regression.
Notice that in all of these examples, not only do we predict the class of each case, we also measure the probability of a case belonging to a specific class.
There are different machine algorithms which can classify or estimate a variable.
The question is, when should we use Logistic Regression?
Here are four situations in which Logistic regression is a good candidate:
First, when the target field in your data is categorical, or specifically, is binary, such as 0/1, yes/no, churn or no churn, positive/negative, and so on.
Second, you need the probability of your prediction, for example, if you want to know what the probability is, of a customer buying a product.
Logistic regression returns a probability score between 0 and 1 for a given sample of data.

In fact, logistic regressing predicts the probability of that sample, and we map the cases to a discrete class based on that probability.
Third, if your data is linearly separable.
The decision boundary of logistic regression is a line or a plane or a hyper-plane.
A classifier will classify all the points on one side of the decision boundary as belonging to one class and all those on the other side as belonging to the other class.
For example, if we have just two features (and are not applying any polynomial processing), we can obtain an inequality like $\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$, which is a half-plane, easily plottable.
Please note that in using logistic regression, we can also achieve a complex decision boundary using polynomial processing as well, which is out of scope here.
You'll get more insight from decision boundaries when you understand how logistic regression works.
Fourth, you need to understand the impact of a feature.
You can select the best features based on the statistical significance of the logistic regression model coefficients or parameters.
That is, after finding the optimum parameters, a feature x with the weight θ_1 close to 0, has a smaller effect on the prediction, than features with large absolute values of θ_1.
Indeed, it allows us to understand the impact an independent variable has on the dependent variable while controlling other independent variables.
Let's look at our dataset again.
We define the independent variables as X, and dependent variable as Y.
Notice that, for the sake of simplicity, we can code the target or dependent values to 0 or 1.
The goal of logistic regression is to build a model to predict the class of each sample (which in this case is a customer) as well as the probability of each sample belonging to a class.
Given that, let's start to formalize the problem.
X is our dataset, in the space of real numbers of m by n, that is, of m dimensions or features and n records.
And y is the class that we want to predict, which can be either zero or one.
Ideally, a logistic regression model, so called $\hat{y}$ (y-hat), can predict that the class of a customer is 1, given its features x.
It can also be shown quite easily, that the probability of a customer being in class 0 can be calculated as 1 minus the probability that the class of the customer is 1.
Thanks for watching this video.

K-Nearest Neighbors

Hello and welcome! In this video, we'll be covering the k-nearest neighbors algorithm. So let's get started. Imagine that a telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups.

If demographic data can be used to predict group membership, the company can customize offers for individual prospective customers.
This is a classification problem.
That is, given the dataset, with predefined labels, we need to build a model to be used to predict the class of a new or unknown case.
The example focuses on using demographic data, such as region, age, and marital status, to predict usage patterns.
The target field, called custcat, has four possible values that correspond to the four customer groups, as follows: Basic Service, E-Service, Plus Service, and Total Service.
Our objective is to build a classifier, for example using the rows 0 to 7, to predict the class of row 8.
We will use a specific type of classification called K-nearest neighbor.
Just for sake of demonstration, let's use only two fields as predictors - specifically, Age and Income, and then plot the customers based on their group membership.
Now, let's say that we have a new customer, for example, record number 8 with a known age and income.
How can we find the class of this customer?
Can we find one of the closest cases and assign the same class label to our new customer?
Can we also say that the class of our new customer is most probably group 4 (i.e. total service) because its nearest neighbor is also of class 4?
Yes, we can.
In fact, it is the first-nearest neighbor.
Now, the question is, "To what extent can we trust our judgment, which is based on the first nearest neighbor?"
It might be a poor judgment, especially if the first nearest neighbor is a very specific case, or an outlier, correct?
Now, let's look at our scatter plot again.
Rather than choose the first nearest neighbor, what if we chose the five nearest neighbors, and did a majority vote among them to define the class of our new customer?
In this case, we'd see that three out of five nearest neighbors tell us to go for class 3, which is "Plus service."
Doesn't this make more sense?
Yes, in fact, it does!
In this case, the value of K in the k-nearest neighbors algorithm is 5.
This example highlights the intuition behind the k-nearest neighbors algorithm.
Now, let's define the k-nearest neighbors.
The k-nearest-neighbors algorithm is a classification algorithm that takes a bunch of labelled points
and uses them to learn how to label other points.
This algorithm classifies cases based on their similarity to other cases.
In k-nearest neighbors, data points that are near each other are said to be "neighbors."
K-nearest neighbors is based on this paradigm: "Similar cases with the same class labels are near each other."
Thus, the distance between two cases is a measure of their dissimilarity.

There are different ways to calculate the similarity, or conversely, the distance or dissimilarity of two data points.
For example, this can be done using Euclidean distance.

Now, let's see how the k-nearest neighbors algorithm actually works. In a classification problem, the k-nearest neighbors algorithm works as follows:

1. Pick a value for K.
2. Calculate the distance from the new case (holdout from each of the cases in the dataset).
3. Search for the K observations in the training data that are 'nearest' to the measurements of the unknown data point.
4, predict the response of the unknown data point using the most popular response value from the K nearest neighbors.

There are two parts in this algorithm that might be a bit confusing.

1. First, how to select the correct K;
2. Second, how to compute the similarity between cases, for example, among customers?

Let's first start with second concern, that is, how can we calculate the similarity between two data points? Assume that we have two customers, customer 1 and customer 2. And, for a moment, assume that these 2 customers have only one feature, Age. We can easily use a specific type of Minkowski distance to calculate the distance of these 2 customers.
It is indeed, the Euclidian distance.
Distance of x1 from x2 is root of 34 minus 30 to power of 2, which is 4.
What about if we have more than one feature, for example Age and Income?
If we have income and age for each customer, we can still use the same formula, but this time, we're using it in a 2-dimensional space.
We can also use the same distance matrix for multi-dimensional vectors.
Of course, we have to normalize our feature set to get the accurate dissimilarity measure.
There are other dissimilarity measures as well that can be used for this purpose but, as mentioned, it is highly dependent on data type and also the domain that classification is done for it.
As mentioned, K in k-nearest neighbors, is the number of nearest neighbors to examine.
It is supposed to be specified by the user.
So, how do we choose the right K?
Assume that we want to find the class of the customer noted as question mark on the chart.
What happens if we choose a very low value of K, let's say, k=1?
The first nearest point would be Blue, which is class 1.
This would be a bad prediction, since more of the points around it are Magenta, or class 4.
In fact, since its nearest neighbor is Blue, we can say that we captured the noise in the data, or we chose one of the points that was an anomaly in the data.
A low value of K causes a highly complex model as well, which might result in over-fitting of the model.
It means the prediction process is not generalized enough to be used for out-of-sample cases.

Out-of-sample data is data that is outside of the dataset used to train the model.
In other words, it cannot be trusted to be used for prediction of unknown samples.
It's important to remember that over-fitting is bad, as we want a general model that works for any data, not just the data used for training.
Now, on the opposite side of the spectrum, if we choose a very high value of K, such as K=20, then the model becomes overly generalized.
So, how we can find the best value for K?
The general solution is to reserve a part of your data for testing the accuracy of the model.
Once you've done so, choose k =1, and then use the training part for modeling, and calculate the accuracy of prediction using all samples in your test set.
Repeat this process, increasing the k, and see which k is best for your model.
For example, in our case, k=4 will give us the best accuracy.
Nearest neighbors analysis can also be used to compute values for a continuous target.
In this situation, the average or median target value of the nearest neighbors is used to obtain the predicted value for the new case.
For example, assume that you are predicting the price of a home based on its feature set, such as number of rooms, square footage, the year it was built, and so on.
You can easily find the three nearest neighbor houses, of course -- not only based on distance, but also based on all the attributes, and then predict the price of the house, as the median of neighbors.

