Introduction to machine learning

Machine learning is a data analysis method that automates analytical model building by allowing computers to find insights in data in a way they are not explicitly programmed to using iterative methods.

Machine learning cycle is:

* Data acquisition
* Data cleaning
* Split data into test and training sets
* Train the model created
* Test the model
* Deploy

Machine learning is used in different areas from fraud detection, recommender systems, customer classification etc.

There are 3 main types of ml algorithms:

* Supervised learning: using previously labeled data to try to predict a label based off known features.
* Unsupervised learning: using unlabeled data to group similar data points based off of features.
* Reinforcement learning: training a model to perform a task based on prior experience.

# Supervised learning

* The algorithm is trained using labelled data and then used to predict future events from unlabeled data.
* Uses regression, classification etc.
* Is used when historical data can predict future events.

# Unsupervised learning

Uses unlabeled data for exploration and finding meaning in it

Popular methods include k=mean clustering, nearest neighbour, etc.

# Reinforcement learning

Used in gaming and similar fields that need to act on a decision

The algorithm learns by trial and error and subsequently choosing the best action (policy).

# Scikit learn

For this course, scikit lean will be used

* To install use pip install scikit-learn
* To use scikit learn, you need to import it with the general form from sklearn.family import Model where family is the family such as linear\_model and Model the specific estimator such as LinearRegression.

An example with supervised learning is below:

* To split data into test set and training set, you import train test split from cross validation with from sklearn.cross\_validation import train\_test\_split. Then, pass X\_train, X\_test, y\_train, y\_test =train\_test\_split (X, y, test\_size=a) where a is a decimal reflecting the proportion of the data to be set as test data.
* You can call each of the data sets i.e. X\_test, y\_train etc.later
* To train your model, use the model.fit (X\_train, y\_train)
* You can then test the model with model.predict (X\_train) to get the predictions.
* You can evaluate model by comparing prediction to correct values.

The scikit learn module tries to ensure uniformity across all models by making the methods similar. The methods are as follows:

## All estimators

* model.fit which accepts 1 and 2 arguments for unsupervised and supervised learning respectively.

## Supervised estimators

* model.predict () which accepts one argument, X\_new.
* model.predict\_proba () which returns the prediction with the highest probability.
* model.score () which returns a value between 0 and 1 and the higher the result, the better the fit.

## Unsupervised estimators

* model.predict ()
* model.transform() whoch transforms data into a new basis and accepts only one argument, X\_new
* model.fit\_transform () which transforms and fit the data simultaneously.

# Choosing an algorithm

* download scikit learn algorithm cheat sheet.

