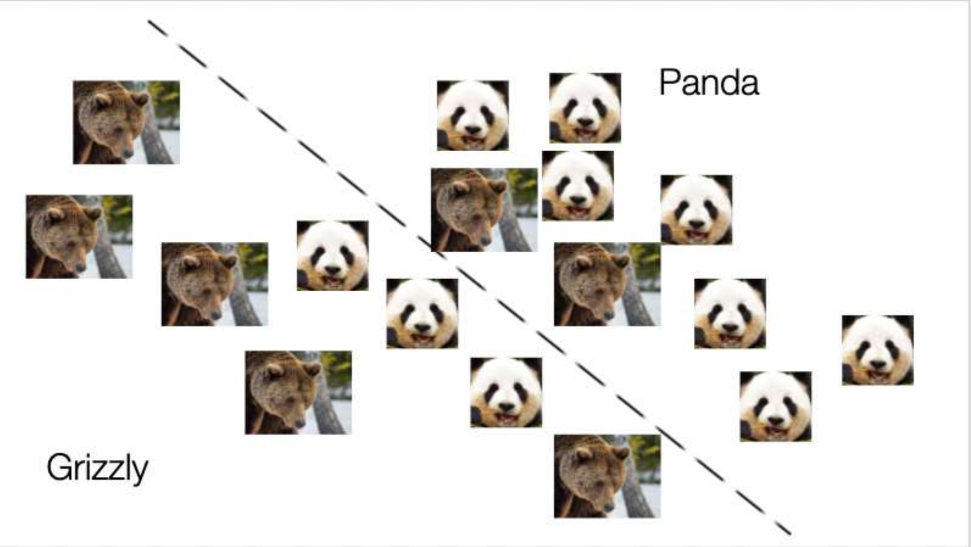
Module 5-1: Clustering

Before moving on to the final type of machine learning model, let's quickly review the ones we have learned up until now.

Classification

Here we have an example of linear classification. The **linear classifier** (the dotted line in the middle) aims to classify bears as being Grizzly or Panda based on different attributes.

Notice that the model is not perfect and will sometimes misclassify data. Since the model is not perfect, sometimes a Panda that has similar features to a Grizzly will be classified as a Grizzly. Conversely, a Grizzly that has similar features to a Panda might be classified as a Panda.

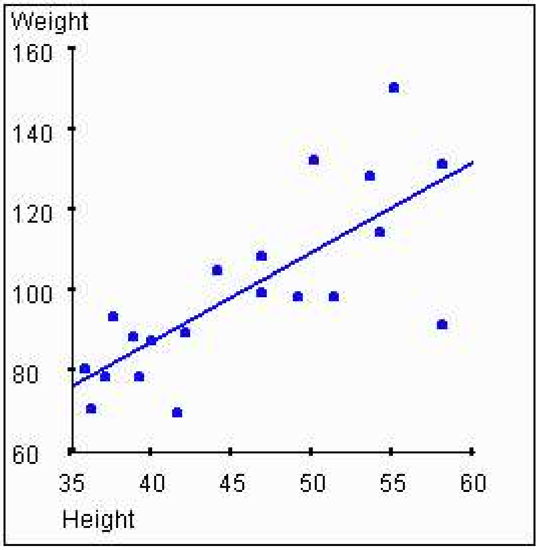


Regression

Here we have an example of Linear Regression. Each blue square represents a mouse. Each mouse is plotted based on their height and weight.

Send Issue

The blue line represents the linear regression model. We can use this model to predict a mouse's Weight depending on its height. This is done by first plotting the height of the mouse onto the x axis, and then using the blue line's y value at that x to represent the weight.



Supervised vs Unsupervised Training

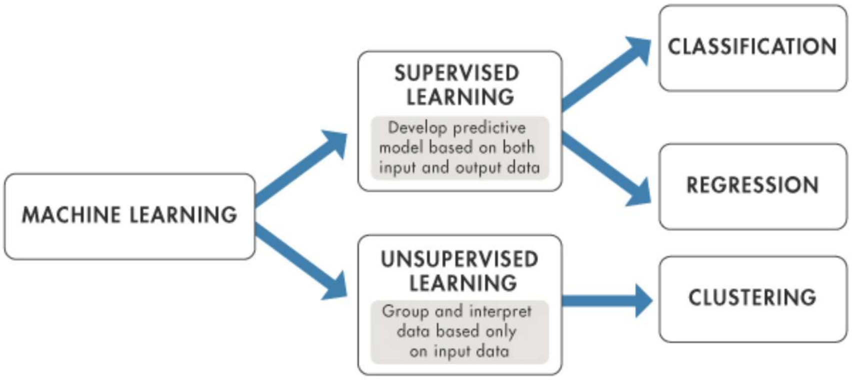
So far, we have only learned about supervised training models. These were regression and classification. However, there are many other machine learning models. The next model we will learn about is clustering, which is an unsupervised training model.

But what's the difference between a supervised training model and an unsupervised training model?

In **supervised training**, the algorithms improve themselves by learning from training data that already contains data that has been labeled. The training data contains an input and an output.

