MLGym Onboarding Session

- Welcome
- Objective
  - cover the basics skills need to do machine learning
  - hopefully you will be able to handle the titanic dataset by end of session
- This is a crash course
  - I advise you to supplement the information you receive today

Section 1

Setup and Overview of machine learning

What is machine learning

- This is not the terminator
  - people have many crazy ideas about what this is.
  - It won't suddenly turn into a human killing robot from the future..
  - this is not magic, or scifi…
- It is an applied form of statical modeling
  - Just maths, data and statics
- It is the core set of techniques, models and algorithms that allow a computer programs to flexibly learn and adapt to a set of tasks and its changes over time
  - NOT hand tuned rules!

The basic ML process

- Gather data
- Analyse data
- Clean data
- Build model
- Fit model
- Deploy model

Installation

- Please install python and pandas
- Ubuntu

```
sudo apt-get install python python-dev
sudo apt-get install python-pip
sudo pip install numpy sklearn pandas
sudo pip install jupyter
sudo pip install tensorflow keras
```

- MacOS

```
TODO confirm

sudo easy-install python python-dev
sudo easy-install install python-pip
sudo pip install numpy sklearn pandas
sudo pip install jupyter
sudo pip install tensorflow keras
```

Section 2

Introduction to Data Anaylsis

Why its important

- Machine learning is applied statical anaylsis,
- If you dont understand the data your handling you have very little chance of succeeding in modeling it
  - Note that “understand the data” is not the domain specific understanding but the “data scientist” understanding of a dataset.

Glossary

- Feature
  - Simply an input. A column of the input table, A pixel in a certain location etc
- Numerical data
  - Number based data
  - Examples: ages, weights, etc
- Class data
  - A set of data that consists of labels
  - Be careful some class data **looks** like numerical data
  - Examples: a rank or title, sex, eye/hair color etc
- Textual/Raw data
  - A raw blog of text.
  - it that may be mined for something useful
  - It may be injested via a sequance based model
  - Examples: peoples names, address etc

Techniques we will cover

- Check for missing data
- Computing mean, variance
- Computing percentile and quartiles of data
- Pivot Tables
- Rendering Histograms
- Plotting boxplots

Over to the Juypter Notebook

Please open your laptops and work along with me

Section 3

Introduction to Data Cleaning

Data rules of thumb

- The curse of dimensionality
  - more features the worse fit
  - Modelling confusion
- The significance rule of thumb:
  - 30(ish) representative samples are needed for a significant statistical step/sampling
- Problems in the data are often the source of model quality problems

Examples of significance effect

- with 1 input feature and 300 samples
  - your accuracy of fit is 30/300*1 = 10%
- with 2 input features and 300 samples
  - your accuracy of fit is 30/300*2 = 20%
- with 3 input features and 3000 samples
  - your accuracy of fit is 30/3000*3 = 3%
- The 30 number is a wet thumb number.. it varies a lot in practice

Techniques we will cover

- Merging data
- Dropping pointless data
- Converting categorical data into a usable form
- Handling missing data
- Segmenting data

Section 4

Introduction to Modelling

What is Modeling

- Modelling is the process of taking a dataset and trying to shape a function so that it produces results that match the dataset.
- By definition the model we are trying to fit on to the data is often a simplification and generalization of what we believe to be the underlying truth of the dataset.
- Ideally we want the model to generalize

Modeling loss

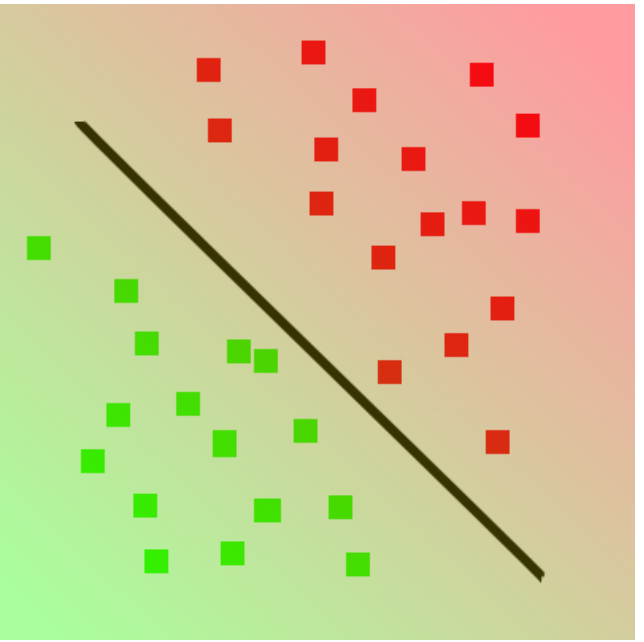
- E= mc^2
  - right?
  - so a photon - speed of light and is massless.. and by the above equation that means it has no energy..
  - yea something is wrong
  - E^2 = m^2c^4 + p^2c^2
- What we are doing is an inductive process, the conversion from a specific set of examples to the general form.
  - this means it can be incorrect (often drastically so)

