Unsupervised learning

The data is not labelled as right or wrong

The Idea is to find the structure in the data, may be into 2 clusters **(this is one type of unsupervised algorithm)**

|  |
| --- |
| How exactly is that accomplished? |

Example

This algorithm is used in google news

So basically it clusters similar news into a bunch of groups

Cluster Gene

Given Human Genome, can tell how much of different types of genes one has

Certain individuals do or do not have a specific gene

And then the individuals can be grouped

This is unsupervised because we are not telling in advance if these are type 1 people or etc.

Few other areas where you can use

Social network, market segment, astronomical and computing

We can find which 2 items have similarity and then group them so they can interact or can be served efficiently

|  |
| --- |
| Here is important point of behaviour modification |

Cocktail party algorithm

Here it is possible to cluster one person’s and another person’s voice and separate it out using a single audio recording where both are speaking

|  |
| --- |
| Understand the shit well, and think where else or how else can you use this?  I want to use this to separate the various instruments in a song  But I guess there is already a software for that which is paid  Or maybe use this remove noise from the communication  Or maybe use this into the noise cancelling phones etc. |

The idea is you keep 2 microphones are different distances to the speaker, so that one type of sounds comes more strong on one microphone and other type on other

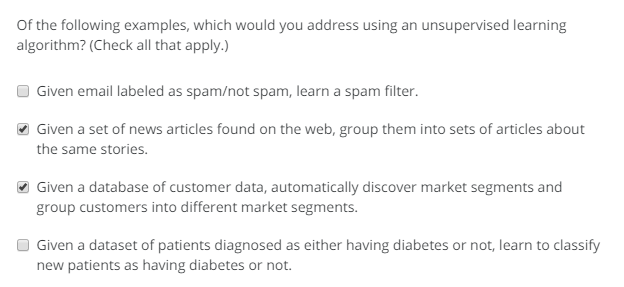
So you can differentiate

Use **Octave** programming environment and implement the algorithms

It seems you can learn much faster if you use this

This guy, who is incredibly smart and yet is able to explain things in such simple terms, is really fascinating.

I guess you need to be really very smart to explain things in very simple terms.



**Linear regression**

Why linear regression is called linear?

In LR the output can be written as

Y = C + Ax

Where C is a constant and A is a parameter and x is predictor

So if you can write anything in this form then it is linear

Non-linear are something like sine, cosine or squared etc.

Y = bo + b1X1 + b2X12

This equation is also linear, since it is of the form constant and product of parameter and predictor

Non-linear are something that does not fit linear model

<http://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-difference-between-linear-and-nonlinear-equations-in-regression-analysis>

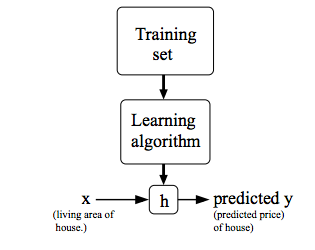
Regression = predicting real-valued output

Classification = predict discrete value output

Training set🡪Algorithm🡪hypothesis (function)

And this function will take the input and produce the output

|  |  |
| --- | --- |
| Hypothesis  Suggested explanation for a phenomenon or prediction of a phenomena  Not well tested | Theory  Well substantiated explanation for a proven hypothesis  Well tested and substantiated |



Formal definition:

Given a training set, learn the function h: x🡪y, such that h(x) is a good predictor of corresponding value of y

How to do a minimization problem

How to know if you have understood the concept by solving the problems

That is the only way

Or apply that concept somewhere else

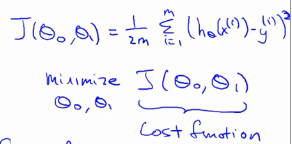
**Cost Function**

For the training set, how do you come up with the values for parameters such it corresponds to a straight line that fits the values?

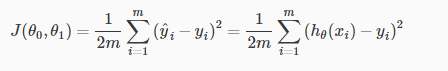
Another way to say is, get the parameter values for the best fit line for the data

Our prediction h(x) should be close to actual y for the training examples

Minimize the J is the cost function and it is also called squared error function



In other words, the accuracy of our hypothesis is computed using the cost function



Half of mean is taken for the convenience and it eventually gets cancelled out when taken the derivative

….