Learning Objectives

In this lesson you will learn about:

- K-Nearest Neighbors
- Decision Trees
- Support Vector Machines
- Logistic Regression

NOTE: The **Logistic Regression - Training** video is optional and the content from this video is not included in the final exam.

Intro to Classification

Hello!
In this video, we'll give you an introduction to Classification.
So let's get started.
In Machine Learning, classification is a supervised learning approach, which can be thought of

as a means of categorizing or "classifying" some unknown items into a discrete set of "classes."
Classification attempts to learn the relationship between a set of feature variables and a target variable of interest.
The target attribute in classification is a categorical variable with discrete values.
So, how does classification and classifiers work?
Given a set of training data points, along with the target labels, classification determines the class label for an unlabeled test case.
Let's explain this with an example.
A good sample of classification is the loan default prediction.
Suppose a bank is concerned about the potential for loans not to be repaid.
If previous loan default data can be used to predict which customers are likely to have problems repaying loans, these "bad risk" customers can either have their loan application declined or offered alternative products.
The goal of a loan default predictor is to use existing loan default data, which is information about the customers (such as age, income, education, etc.), to build a classifier, pass a new customer or potential future defaulter to the model, and then label it (i.e. the data points) as "Defaulter" or "Not Defaulter", or for example, 0 or 1.
This is how a classifier predicts an unlabeled test case.
Please notice that this specific example was about a binary classifier with two values.
We can also build classifier models for both binary classification and multi-class classification.
For example, imagine that you collected data about a set of patients, all of whom suffered from the same illness.
During their course of treatment, each patient responded to one of three medications.
You can use this labeled dataset, with a classification algorithm, to build a classification model.
Then you can use it to find out which drug might be appropriate for a future patient with the same illness.
As you can see, it is a sample of multi-class classification.
Classification has different business use cases as well, for example:
To predict the category to which a customer belongs;
For Churn detection, where we predict whether a customer switches to another provider or brand; Or to predict whether or not a customer responds to a particular advertising campaign.
Data classification has several applications in a wide variety of industries.
Essentially, many problems can be expressed as associations between feature and target variables, especially when labeled data is available.
This provides a broad range of applicability for classification.
For example, classification can be used for email filtering, speech recognition, handwriting recognition, bio-metric identification, document classification, and much more.
Here we have the types of classification algorithms in machine learning.
They include: Decision Trees, Naïve Bayes, Linear Discriminant Analysis, K-nearest neighbor, Logistic regression, Neural Networks, and Support Vector Machines.
There are many types of classification algorithms.
We will only cover a few in this course.

Thanks for watching.

Logistic Regression vs Linear Regression

Hello, and welcome! In this video we will learn the difference
between linear regression and logistic regression. We go over linear regression and see why it
cannot be used properly for some binary classification problems.
We also look at the Sigmoid function, which is the main part of logistic regression.
Let's start.
Let's look at the telecommunication dataset again.
The goal of logistic regression is to build a model to predict the class of each customer,
and also the probability of each sample belonging to a class.
Ideally, we want to build a model, y^, that can estimate that the class of a customer
is 1, given its features, x. I want to emphasize that y is the "labels
vector," also called "actual values" that we would like to predict, and y^ is the
vector of the predicted values by our model. Mapping the class labels to integer numbers,
can we use linear regression to solve this problem?
First, let's recall how linear regression works to better understand logistic regression.
Forget about the churn prediction for a minute, and assume our goal is to predict the income
of customers in the dataset. This means that instead of predicting churn,
which is a categorical value, let's predict income, which is a continuous value.
So, how can we do this? Let's select an independent variable, such
as customer age, and predict a dependent variable, such as income.
Of course we can have more features, but for the sake of simplicity, let's just take
one feature here. We can plot it, and show age as an independent
variable, and income as the target value we would like to predict.
With linear regression, you can fit a line or polynomial through the data.
We can find this line through the training of our model, or calculating it mathematically
based on the sample sets. We'll say this is a straight line through
the sample set. This line has an equation shown as a+b_x1.
Now, use this line to predict the continuous value y, that is, use this line to predict
the income of an unknown customer based on his or her age.
And it is done.
What if we want to predict churn? Can we use the same technique to predict a
categorical field such as churn? OK, let's see.
Say we're given data on customer churn and our goal this time is to predict the churn
of customers based on their age. We have a feature, age denoted as x1, and
a categorical feature, churn, with two classes: churn is yes and churn is no.
As mentioned, we can map yes and no to integer values, 0 and 1.
How can we model it now? Well, graphically, we could represent our
data with a scatter plot. But this time we have only 2 values for the
y axis. In this plot, class zero is denoted in red

and class one is denoted in blue. Our goal here is to make a model based on existing data, to predict if a new customer is red or blue.
Let's do the same technique that we used for linear regression here to see if we can solve the problem for a categorical attribute, such as churn.
With linear regression, you again can fit a polynomial through the data, which is shown traditionally as a + b_x. This polynomial can also be shown traditionally as θ_0 + θ_1 x_1. This line has 2 parameters, which are shown with vector θ, where the values of the vector are θ_0 and θ_1.
We can also show the equation of this line formally as θ^T X.
And generally, we can show the equation for a multi-dimensional space as θ^T X, where θ is the parameters of the line in 2-dimensional space, or parameters of a plane in 3-dimensional space, and so on. As θ is a vector of parameters, and is supposed to be multiplied by X, it is shown conventionally as T^θ.
θ is also called the "weight vector" or "confidences of the equation" -- with both of these terms used interchangeably. And X is the feature set, which represents a customer. Anyway, given a dataset, all the feature sets X, θ parameters, can be calculated through an optimization algorithm or mathematically, which results in the equation of the fitting line.
For example, the parameters of this line are -1 and 0.1 and the equation for the line is -1 + 0.1 x1.
Now, we can use this regression line to predict the churn of the new customer.
For example, for our customer, or let's say a data point with x value of age=13, we can plug the value into the line formula, and the y value is calculated and returns a number. For instance, for p1 point we have:
θ^T X = -1 + 0.1 x x_1 = -1 + 0.1 x 13
= 0.3 We can show it on our graph.
Now we can define a threshold here, for example at 0.5, to define the class.
So, we write a rule here for our model, y^, which allows us to separate class 0 from class 1.
If the value of θ^T X is less than 0.5, then the class is 0, otherwise, if the value of θ^T X is more than 0.5, then the class is 1.
And because our customer's y value is less than the threshold, we can say it belongs to class 0, based on our model. But there is one problem here; what is the probability that this customer belongs to class 0?
As you can see, it's not the best model to solve this problem.
Also, there are some other issues, which verify that linear regression is not the proper method for classification problems.
So, as mentioned, if we use the regression line to calculate the class of a point, it always returns a number, such as 3, or -2, and so on.
Then, we should use a threshold, for example, 0.5, to assign that point to either class of 0 or 1. This threshold works as a step function that outputs 0 or 1, regardless of how big or small, positive or negative, the input is.
So, using the threshold, we can find the class of a record.