For example, we can train a linear regression model that has training data which contains flower color and length as input and the type of flower as output. Then when the model makes predictions, it predicts an output that we have defined. In this case, the model will predict the type of flower.

In **unsupervised training**, the algorithms use patterns within the data to predict an unknown output. All the algorithm has to work off of is preexisting input; they do not have an actual expected output.

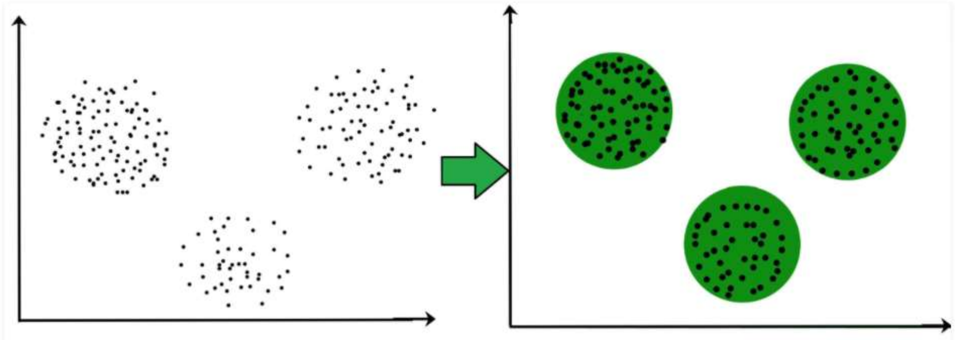

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Clustering

Clustering is the assignment of data into "clusters" or groupings, so that data in the same cluster are similar to each other. The goal of clustering is to categorize all of the data within different groups. Clustering works by looking at the features within the individual data points and then infers how they should be grouped.

Just like how there are different types of algorithms for classification (linear classification, svm, etc.) or regression (linear regression, polynomial regression, etc.), there are different types of algorithms for clustering.

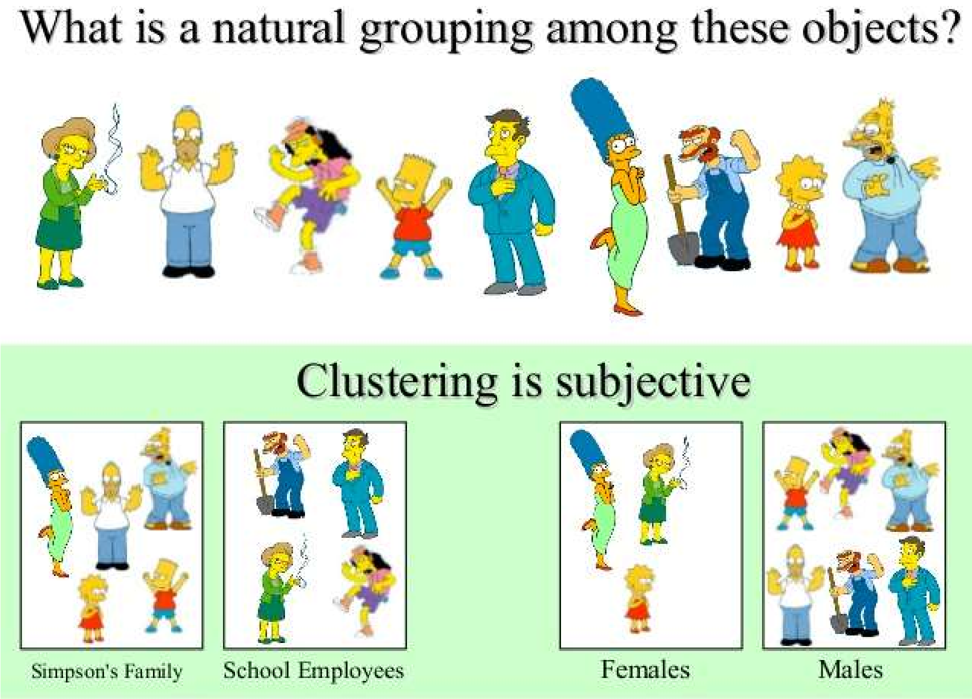
The kind of clustering algorithm you decide to use will greatly affect the way that the data is clustered. Let's look at some examples of clustering in action!



