# Machine Learning - Stanford

## Examples of Machine Learning

* + DB Mining : Web Click Data, Medical Records, Biology
  + Applications that cannot be written by hand : NLP, Computer Vision , Handwriting Recognition
  + Self-Customizing Programs : Amazon, Netflix product recommendations
  + Understanding the human brain : Real brain analysis

## What is Machine Learning

Arthur Samuel : “Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed”

Tom Mitchell : “A computer program is said to learn from Experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P improves with experience E”

### Learning types

* Supervised Learning
* Unsupervised learning
* Reinforcement learning
* Recommender systems

### Learning Models

* Classification
* Regression
* Clustering
* Non-Clustering

#### Supervised Learning :

Learning which has the “Correct Answers” inputted e.g. regression models are often built on datasets of pre-existing data and then trained to be able to predict features, similarly classification models are classified and the computer will pull data / draw correlations from them

Unsupervised learning:  
Learning which allows the computer to decide how to cluster or group the data by itself and do not take any input; the computer will then teach itself on this. There is no feedback based on prediction results

##### Cocktail Party Algorithm

Imagine there a two people in a room each counting to ten in different languages ; and there are two microphones which each record one person loudly and the other quietly ; This algorithm can separate each source into two separate distinct audios this can be done with the following line

## Model Representation and Cost Function

Notation :

* M = Number of Training Examples
* X’s = “input” variables
* Y’s = “output” variables
* (x,y) – Training Examples
* (xi,yi) = ith Training Examples
* h = hypothesis function mapping x to y e.g. “H : X -> Y”
* (Univariate Linear Regression) ->

### Cost Function ( Mean Squared Error )

A picture containing chart

Description automatically generatedSometimes referred to as Squared error function, this can measure the accuracy of our hypothesis function; this plots a quadratic and finds its minimum point to minimise the value

Choose θ0 , θ1 so that h0(x) is close to y for your training example (x,y) , minimise J(θ0 , θ1)

0 θ1) = (x[i])– y[i])2

When measuring both θ0 and θ1 we use contour plots to show the correlation , the closer to the centre of the contour the more accurate it will be

### Text, letter Description automatically generatedGradient Descent Algorithm

Initialise variables and then repeat until you reach a local optimum (minimum) within the following equation simultaneous updates should be used

α = Learning Rate

If α is too small, GD algorithm can be slow

If α is too large, GD can overshoot the minimum and fail to converge or even diverge

As GD runs for longer GD will take smaller steps as derivative becomes smaller

### Gradient Descent with Linear Regression

Within Linear Regression all functions are convex functions and thus there is only one local optimum the aforementioned variant of GD is using Batch GD.

## Linear Algebra

Matrixes are a 2D array of numbers typically represented as MxN where M and N are integers e.g 4x2 Matrix and where M is the number of rows and N the number of columns it can also be represented a R^MxN

Matrix Indexing AIJ is the Ith row and the Jth column

Vectors are a form of matrixes represented in Nx1 these can be represented as R^N

These can be either zero-indexed or one-indexed

### Multiplication and Addition Within Matrices

Matrixes are added via layer addition e.g., adding where the two values overlap and then summing them , matrixes of different lengths cannot be added. Matrices can also be multiplied by scaler quantities

Text

Description automatically generatedVectors and Matrices can be multiplied by cancelling out the N ; Matrices are non-commutative however they are Associative, Identity matrix denoted as IIxI which is zero except for in y = -x

### Diagram Description automatically generatedMatrix Inverse and Transpose

As 3-1 is the inverse of 3, each number has inverses

A picture containing text, receipt

Description automatically generated

Diagram

Description automatically generated

Diagram, text, whiteboard

Description automatically generatedA picture containing text, clock, watch

Description automatically generated

## Multivariate Linear Regression

n = number of features

x[i] = input features of Ith training examples

yj[i] = value of feature j in Ith training examples

x0[i] = 1

### Gradient Descent

### Feature Scaling

Making all features be on a similar scale means you get a faster result from GD , e.g. normalizing values; usually all features should be in a range from -1 -> 1

### Mean Normalization

You may do the following make features have exactly zero mean

Or you could do the following

Where muon is the average value of x from training set and s is either the standard deviation or range

### Learning Rate

Debugging GD : plotting the amount of iterations with the level of j(*θ) to ensure that value of alpha is right*

*Automatic Convergence testing : this states that if change by ε is observed in* j(*θ) where ε is a value like 10-3*

### Polynomial Regression

*You can use feature modelling to create new features and optimise them; this also allows you to use a different non-linear method e.g quadratic or cubic graphs which better fit the data*

## Normal Equation

*Method to solve for theta analytically*

*1D :*

*Using calculus the global minimum of J can be found*

*1 > D :*

*Using calculus on all values of J a global minimum can be found*

### Normal Equation

*Advantages and Disadvantages in Comparison to GD*

|  |  |
| --- | --- |
| *GD* | *Normal Equation* |
| * *Needs to Choose Alpha* | *+ No Need to choose Alpha or to iterate* |
| * *Needs Multiple Iterations* | * *Need to Compute* |
| *+ works when N is large* | * *Slow if N is large* |
| *O(KN2)* | *O(n3)* |

### Non invertibility

*Provided that Matrix X is degenerate; features should be changed*

*This can be caused by the following*

* *Linearly Dependent / Redundant Features*
* *Too many features where (M =< n)*