Notice that in the step function, no matter how big the value is, as long as it's greater than 0.5, it simply equals 1. And vice versa, regardless of how small the value y is, the output would be zero if it is less than 0.5. In other words, there is no difference between a customer who has a value of one or 1000; the outcome would be 1.
Instead of having this step function, wouldn't it be nice if we had a smoother line - one that would project these values between zero and one?
Indeed, the existing method does not really give us the probability of a customer belonging to a class, which is very desirable. We need a method that can give us the probability of falling in a class as well. So, what is the scientific solution here?
Well, if instead of using θ^T X we use a
specific function called sigmoid, then, sigmoid of θ^T X gives us the probability of a point belonging to a class, instead of the value of y directly.
I'll explain this sigmoid function in a second, but for now, please accept that it will do the trick. Instead of calculating the value of θ^T X
directly, it returns the probability that a θ^T X is very big or very small.
It always returns a value between 0 and 1 depending on how large the θ^T X actually is. Now, our model is σ(θ^T X), which represents
the probability that the output is 1, given x.
Now the question is, "What is the sigmoid function?"
Let me explain in detail what sigmoid really is.
The sigmoid function, also called the logistic function, resembles the step function and is used by the following expression in the logistic regression.
The sigmoid function looks a bit complicated at first, but don't worry about remembering this equation. It'll make sense to you after working with it.
Notice that in the sigmoid equation, when
θ^T X gets very big, the [e]^(-θ^T X) in the denominator of the fraction becomes almost zero, and the value of the sigmoid function gets closer to 1.
If θ^T X is very small, the sigmoid function gets closer to zero.
Depicting on the in sigmoid plot, when θ^T X , gets bigger, the value of the sigmoid function gets closer to 1, and also, if the θ^T X is very small, the sigmoid function gets closer to zero. So, the sigmoid function's output is always
between 0 and 1, which makes it proper to interpret the results as probabilities.
It is obvious that when the outcome of the sigmoid function gets closer to 1, the probability of y=1, given x, goes up, and in contrast, when the sigmoid value is closer to zero, the probability of y=1, given x, is very small.
So what is the output of our model when we use the sigmoid function?
In logistic regression, we model the probability that an input (X) belongs to the default class (Y=1), and we can write this formally as, P(Y=1|X).
We can also write P(y=0|X) = 1 -P(y=1|x). For example, the probability of a customer staying with the company can be shown as probability of churn equals 1 given a customer's income
and age, which can be, for instance, 0.8. And the probability of churn is 0, for the same customer, given a customer's income and age can be calculated as 1-0.8=0.2.
So, now, our job is to train the model to set its parameter values in such a way that

our model is a good estimate of P(y=1|x). In fact, this is what a good classifier model built by logistic regression is supposed to do for us.
Also, it should be a good estimate of P(y=0|x) that can be shown as 1-σ(θ^T X).
Now, the question is: "How we can achieve this?"
We can find θ through the training process, so let's see what the training process is.
Step 1. Initialize θ vector with random values, as with most machine learning algorithms, for example −1 or 2.
Step 2. Calculate the model output, which is σ(θ^T X), for a sample customer in your training set. X in θ^T X is the feature vector values -- for
example, the age and income of the customer, for instance [2,5].
And θ is the confidence or weight that you've set in the previous step.
The output of this equation is the prediction value … in other words, the probability that the customer belongs to class 1.
Step 3. Compare the output of our model, y^, which could be a value of, let's say, 0.7, with the actual label of the customer, which is for example, 1 for churn.
Then, record the difference as our model's error for this customer, which would be 1-0.7, which of course equals 0.3. This is the error for only one customer out of all the customers in the training set.
Step. 4. Calculate the error for all customers as we did in the previous steps, and add up these errors. The total error is the cost of your model,
and is calculated by the model's cost function. The cost function, by the way, basically represents
how to calculate the error of the model, which is the difference between the actual and the model's predicted values. So, the cost shows how poorly the model is
estimating the customer's labels. Therefore, the lower the cost, the better
the model is at estimating the customer's labels correctly.
And so, what we want to do is to try to minimize this cost.
Step 5. But, because the initial values for θ were chosen randomly, it's very likely that the cost function is very high. So, we change the θ in such a way to hopefully reduce the total cost.
Step 6. After changing the values of θ, we go back to step 2.
Then we start another iteration, and calculate the cost of the model again.
And we keep doing those steps over and over, changing the values of θ each time, until the cost is low enough. So, this brings up two questions: first, "How
can we change the values of θ so that the cost is reduced across iterations?"
And second, "When should we stop the iterations?" There are different ways to change the values
of θ, but one of the most popular ways is gradient descent.
Also, there are various ways to stop iterations, but essentially you stop training by calculating the accuracy of your model, and stop it when it's satisfactory.
Thanks for watching this video!

Support Vector Machines

Hello, and welcome!
In this video we will learn a machine learning method called Support Vector Machine (or SVM), which is used for classification.
So let's get started.
Imagine that you've obtained a dataset containing characteristics of thousands of human cell samples extracted from patients who were believed to be at risk of developing cancer.
Analysis of the original data showed that many of the characteristics differed significantly between benign and malignant samples.
You can use the values of these cell characteristics in samples from other patients to give an early indication of whether a new sample might be benign or malignant.
You can use support vector machine, or SVM, as a classifier, to train your model to understand patterns within the data, that might show benign or malignant cells.
Once the model has been trained, it can be used to predict your new or unknown cell with rather high accuracy.
Now, let me give you a formal definition of SVM.
A Support Vector Machine is a supervised algorithm that can classify cases by finding a separator.
SVM works by first, mapping data to a high-dimensional feature space so that data points can be categorized,
even when the data are not otherwise linearly separable.
Then, a separator is estimated for the data.
The data should be transformed in such a way that a separator could be drawn as a hyperplane.
For example, consider the following figure, which shows the distribution of a small set
of cells, only based on their Unit Size and Clump thickness.
As you can see, the data points fall into two different categories.
It represents a linearly, non-separable, dataset.
The two categories can be separated with a curve, but not a line.
That is, it represents a linearly, non-separable dataset, which is the case for most real-world datasets.
We can transfer this data to a higher dimensional space … for example, mapping it to a 3-dimensional
space.
After the transformation, the boundary between the two categories can be defined by a hyperplane.
As we are now in 3-dimensional space, the separator is shown as a plane.
This plane can be used to classify new or unknown cases.
Therefore, the SVM algorithm outputs an optimal hyperplane that categorizes new examples.
Now, there are two challenging questions to consider:
1) How do we transfer data in such a way that a separator could be drawn as a hyperplane?
and 2) How can we find the best/optimized hyperplane
separator after transformation?
Let's first look at "transforming data" to see how it works.
For the sake of simplicity, imagine that our dataset is 1-dimensional data, this means,
we have only one feature x.

