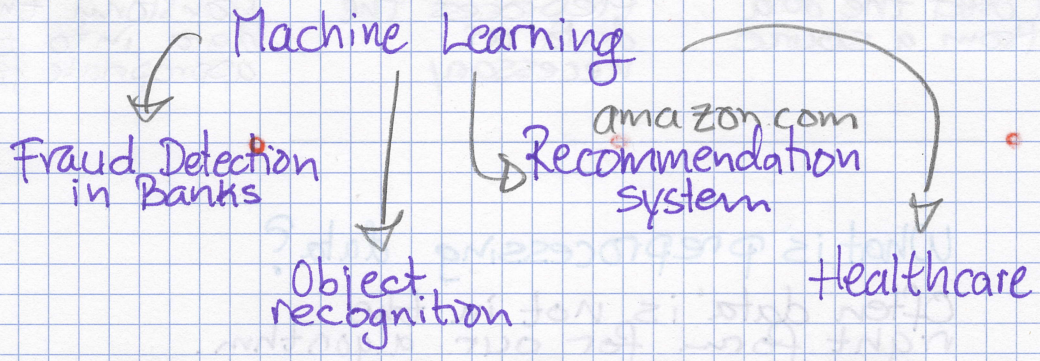
The Purposes of Machine Learning

- Why do we use ML
- Variety of data types
- Variety of ML algorithms
- Simple classification task
- Why should we choose C++ for coding

Why Do we Use ML?

What is Machine Learning?

- Machine Learning is a broad discipline but generally it involves making a prediction of an output given an input to a system
- An example might be the prediction of stock price given previous stock prices.



- Implementation of ML algorithms because of the advent of faster computer hardware
- Computers are incredibly efficient - saving time and money for an employer.
- We can process a lot of data

Variety of Data Types

In what way is data represented?

Data comes in many different shapes & formats

Where can I get Data from?

- Collating data is time consuming
- Amazon Mechanical Turk allows you to outsource the task
- The internet contains a vast amount of data
- Twitter, Youtube, Facebook and so on allows easy access to information and trends
- Academic institutions - universities often host free copies of the datasets they used to train their algorithms.

Images

Databases

Signals

Step 1
Collect the data from a source

Step 2
Preprocess the data as necessary

Step 3
Transform the data into the appropriate form.

What is preprocessing data?

Often data is not in the right form for our algorithm.

- Data entries can sometime be missing in a database.
- Sometimes there is just too much data to deal with, therefore we need to sample
- We can aggregate data sources together to create richer models or remove redundant data

What is sampling?

If the data source is too large, we need to sample

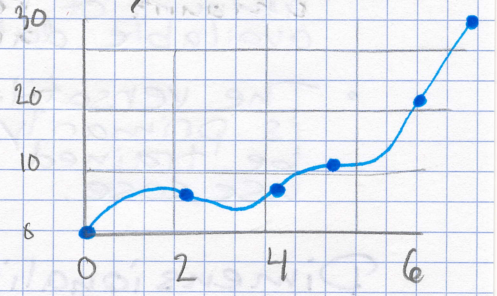
- Sampling is the process of selecting a small portion of the corpus that gives an accurate representation of all of the data

Variety of ML Algorithms

Polynomial Regression

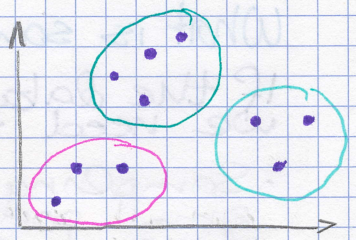
- Finding a line of best fit on a set of data points
- However, the line doesn't always have to be straight (known as linear regression) and can be generalised as polynomial regression
- Not to be confused with logistics regression - this is something very different.

Polynomial Regression (Overfitting)



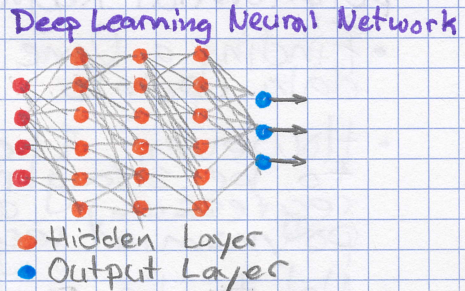
- Overfitting is when we effectively make the model too complicated for the system
- An example may be hypothesising that the weight of a ball is proportional to the size, color, smell, and the name of the shop attendant you bought the ball from.
- To alleviate overfitting, we introduce regularization.

Clustering



- Process of creating groups that exhibit similar characteristics within the data
- It is predominantly an unsupervised Machine learning algorithm
- K-Means is a clustering technique that we will implement.

Deep Learning



- It is a supervised Machine Learning method and requires a large amount of training data
- Used almost exclusively for image recognition due to their high accuracy and the vast amount of readily available data
- The versatility of a Neural Network is primarily due to its ability to be trained very accurately to certain use case.

Dimensionality Reduction

- Method for removing unwanted dimensions from a dataset
- It is extremely useful to help determine correlations in the data and reduce the complexity of the dataset.
- PCA - transforms the data based on the size of the eigenvalues

Simple Classification Task

Why is classification useful?

Ideal for certain charts and images (or flow charts)

Object Recognition
|
Fraud Detection
|
Health Diagnosis

Object Detection

- Self driving cars need to be able to identify different objects
- Postal service scans thousands of letters and directs them to the right location

Fraud Detection

- It's vital that banks can detect if a transaction is fraudulent
- Banks may also find it useful to classify certain customers into different classes.

Health Diagnosis

- New health diagnosis apps need to be able to classify the ailment corresponding to your symptoms.

Where is classification used?

- We will look at the postal service example of classification
- Goal - Classify an image of the handwritten digits into one of 10 classes (0-9)
- Input - 28x28 gray scale image
- Output - A single digit

- An Idea! If two images are the same, then subtracting their pixel values in an element-wise manner and summing the square of the values should give zero.
- If we have a set of example images then we can perform the above step on each example and choose the class of the lowest value to be the chosen class.
- This is effectively a K-Nearest Neighbor (KNN) classifier with K=1.

So classification is easy?

Not quite...

There are several issues with this algorithm

- What if the image is rotated?
- What if the pen is thinner or larger than the example?
- What if our example data is bad?

Why Choose C++?

Object Oriented
Code can be structured in an effective manner

Speed
C++ is fast!

Libraries
There are a plethora of open source C++ machine libraries