**Gradient Descent**

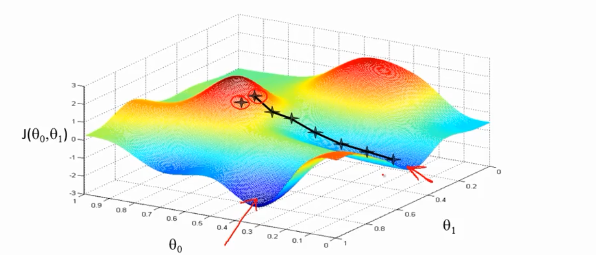
This is an algorithm that is used to minimize the cost function

Not just that but many other functions in ML

So the algorithm

1. Start with random values, for example theta 0=0, and theta 1=0

And then keep trying different values to minimize the cost function



Here we start from some random point

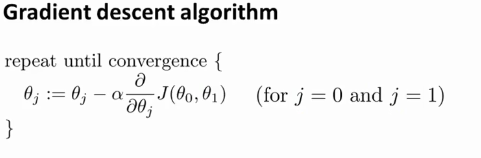
And then going downhill from that point to the local minimum

Here if you start at a different point your local minimum will be different

|  |
| --- |
| Why exactly is that?  It is apparently the property of gradient descent |

A: = b (assignment)

A = b (check if equals)



Alpha is called the learning rate and it controls how much bigger step we take towards the descent

Derivative term

We need to update the values of theta o and theta 1 for j = 0 and j = 1

And we need to that simultaneously

Compute right hand side and **simultaneously** update both the values

Derivative is basically the tangential line to the cost function

Slope of the tangent is the derivative at that point

And this gives the direction to move towards

So the answer to my earlier question

The direction of the step is determined by the partial derivative of J (theta 0 and theta 1)

|  |
| --- |
| How exactly does the partial derivative give the tangential line to a function?  Meaning, why is partial derivative is equivalent to slope of the line |

Derivative in this case slope of the line that is tangent to the function

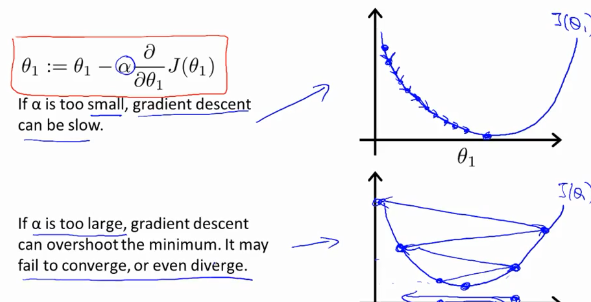
And that slope can either be positive or negative

The idea of Gradient Descent is it will guide us towards the right direction leading us to the minimization of cost function

It is very important to take right values for the alpha

Because if it is too small then you will take many learning steps to reach the minimum

And if it’s too large, you may overshoot the minimum and then may fail to converge or even diverge



If you are already at the local minimum

Then gradient decent will keep the value unchanged

In gradient descent as we approach the minimum we will automatically take smaller steps

That is because the slope gets smaller and smaller

Apply the gradient descent algorithm to the Linear Regression Model to minimize the cost function

Gradient descent can have different local minimum but

Cost function for **linear regression** is always a **convex function** so that will have only one local optimum not many like gradient descent

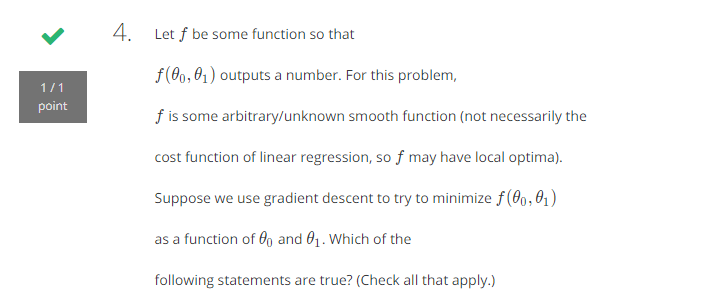
|  |
| --- |
| One of the problem with supervised learning is, you need accurate data  How can you ensure authenticity and legitimacy of the data  At the moment, the data is public but not in a usable form or the data is structured and sold by companies or some data is just private information extracted by some players and used to their advantage like google  And also many data are not captured  Best way to get details of the data is to follow the money trail |

|  |
| --- |
| How is the Machine Learning Linear regression different from statistics?  Hard to answer, some discussion here  <https://stats.stackexchange.com/questions/268755/when-should-linear-regression-be-called-machine-learning> |

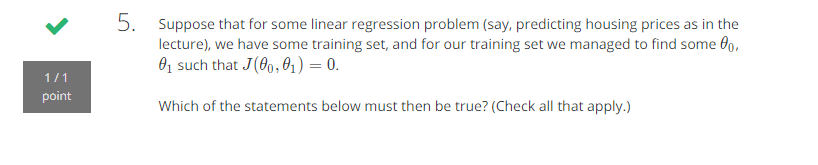
The gradient descent we are using is called batch gradient descent, because at any point we are using whole/entire of the training examples

You can find the minimum values for theta without needing the partial derivative and multiple steps using the method of statistics

The thing is the gradient descent scales better for large data set and large parameters than the normal method



What are true?



<https://www.coursera.org/learn/machine-learning/lecture/38jIT/matrices-and-vectors>