As you can see, it is not linearly separable.
So, what we can do here?
Well, we can transfer it into a 2-dimensional space.
For example, your can increase the dimension of data by mapping x into a new space using a function, with outputs x and x-squared.
Now, the data is linearly separable, right?
Notice that, as we are in a two dimensional space, the hyperplane is a line dividing a plane into two parts where each class lays on either side.
Now we can use this line to classify new cases.
Basically, mapping data into a higher dimensional space is called kernelling.
The mathematical function used for the transformation is known as the kernel function, and can be
of different types, such as: Linear, Polynomial, Radial basis function (or RBF), and Sigmoid.
Each of these functions has its own characteristics, its pros and cons, and its equation, but the good news is that you don't need to know them, as most of them are already implemented in libraries of data science programming languages.
Also, as there's no easy way of knowing which function performs best with any given dataset, we usually choose different functions in turn and compare the results.
Now, we get to another question, specifically, "How do we find the right or optimized separator after transformation?"
Basically, SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown here.
As we're in a 2-dimensional space, you can think of the hyperplane as a line that linearly separates the blue points from the red points.
One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes.
So, the goal is to choose a hyperplane with as big a margin as possible.
Examples closest to the hyperplane are support vectors.
It is intuitive that only support vectors matter for achieving our goal; and thus, other training examples can be ignored.
We try to find the hyperplane in such a way that it has the maximum distance to support vectors.
Please note, that the hyperplane and boundary decision lines have their own equations.
So, finding the optimized hyperplane can be formalized using an equation which involves quite a bit more math, so I'm not going to go through it here, in detail.
That said, the hyperplane is learned from training data using an optimization procedure that maximizes the margin; and like many other problems, this optimization problem can also be solved by gradient descent, which is out of scope of this video.
Therefore, the output of the algorithm is the values 'w' and 'b' for the line.
You can make classifications using this estimated line.
It is enough to plug in input values into the line equation, then, you can calculate whether an unknown point is above or below the line.
If the equation returns a value greater than 0, then the point belongs to the first class, which is above the line, and vice versa.

The two main advantages of support vector machines are that they're accurate in high dimensional spaces; and, they use a subset of training points in the decision function (called support vectors), so it's also memory efficient.
The disadvantages of support vector machines include the fact that the algorithm is prone for over-fitting, if the number of features is much greater than the number of samples.
Also, SVMs do not directly provide probability estimates, which are desirable in most classification
problems.
And finally, SVMs are not very efficient computationally, if your dataset is very big, such as when you have more than one thousand rows.
And now, our final question is, "In which situation should I use SVM?"
Well, SVM is good for image analysis tasks, such as image classification and handwritten digit recognition.
Also SVM is very effective in text-mining tasks, particularly due to its effectiveness in dealing with high-dimensional data.
For example, it is used for detecting spam, text category assignment, and sentiment analysis.
Another application of SVM is in Gene Expression data classification, again, because of its power in high dimensional data classification.
SVM can also be used for other types of machine learning problems, such as regression, outlier detection, and clustering.
I'll leave it to you to explore more about these particular problems.
This concludes this video … Thanks for watching!

Intro to Clustering

Hello, and welcome! In this video, we'll give you a high level
introduction to clustering, its applications, and different types of clustering algorithms.
Let's get started.
Imagine that you have a customer dataset, and you need to apply customer segmentation on this historical data. Customer segmentation is the practice of partitioning
a customer base into groups of individuals that have similar characteristics.
It is a significant strategy as it allows a business to target specific groups of customers so as to more effectively allocate marketing resources.
For example, one group might contain customers who are high-profit and low-risk, that is, more likely to purchase products, or subscribe for a service.
Knowing this information allows a business to devote more time and attention to retaining these customers. Another group might include customers from
non-profit organizations, and so on. A general segmentation process is not usually feasible for large volumes of varied data. Therefore, you need an analytical approach to deriving segments and groups from large data sets.
Customers can be grouped based on several factors: including age, gender, interests, spending habits, and so on. The important requirement is to use the available
data to understand and identify how customers are similar to each other.

Let's learn how to divide a set of customers into categories, based on characteristics they share. One of the most adopted approaches that can be used for customer segmentation is clustering. Clustering can group data only "unsupervised," based on the similarity of customers to each other.

It will partition your customers into mutually exclusive groups, for example, into 3 clusters. The customers in each cluster are similar to each other demographically.

Now we can create a profile for each group, considering the common characteristics of each cluster. For example, the first group is made up of AFFULUENT AND MIDDLE AGED customers. The second is made up of YOUNG EDUCATED AND MIDDLE INCOME customers. And the third group includes YOUNG AND LOW INCOME customers. Finally, we can assign each individual in our dataset to one of these groups or segments of customers.

Now imagine that you cross-join this segmented dataset, with the dataset of the product or services that customers purchase from your company.

This information would really help to understand and predict the differences in individual customers' preferences and their buying behaviors across various products.

Indeed, having this information would allow your company to develop highly personalized experiences for each segment. Customer segmentation is one of the popular usages of clustering. Cluster analysis also has many other applications in different domains. So let's first define clustering, and then we'll look at other applications.

Clustering means finding clusters in a dataset, unsupervised.

So, what is a cluster? A cluster is group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to data points in other clusters. Now, the question is, "What is different between clustering and classification?"

Let's look at our customer dataset again. Classification algorithms predict categorical class labels. This means, assigning instances to pre-defined classes such as "Defaulted" or "Non-Defaulted." For example, if an analyst wants to analyze customer data in order to know which customers might default on their payments, she uses a labeled dataset as training data, and uses classification approaches such as a decision tree, Support Vector Machines (or SVM), or, logistic regression to predict the default value for a new, or unknown customer. Generally speaking, classification is a supervised learning where each training data instance belongs to a particular class.

In clustering, however, the data is unlabelled and the process is unsupervised.

For example, we can use a clustering algorithm such as k-Means, to group similar customers as mentioned, and assign them to a cluster, based on whether they share similar attributes, such as age, education, and so on. While I'll be giving you some examples in different industries, I'd like you to think about more samples of clustering.

In the Retail industry, clustering is used to find associations among customers based on their demographic characteristics and use that information to identify buying patterns of various customer groups. Also, it can be used in recommendation systems to find a group of similar items or similar users, and use it for collaborative filtering,

to recommend things like books or movies to customers.
In Banking, analysts find clusters of normal transactions to find the patterns of fraudulent credit card usage. Also, they use clustering to identify clusters
of customers, for instance, to find loyal customers, versus churn customers.
In the Insurance industry, clustering is used for fraud detection in claims analysis, or to evaluate the insurance risk of certain customers based on their segments.
In Publication Media, clustering is used to auto-categorize news based on its content, or to tag news, then cluster it, so as to recommend similar news articles to readers.
In Medicine: it can be used to characterize patient behavior, based on their similar characteristics,
so as to identify successful medical therapies for different illnesses.
Or, in Biology: clustering is used to group genes with similar expression patterns, or to cluster genetic markers to identify family ties.
If you look around, you can find many other applications of clustering, but generally, clustering can be used for one of the following purposes: exploratory data analysis, summary generation or reducing the scale, outlier detection, especially to be used for fraud detection, or noise removal, finding duplicates in datasets, or, as a pre-processing step for either prediction, other data mining tasks, or, as part of a complex system.
Let's briefly look at different clustering algorithms and their characteristics.
Partitioned-based clustering is a group of clustering algorithms that produces sphere-like clusters, such as k-Means, k-Median, or Fuzzy c-Means.
These algorithms are relatively efficient and are used for Medium and Large sized databases.
Hierarchical clustering algorithms produce trees of clusters, such as Agglomerative and Divisive algorithms. This group of algorithms are very intuitive
and are generally good for use with small size datasets.
Density based clustering algorithms produce arbitrary shaped clusters.
They are especially good when dealing with spatial clusters or when there is noise in your dataset, for example, the DBSCAN algorithm.
This concludes our video. Thanks for watching!