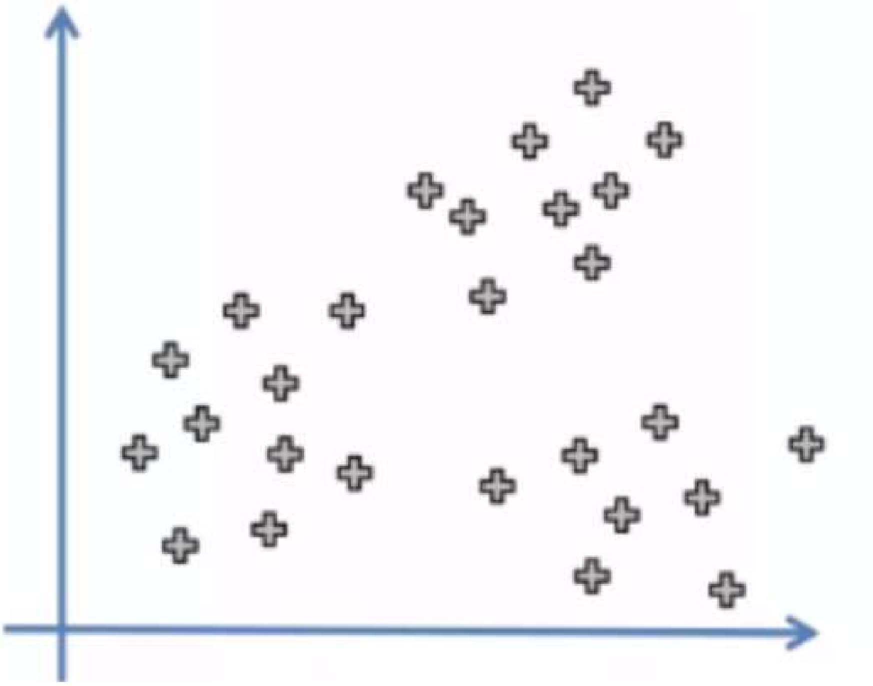
On the left you can see the data that is not clustered, and then on the right you can see what the data looks like after it gets clustered. The data was grouped into three different clusters based on existing characteristics.

When the clustering model is trained, the criteria for placing data into certain clusters is determined completely by the attributes of those features and how similar they are to other data. This is different from classification, because when a classification model gets trained, we specify exactly what class each data point should belong in. It is not determined by the attributes of the features of the data.

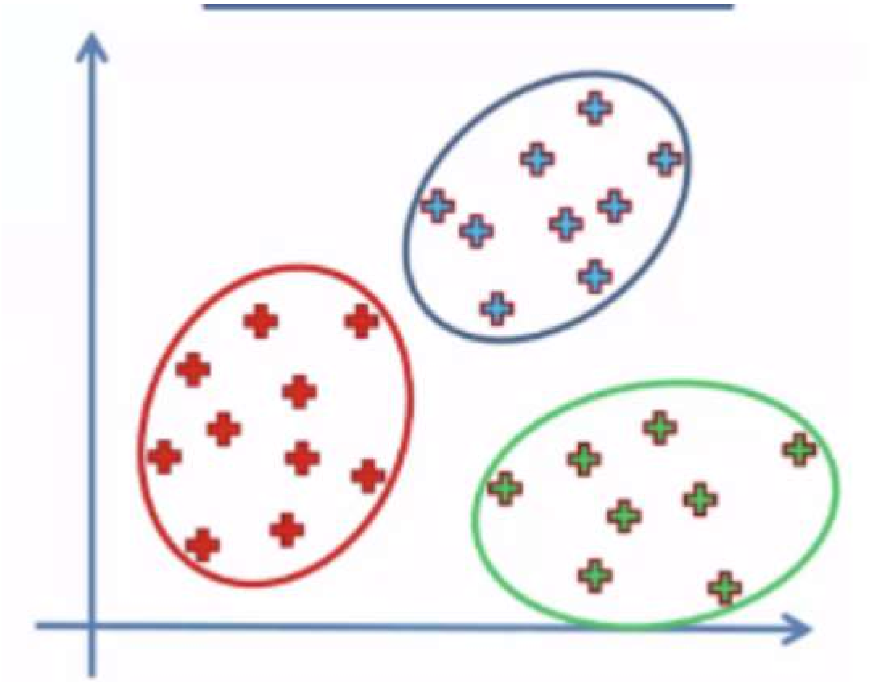
This Simpsons example is a perfect way to show the flexibility of clustering. On the left, the Simpsons characters are clustered based on whether or not they belong to the Simpson's family or if they are school employees. On the right the Simpsons characters are clustered based on whether or not they are male or female. The idea is that we can feed the model only certain attributes about the data we want to cluster, and this will affect the actual clustering process.



One of the biggest advantages of clustering is the ability to find underlying characteristics within data that might not be easily visible. Let's say that you are in charge of a business that sells limited edition sneakers and want to increase sales. Through surveys and monitoring sales data you plot the following data.



Each cross represents a customer as well as how much they spend at the store and what time they came in. You then use clustering to find some trends within your customers.



After grouping the data into three different clusters you realize some similarities between the customers. Customers who come in earlier tend to spend more money and fall into one category. Conversely, customers who come in later tend to spend less money and all fall into another category.

After examining more data about the customers in each of these clusters, you come to the conclusion that the customers who come later in the day are also students so they have less money to spend.

Now that you were able to identify your customer base, you implement a student discount for certain items that will entice students to invite their friends and buy more sneakers. Clustering allowed you to find a trend you had no idea existed within your customers. Now you have better tapped into the student demographic.

K Means Clustering

Here we have another example of unclustered and clustered data. However, now we will talk specifically about K means clustering. **K means clustering** is just a specific clustering algorithm the same way linear regression is a form of regression algorithm and linear classification is a form of classification algorithm.

In K means clustering, K stands for the amount of clusters you want the machine learning model to create. In this example, K would be