Deep frying data

- even cardboard is eatable if deep fried
- any model will to some extent fit the provided data
  - Newtonian physics vs relativistic physics
- Try to avoid making outrageous claims because a model you think is perfect fits the data with a very high accuracy
  - Ie we could have missed measuring a critical datapoint etc

Types of models

- Classification Model
  - These models make a choice of class, Is the data A, B or C
- Regression Model
  - These models attempt to map a function y = f(x)
- Both models have a fundamental interrelationship
  - A classification model can be made using a regression model that outputs how “strong” a fit to class “A” the inputs are

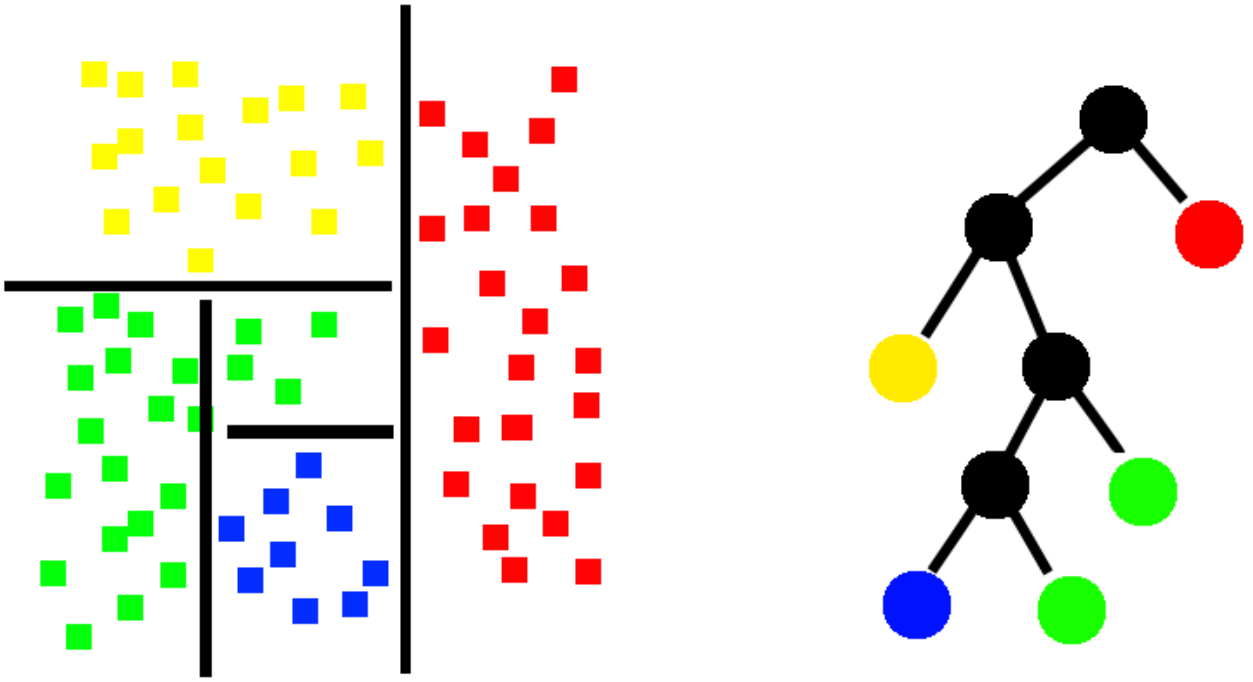


Supervised vs Unsupervised learning

- Supervised
  - Supervised models use a known set of output labels and attempt map from inputs to outputs
  - y = f(x0,x1...)
- Unsupervised
  - Unsupervised model use a set of data that has no clearly labeled outputs
  - f(x0,x1...)
- Both learning methods have a fundamental interrelationship
  - y = f(x0,x1…) <=> f(y,x0,x1…)
  - Hint: this idea is the key to more advanced data cleaning techniques

Random forest

- Random forest is a rather simple “if a then x” decision tree it looks at the spread of data and makes a decision that divides the data with the largest information gain as possible



Over to the Juypter Notebook

Please open your laptops and work along with me