DBSCAN Clustering

Hello, and welcome! In this video, we'll be covering DBSCAN,
a density-based clustering algorithm, which is appropriate to use when examining spatial data. So let's get started.
Most of the traditional clustering techniques, such as k-means, hierarchical, and fuzzy clustering,
can be used to group data in an un-supervised way.
However, when applied to tasks with arbitrary shape clusters, or clusters within clusters, traditional techniques might not be able to achieve good results.
That is, elements in the same cluster might not share enough similarity -- or the performance may be poor.
Additionally, while partitioning-based algorithms, such as K-Means, may be easy to understand

and implement in practice, the algorithm has no notion of outliers.
That is, all points are assigned to a cluster, even if they do not belong in any.
In the domain of anomaly detection, this causes problems as anomalous points will be assigned to the same cluster as "normal" data points. The anomalous points pull the cluster centroid towards them, making it harder to classify them as anomalous points.
In contrast, Density-based clustering locates regions of high density that are separated from one another by regions of low density. Density, in this context, is defined as the number of points within a specified radius. A specific and very popular type of density-based clustering is DBSCAN. DBSCAN is particularly effective for tasks like class identification on a spatial context. The wonderful attribute of the DBSCAN algorithm is that it can find out any arbitrary shape cluster without getting affected by noise.
For example, this map shows the location of weather stations in Canada.
DBSCAN can be used here to find the group of stations, which show the same weather conditions.
As you can see, it not only finds different arbitrary shaped clusters, it can find the denser part of data-centered samples by ignoring less-dense areas or noises.
Now, let's look at this clustering algorithm to see how it works.
DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise.
This technique is one of the most common clustering algorithms, which works based on density of
object. DBSCAN works on the idea is that if a particular
point belongs to a cluster, it should be near to lots of other points in that cluster.
It works based on 2 parameters: Radius and Minimum Points.
R determines a specified radius that, if it includes enough points within it, we call it a "dense area." M determines the minimum number of data points we want in a neighborhood to define a cluster.
Let’s define radius as 2 units. For the sake of simplicity, assume it as radius of 2 centimeters around a point of interest. Also, let’s set the minimum point, or M, to be 6 points including the point of interest. To see how DBSCAN works, we have to determine the type of points. Each point in our dataset can be either a
core, border, or outlier point. Don’t worry, I’ll explain what these points are, in a moment. But the whole idea behind the DBSCAN algorithm is to visit each point, and find its type first.
Then we group points as clusters based on their types.
Let’s pick a point randomly. First we check to see whether it’s a core data point.
So, what is a core point? A data point is a core point if, within R-neighborhood of the point, there are at least M points. For example, as there are 6 points in the 2-centimeter neighbor of the red point, we mark this point as a core point.
Ok, what happens if it’s NOT a core point?
Let’s look at another point. Is this point a core point? No.
As you can see, there are only 5 points in this neighborhood, including the yellow point.
So, what kind of point is this one? In fact, it is a "border" point.
What is a border point? A data point is a BORDER point if:

a. Its neighborhood contains less than M data points, or
b. It is reachable from some core point. Here, Reachability means it is within R-distance from a core point. It means that even though the yellow point is within the 2-centimeter neighborhood of the red point, it is not by itself a core point, because it does not have at least 6 points in its neighborhood.
We continue with the next point. As you can see it is also a core point.
And all points around it, which are not core points, are border points.
Next core point.
And next core point.
Let's take this point. You can see it is not a core point, nor is it a border point. So, we'd label it as an outlier.
What is an outlier? An outlier is a point that: Is not a core point, and also, is not close enough to be reachable from a core point.
We continue and visit all the points in the dataset and label them as either Core, Border, or Outlier.
The next step is to connect core points that are neighbors, and put them in the same cluster.
So, a cluster is formed as at least one core point, plus all reachable core points, plus all their borders. It simply shapes all the clusters and finds outliers as well.
Let's review this one more time to see why DBSCAN is cool.
DBSCAN can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by a different cluster. DBSCAN has a notion of noise, and is robust to outliers. On top of that, DBSCAN makes it very practical for use in many really world problems because it does not require one to specify the number of clusters, such as K in k-Means.
This concludes this video. Thanks for watching!

Hierarchical Clustering

Hello, and welcome! In this video, we'll be covering Hierarchical clustering. So let's get started.
Let's look at this chart. An international team of scientists, led by UCLA biologists, used this dendrogram to report genetic data from more than 900 dogs from 85 breeds -- and more than 200 wild gray wolves worldwide, including populations from North America, Europe, the Middle East, and East Asia.
They used molecular genetic techniques to analyze more than 48,000 genetic markers.
This diagram, shows hierarchical clustering of these animals based on the similarity in their genetic data. Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consisting of the clusters of its daughter nodes.
Strategies for hierarchical clustering generally fall into two types: Divisive and Agglomerative.
Divisive is top-down, so you start with all observations in a large cluster and break it down into smaller pieces. Think about divisive as "dividing" the cluster.
Agglomerative is the opposite of divisive, so it is bottom-up, where each observation

starts in its own cluster and pairs of clusters are merged together as they move up the hierarchy. Agglomeration means to amass or collect things, which is exactly what this does with the cluster.
The Agglomerative approach is more popular among data scientists and so it is the main subject of this video.
Let's look at a sample of Agglomerative clustering.
This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, let's say we want to cluster
6 cities in Canada based on their distances from one another.
They are: Toronto, Ottawa, Vancouver, Montreal, Winnipeg, and Edmonton.
We construct a distance matrix at this stage, where the numbers in the row i column j is the distance between the i and j cities. In fact, this table shows the distances between each pair of cities.
The algorithm is started by assigning each city to its own cluster.
So, if we have 6 cities, we have 6 clusters, each containing just one city.
Let's note each city by showing the first two characters of its name.
The first step is to determine which cities -- let's call them clusters from now on -- to merge into a cluster. Usually, we want to take the two closest clusters according to the chosen distance. Looking at the distance matrix, Montreal and Ottawa are the closest clusters. So, we make a cluster out of them.
Please notice that we just use a simple 1-dimentional distance feature here, but our object can be multi-dimensional, and distance measurement can be either Euclidean, Pearson, average distance, or many others, depending on data type and domain knowledge.
Anyhow, we have to merge these two closest cities in the distance matrix as well.
So, rows and columns are merged as the cluster is constructed.
As you can see in the distance matrix, rows and columns related to Montreal and Ottawa cities are merged as the cluster is constructed. Then, the distances from all cities to this new merged cluster get updated. But how? For example, how do we calculate the distance from Winnipeg to the Ottawa-Montreal cluster? Well, there are different approaches, but let's assume, for example, we just select the distance from the centre of the Ottawa-Montreal cluster to Winnipeg. Updating the distance matrix, we now have one less cluster. Next, we look for the closest clusters once again. In this case, Ottawa-Montreal and Toronto are the closest ones, which creates another cluster.
In the next step, the closest distance is between the Vancouver cluster and the Edmonton cluster. Forming a new cluster, their data in the matrix table gets updated. Essentially, the rows and columns are merged as the clusters are merged and the distance updated.
This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters.
In the same way, agglomerative algorithm proceeds by merging clusters.
And we repeat it until all clusters are merged and the tree becomes completed.
It means, until all cities are clustered into a single cluster of size 6.
Hierarchical clustering is typically visualized as a dendrogram as shown on this slide.

Each merge is represented by a horizontal line.
The y-coordinate of the horizontal line is the similarity of the two clusters that were merged, where cities are viewed as singleton clusters.
By moving up from the bottom layer to the top node, a dendrogram allows us to reconstruct the history of merges that resulted in the depicted clustering.
Essentially, Hierarchical clustering does not require a pre-specified number of clusters.
However, in some applications we want a partition of disjoint clusters just as in flat clustering.
In those cases, the hierarchy needs to be cut at some point.
For example here, cutting in a specific level of similarity, we create 3 clusters of similar cities.
This concludes this video. Thanks for watching.

K-Means Clustering

Hello, and welcome! In this video, we'll be covering k-Means clustering. So let's get started.
Imagine that you have a customer dataset, and you need to apply customer segmentation on this historical data. Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics.
One of the algorithms that can be used for customer segmentation is k-Means clustering.
k-Means can group data only "unsupervised," based on the similarity of customers to each other.
Let's define this technique more formally.
There are various types of clustering algorithms, such as partitioning, hierarchical, or density-based clustering. k-Means is a type of partitioning clustering, that is, it divides the data into k non-overlapping subsets (or clusters) without any cluster-internal structure, or labels. This means, it's an unsupervised algorithm.
Objects within a cluster are very similar and objects across different clusters are very different or dissimilar. As you can see, for using k-Means, we have to find similar samples (for example, similar customers).
Now we face a couple of key questions. First, "How can we find the similarity of samples in clustering?" And then, "How do we measure how similar two customers are with regard to their demographics?"
Though the objective of k-Means is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters, it can be shown that instead of a similarity metric, we can use dissimilarity metrics.
In other words, conventionally, the distance of samples from each other is used to shape the clusters. So, we can say, k-Means tries to minimize the "intra-cluster" distances and maximize the "inter-cluster" distances.
Now, the question is, "How we can calculate the dissimilarity or distance of two cases, such as two customers?"
Assume that we have two customers, we'll call them customer 1 and 2.

Let’s also assume that we have only one feature for each of these two customers, and that feature is Age. We can easily use a specific type of Minkowski distance to calculate the distance of these two customers. Indeed, it is the Euclidean distance. Distance of x1 from x2 is root of 34 minus 30 power 2, which is 4. What about if we have more than one feature, for example Age and Income?
For example, if we have income and age for each customer, we can still use the same formula, but this time in a 2-dimensional space.
Also, we can use the same distance matrix for multi-dimensional vectors.
Of course, we have to normalize our feature set to get the accurate dissimilarity measure.
There are other dissimilarity measures as well that can be used for this purpose, but it is highly dependent on data type and also the domain that clustering is done for it.
For example, you may use Euclidean distance, cosine similarity, average distance, and so on. Indeed, the similarity measure highly controls how the clusters are formed, so it is recommended to understand the domain knowledge of your dataset, and data type of features, and then choose the meaningful distance measurement.
Now, let’s see how k-Means clustering works. For the sake of simplicity, let’s assume that our dataset has only two features, the age and income of customers.
This means, it’s a 2-dimentional space. We can show the distribution of customers using a scatterplot. The y-axes indicates Age and the x-axes shows Income of customers. We try to cluster the customer dataset into distinct groups (or clusters) based on these two dimensions.
In the first step, we should determine the number of clusters.
The key concept of the k-Means algorithm is that it randomly picks a center point for each cluster. It means, we must initialize k, which represents "number of clusters." Essentially, determining the number of clusters in a data set, or k, is a hard problem in k-Means that we will discuss later.
For now, let’s put k equals 3 here, for our sample dataset.
It is like we have 3 representative points for our clusters.
These 3 data points are called centroids of clusters, and should be of same feature size of our customer feature set. There are two approaches to choose these centroids:
1) We can randomly choose 3 observations out of the dataset and use these observations as the initial means. Or, 2) We can create 3 random points as centroids of the clusters, which is our choice that is shown in this plot with red color.
After the initialization step, which was defining the centroid of each cluster, we have to assign each customer to the closest center. For this purpose, we have to calculate the distance of each data point (or in our case, each customer) from the centroid points.
As mentioned before, depending on the nature of the data and the purpose for which clustering is being used, different measures of distance may be used to place items into clusters.
Therefore, you will form a matrix where each row represents the distance of a customer from each centroid. It is called the "distance-matrix."
The main objective of k-Means clustering is to minimize the distance of data points from the centroid of its cluster and maximize the distance from other cluster centroids.
So, in this step we have to find the closest centroid to each data point.

We can use the distance-matrix to find the nearest centroid to data points.
Finding the closest centroids for each data point, we assign each data point to that cluster.
In other words, all the customers will fall to a cluster, based on their distance from centroids. We can easily say that it does not result
in good clusters, because the centroids were chosen randomly from the first.
Indeed, the model would have a high error. Here, error is the total distance of each point from its centroid. It can be shown as within-cluster sum of squares error. Intuitively, we try to reduce this error.
It means we should shape clusters in such a way that the total distance of all members of a cluster from its centroid be minimized. Now, the question is, "How we can turn it into better clusters, with less error?"
Okay, we move centroids. In the next step, each cluster center will
be updated to be the mean for data points in its cluster. Indeed, each centroid moves according to their cluster members. In other words, the centroid of each of the 3 clusters becomes the new mean. For example, if Point A coordination is 7.4 and 3.6, and Point B features are 7.8 and 3.8, the new centroid of this cluster with 2 points, would be the average of them, which is 7.6 and 3.7.
Now we have new centroids. As you can guess, once again, we will have
to calculate the distance of all points from the new centroids.
The points are re-clustered and the centroids move again.
This continues until the centroids no longer move.
Please note that whenever a centroid moves, each point's distance to the centroid needs to be measured again.
Yes, k-Means is an iterative algorithm, and we have to repeat steps 2 to 4 until the algorithm converges. In each iteration, it will move the centroids,
calculate the distances from new centroids, and assign the data points to the nearest centroid. It results in the clusters with minimum error,
or the most dense clusters. However, as it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial clusters. It means this algorithm is guaranteed to converge to a result, but the result may be a local optimum (i.e. not necessarily the best possible outcome). To solve this problem, it is common to run
the whole process, multiple times, with different starting conditions.
This means, with randomized starting centroids, it may give a better outcome.
And as the algorithm is usually very fast, it wouldn't be any problem to run it multiple times.
Thanks for watching this video!

More on Hierarchical Clustering

Hello, and welcome! In this video, we'll be covering more details
about Hierarchical clustering. Let's get started.
Let's look at Agglomerative algorithm for Hierarchical Clustering.
Remember that Agglomerative clustering is a bottom-up approach.

Let's say our dataset has n data points. First, we want to create n clusters, one for each data point. Then each point is assigned as a cluster.
Next, we want to compute the distance/proximity matrix, which will be an n by n table.
After that, we want to iteratively run the following steps until the specified cluster number is reached, or until there is only one cluster left.
First, MERGE the two nearest clusters. (Distances are computed already in the proximity matrix.)
Second, UPDATE the proximity matrix with the new values.
We stop after we've reached the specified number of clusters, or there is only one cluster remaining, with the result stored in a dendrogram. So, in the proximity matrix, we have to measure
the distances between clusters, and also merge the clusters that are "nearest."
So, the key operation is the computation of the proximity between the clusters with one point, and also clusters with multiple data points.
At this point, there are a number of key questions that need to be answered.
For instance, "How do we measure the distances between these clusters and How do we define the 'nearest' among clusters?" We also can ask, "Which points do we use?"
First, let's see how to calculate the distance between 2 clusters with 1 point each.
Let's assume that we have a dataset of patients, and we want to cluster them using hierarchy clustering. So, our data points are patients, with a feature
set of 3 dimensions. For example, Age, Body Mass Index (or BMI),
and Blood Pressure. We can use different distance measurements
to calculate the proximity matrix. For instance, Euclidean distance.
So, if we have a dataset of n patients, we can build an n by n dissimilarly-distance matrix. It will give us the distance of clusters with
1 data point. However, as mentioned, we merge clusters in
Agglomerative clustering. Now, the question is, "How can we calculate
the distance between clusters when there are multiple patients in each cluster?"
We can use different criteria to find the closest clusters, and merge them.
In general, it completely depends on the data type, dimensionality of data, and most importantly, the domain knowledge of the dataset. In fact, different approaches to defining
the distance between clusters, distinguish the different algorithms.
As you might imagine, there are multiple ways we can do this.
The first one is called Single-Linkage Clustering. Single linkage is defined as the shortest distance between 2 points in each cluster, such as point "a" and "b".
Next up is Complete-Linkage Clustering. This time, we are finding the longest distance between points in each cluster, such as the distance between point "a" and "b".
The third type of linkage is Average Linkage Clustering, or the mean distance.
This means we're looking at the average distance of each point from one cluster to every point in another cluster. The final linkage type to be reviewed is Centroid Linkage Clustering. Centroid is the average of the feature sets
of points in a cluster. This linkage takes into account the centroid
of each cluster when determining the minimum distance.
There are 3 main advantages to using hierarchical clustering.

First, we do not need to specify the number of clusters required for the algorithm.
Second, hierarchical clustering is easy to implement.
And third, the dendrogram produced is very useful in understanding the data.
There are some disadvantages as well. First, the algorithm can never undo any previous steps. So for example, the algorithm clusters 2 points,
and later on we see that the connection was not a good one, the program cannot undo that step. Second, the time complexity for the clustering
can result in very long computation times, in comparison with efficient algorithms, such k-Means. Finally, if we have a large dataset, it can
become difficult to determine the correct number of clusters by the dendrogram.
Now, let's compare Hierarchical clustering with k-Means.
K-Means is more efficient for large datasets. In contrast to k-Means, Hierarchical clustering does not require the number of clusters to be specified.
Hierarchical clustering gives more than one partitioning depending on the resolution, whereas k-Means gives only one partitioning of the data.
Hierarchical clustering always generates the same clusters, in contrast with k-Means that returns different clusters each time it is run due to random initialization of centroids.
Thanks for watching!

More on K-Means

Hello, and welcome! In this video, we'll look at k-Means accuracy
and characteristics. Let's get started.
Let's define the algorithm more concretely before we talk about its accuracy.
A k-Means algorithm works by randomly placing k centroids, one for each cluster.
The farther apart the clusters are placed, the better.
The next step is to calculate the distance of each data point (or object) from the centroids.
Euclidean distance is used to measure the distance from the object to the centroid.
Please note, however, that you can also use different types of distance measurements, not just Euclidean distance. Euclidean distance is used because it's
the most popular. Then, assign each data point (or object) to
its closest centroid, creating a group. Next, once each data point has been classified to a group, recalculate the position of the k centroids. The new centroid position is determined by the mean of all points in the group.
Finally, this continues until the centroids no longer move.
Now the question is, "How can we evaluate the 'goodness' of the clusters formed by k-Means?" In other words, "How do we calculate the accuracy
of k-Means clustering?" One way is to compare the clusters with the
ground truth, if it's available. However, because k-Means is an unsupervised algorithm, we usually don't have ground truth in real world problems to be used.
But, there is still a way to say how bad each cluster is, based on the objective of the k-Means. This value is the average distance between
data points within a cluster. Also, average of the distances of data points

from their cluster centroids can be used as a metric of error for the clustering algorithm.
Essentially, determining the number of clusters in a data set, or k, as in the k-Means algorithm, is a frequent problem in data clustering. The correct choice of k is often ambiguous, because it's very dependent on the shape and scale of the distribution of points in a data set. There are some approaches to address this problem, but one of the techniques that is commonly used, is to run the clustering across the different values of K, and looking at a metric of accuracy for clustering.
This metric can be "mean distance between data points and their cluster centroid," which indicate how dense our clusters are, or to what extend we minimized the error of clustering.
Then looking at the change of this metric, we can find the best value for k.
But the problem is that with increasing the number of clusters, the distance of centroids to data points will always reduce. This means, increasing K will always decrease the "error."
So, the value of the metric as a function of K is plotted and the "elbow point" is determined, where the rate of decrease sharply shifts. It is the right K for clustering.
This method is called the "elbow" method.
So, let's recap k-Means clustering: k-Means is a partitioned-based clustering, which is:
a) Relatively efficient on medium and large sized datasets;
b) Produces sphere-like clusters, because the clusters are shaped around the centroids;
and c) Its drawback is that we should pre-specify the number of clusters, and this is not an easy task.
Thanks for watching!

Collaborative Filtering

Hello, and welcome! In this video, we'll be covering a recommender system technique called, Collaborative filtering. So let's get started.
Collaborative filtering is based on the fact that relationships exist between products and people's interests. Many recommendation systems use Collaborative filtering to find these relationships and to give an accurate recommendation of a product that the user might like or be interested in.
Collaborative filtering has basically two approaches: User-based and Item-based.
User-based collaborative filtering is based on the user's similarity or neighborhood.
Item-based collaborative filtering is based on similarity among items.
Let's first look at the intuition behind the "user-based" approach.
In user-based collaborative filtering, we have an active user for whom the recommendation is aimed. The collaborative filtering engine, first looks for users who are similar, that is, users who share the active user's rating patterns. Collaborative filtering bases this similarity on things like history, preference, and choices that users make when buying, watching, or enjoying something. For example, movies that similar users have rated highly. Then, it uses the ratings from these similar users to predict the possible ratings by the active user for a movie that she had not previously watched. For instance, if 2 users are similar or are neighbors, in terms of their interest in movies, we can recommend a movie to the active user

that her neighbor has already seen. Now, let's dive into the algorithm to see how all of this works.

Assume that we have a simple user-item matrix, which shows the ratings of 4 users for 5 different movies. Let's also assume that our active user has watched and rated 3 out of these 5 movies. Let's find out which of the two movies that our active user hasn't watched, should be recommended to her.

The first step is to discover how similar the active user is to the other users.

How do we do this? Well, this can be done through several different statistical and vectorial techniques such as distance or similarity measurements, including Euclidean Distance, Pearson Correlation, Cosine Similarity, and so on.

To calculate the level of similarity between 2 users, we use the 3 movies that both the users have rated in the past. Regardless of what we use for similarity measurement, let's say, for example, the similarity, could be 0.7, 0.9, and 0.4 between the active user and other users. These numbers represent similarity weights, or proximity of the active user to other users in the dataset.

The next step is to create a weighted rating matrix.

We just calculated the similarity of users to our active user in the previous slide.

Now we can use it to calculate the possible opinion of the active user about our 2 target movies. This is achieved by multiplying the similarity weights to the user ratings. It results in a weighted ratings matrix, which represents the user's neighbour's opinion about our 2 candidate movies for recommendation.

In fact, it incorporates the behaviour of other users and gives more weight to the ratings of those users who are more similar to the active user.

Now we can generate the recommendation matrix by aggregating all of the weighted rates.

However, as 3 users rated the first potential movie, and 2 users rated the second movie, we have to normalize the weighted rating values. We do this by dividing it by the sum of the similarity index for users. The result is the potential rating that our active user will give to these movies, based on her similarity to other users.

It is obvious that we can use it to rank the movies for providing recommendation to our active user.

Now, let's examine what's different between "User-based" and "Item-based" Collaborative filtering: In the User-based approach, the recommendation is based on users of the same neighborhood, with whom he or she shares common preferences.

For example, as User1 and User3 both liked Item 3 and Item 4, we consider them as similar or neighbor users, and recommend Item 1, which is positively rated by User1 to User3.

In the item-based approach, similar items build neighborhoods on the behavior of users. (Please note, however, that it is NOT based on their content).

For example, Item 1 and Item 3 are considered neighbors, as they were positively rated by both User1 and User2. So, Item 1 can be recommended to User 3 as he has already shown interest in Item3. Therefore, the recommendations here are based on the items in the neighborhood that a user might prefer.

Collaborative filtering is a very effective recommendation system, however, there are

some challenges with it as well. One of them is Data Sparsity.
Data sparsity happens when you have a large dataset of users, who generally, rate only a limited number of items. As mentioned, collaborative-based recommenders can only predict scoring of an item if there are other users who have rated it.
Due to sparsity, we might not have enough ratings in the user-item dataset, which makes it impossible to provide proper recommendations. Another issue to keep in mind is something called 'cold start.' Cold start refers to the difficulty the recommendation system has when there is a new user and, as such, a profile doesn't exist for them yet.
Cold start can also happen when we have a new item, which has not received a rating.
Scalability can become an issue, as well. As the number of users or items increases and the amount of data expands, Collaborative filtering algorithms will begin to suffer drops in performance, simply due to growth in the similarity computation.
There are some solutions for each of these challenges, such as using hybrid-based recommender systems, but they are out of scope of this course.
Thanks for watching!

Content-Based Recommender Systems

Hello, and welcome! In this video, we'll be covering content-based recommendation systems. So let's get started.
A content-based recommendation system tries to recommend items to users, based on their profile. The user's profile revolves around that user's preferences and tastes. It is shaped based on user ratings, including the number of times that user has clicked on different items or perhaps, even liked those items. The recommendation process is based on the similarity between those items. Similarity, or closeness of items, is measured based on the similarity in the content of those items.
When we say content, we're talking about things like the item's category, tag, genre, and so on. For example, if we have 4 movies, and if the user likes or rates the first 2 items, and if item 3 is similar to item 1, in terms of their genre, the engine will also recommend item 3 to the user.
In essence, this is what content-based recommender system engines do.
Now, let's dive into a content-based recommender system to see how it works.
Let's assume we have a dataset of only 6 movies.
This dataset shows movies that our user has watched, and also the genre of each of the movies. For example, "Batman versus Superman" is in the Adventure, Super Hero genre. And "Guardians of the Galaxy" is in Comedy, Adventure, Super Hero, and Science-Fiction genres.
Let's say the user has watched and rated 3 movies so far and she has given a rating of 2 out of 10 to the first movie, 10 out of 10 to the second movie, and an 8 out of 10 to the third. The task of the recommender engine is to recommend one of the 3 candidate movies to this user. Or, in other words, we want to predict what

the user's possible rating would be, of the 3 candidate movies if she were to watch them.
To achieve this, we have to build the user profile.
First, we create a vector to show the user's ratings for the movies that she's already watched. We call it "input user ratings." Then, we encode the movies through the "One Hot Encoding" approach. Genre of movies are used here as a feature set.
We use the first 3 movies to make this matrix,
which represents the movie 'feature-set' matrix.
If we multiply these 2 matrices, we can get the "weighted feature set" for the movies.
Let's take a look at the result. This matrix is also called the "Weighted Genre Matrix," and represents the interests of the user for each genre based on the movies that she's watched. Now, given the weighted genre matrix, we can shape the profile of our active user. Essentially, we can aggregate the weighted genres, and then normalize them to find the user profile.
It clearly indicates that she likes "super hero" movies more than other genres.
We use this profile to figure out what movie is proper to recommend to this user.
Recall that we also had 3 candidate movies for recommendation, that haven't been watched by the user. We encode these movies as well.
Now we're in the position where we have to figure out which of them is most suited to be recommended to the user.
To do this, we simply multiply the user-profile matrix by the candidate movie matrix, which results in the "weighted movies" matrix. It shows the weight of each genre, with respect to the user profile. Now, if we aggregate these weighted ratings, we get the active user's possible interest-level in these 3 movies.
In essence, it's our "recommendation" list, which we can sort to rank the movies, and recommend them to the user. For example, we can say that the "Hitchhiker's Guide to the Galaxy" has the highest score in our list, and is proper to recommend to the user. Now you can come back and fill the predicted ratings for the user.
So, to recap what we've discussed so far, the recommendation in a content-based system, is based on user's tastes, and the content or feature set items.
Such a model is very efficient. However, in some cases it doesn't work.
For example, assume that we have a movie in the "drama" genre, which the user has never watched. So, this genre would not be in her profile.
Therefore, she'll only get recommendations related to genres that are already in her profile, and the recommender engine may never recommend any movie within other genres.
This problem can be solved by other types of recommender systems such as "Collaborative Filtering."
Thanks for watching!

Intro to Recommender Systems

Hello, and welcome! In this video, we'll be going through a quick introduction to recommendation systems. So, let's get started.

Even though people's tastes may vary, they generally follow patterns.
By that, I mean that there are similarities in the things that people tend to like … or another way to look at it, is that people tend to like things in the same category or things that share the same characteristics. For example, if you've recently purchased a book on Machine Learning in Python and you've enjoyed reading it, it's very likely that you'll also enjoy reading a book on Data Visualization.
People also tend to have similar tastes to those of the people they're close to in their lives. Recommender systems try to capture these patterns and similar behaviors, to help predict what else you might like.
Recommender systems have many applications that I'm sure you're already familiar with. Indeed, Recommender systems are usually at play on many websites. For example, suggesting books on Amazon and movies on Netflix. In fact, everything on Netflix's website is driven by customer selection. If a certain movie gets viewed frequently enough, Netflix's recommender system ensures that that movie gets an increasing number of recommendations.
Another example can be found in a daily-use mobile app, where a recommender engine is used to recommend anything from where to eat, or, what job to apply to.
On social media, sites like Facebook or LinkedIn, regularly recommend friendships.
Recommender systems are even used to personalize your experience on the web.
For example, when you go to a news platform website, a recommender system will make note of the types of stories that you clicked on and make recommendations on which types of stories you might be interested in reading, in future.
There are many of these types of examples and they are growing in number every day.
So, let's take a closer look at the main benefits of using a recommendation system.
One of the main advantages of using recommendation systems is that users get a broader exposure to many different products they might be interested in.
This exposure encourages users towards continual usage or purchase of their product.
Not only does this provide a better experience for the user but it benefits the service provider, as well, with increased potential revenue and better security for its customers.
There are generally 2 main types of recommendation systems: content-based and collaborative filtering.
The main difference between each, can be summed up by the type of statement that a consumer might make. For instance, the main paradigm of a content-based recommendation system is driven by the statement: "Show me more of the same of what I've liked before."
Content-based systems try to figure out what a user's favorite aspects of an item are, and then make recommendations on items that share those aspects.
Collaborative filtering is based on a user saying, "Tell me what's popular among my neighbors because I might like it too." Collaborative filtering techniques find similar groups of users, and provide recommendations based on similar tastes within that group.

In short, it assumes that a user might be interested in what similar users are interested in.
Also, there are Hybrid recommender systems,
which combine various mechanisms.
In terms of implementing recommender systems, there are 2 types: Memory-based and Model-based.
In memory-based approaches, we use the entire user-item dataset to generate a recommendation
system. It uses statistical techniques to approximate
users or items. Examples of these techniques include: Pearson
Correlation, Cosine Similarity and Euclidean Distance, among others.
In model-based approaches, a model of users is developed in an attempt to learn their preferences. Models can be created using Machine Learning
techniques like regression, clustering, classification, and so on.
This is the end of our video. Thanks for watching!