Intro to Machine Learning

Course Modules

1. Introduction and Regression
   1. Simple Linear Regression
   2. Multiple Linear Regression
   3. Regression Trees
2. Classification
   1. Logistic Regression
   2. K Nearest Neighbours
   3. Support Vector Machines
   4. Multiclass Predcition
   5. Decision Trees
3. Clustering
   1. K-Means
4. Final Project

Keywords

Supervised Learning, Unsupervised Learning, Classification, Regression, Clustering.

Key Skills

* Regression
* Classification
* Clustering
* Scikit Learn
* Scipy

Projects

* Cancer detection
* Predicitng Economic Trends
* Predicting Customer Churn
* Recommendation Engines

What is machine learning?

Machine learning is the subfield of computer science that gives computers the ability to learn **without being explicitly programmed.**

Before machine learning, to classify for example dogs, a list of features would be created (tails, legs etc) in a process called feature extraction and then rules would be written to describe and classify those features. However, as the list of features becomes too large or complex, the ability for programs to be written to classify those objects becomes too complex. Machine learning algos write those classifications themselves by iteratively learning and correcting a set of features to provide a classification.

**Different ML Techniques**

* Regression/Estimation
  + Is used for predicting Continuous Values (House prices, CO2 from an engine)
* Classification
  + Is used for predicting the item class/category of a case. (Cancer Cells, Customer will churn)
* Clustering
  + Finding structure of data (customer segmentation etc)
* Associations
  + Associating Frequent co-occurring items/events (Grocery items that are often bought together)
* Anomaly Detection
  + Discovering outliers
* Sequence Mining
  + Predicting the next event (Markov Model, HMM)
* Dimension Reduction
  + Reducing the size of data (PCA)
* Recommendation Systems
  + Recommending Preferences

**What is the difference between AI, ML and DL?**

AI tries to mimic the cognitive functions in humans.

Machine Learning covers the statistical part of AI. It teaches the computer to solve problems by running through an iterative process.

Deep Learning is a specialized version of Machine Learning.

**Supervised vs. Unsupervised Learning**

Supervised Learning basically means to teach the model

We teach it by training it with “Labelled data”. Sine we already know the outcome of the data, we can give the model this data to start with.

The names of rows are the datas **Attributes**

A row is referred to as an **observation**

The colums data is **features**

**Every data point will have an attribute a feature and a label**

Data can either be numerical or categorical

**Classification vs Regression**

There are two types of supervised learning: Classification and Regression

* Classification is the process of predicting a discrete class label or category.
* Regression is the process of predicting a continuous value in classification.

Unsupervised learning is as it sounds. The model works with the data that is unlabelled to discover things that may not be visible to the human eye.

Unsupervised learning algorithms tend to be more complex than supervised algorithms.

Dimension Reduction, Density Estimation, Market Basket Analysis and Clustering and the most commonly used ML Techniques. Dimension Reduction/Feature Selection play a large role in reducing redundant features to make classification easier. Market Basket Analysis works on the premise that if you buy certain things, you will buy certain other things. Density Estimation is used to find structures within data.

Clustering is a very popular unsupervised learning technique, used for grouping data points that are somehow similar. Cluster analysis, can be used to segment customers etc etc

**Regression**

The process of regression looks to predict a continuous value (what is the relationship between things?). In regression there are 2 types of variables, dependent and independent.

Y, the dependent variable is the state target or goal that is looking to be predicted. X is one or more independent variables can be seen as the causes of those states.

We are looking to identify Y as a function of X.

Y must be a continuous variable and not discreet.

X however can be measured on a categorical or continuous measurement scale.

There are two types of regression model:

* Simple Regression: When one independent variable is used o describe one dependent variable.
  + Simple linear regression
  + Simple non-linear regression

The linearity of the relationship is based on the relationship between the two.

* Multiple regression: When multiple independent variables describe one dependent
  + Multiple linear regression
  + Multiple non-linear regression

**Regression Algorithms**

* Ordinal regression
* Poisson Regression
* Fast Forest Quantile Regression
* Linear, Polynomial, Lasso, Stepwise, Ridge Regression
* Bayesian Linear Regression
* Neural Network Regression
* Decision Forest Regression
* Boosted Decision Tree Regression
* K Nearest Neighbours

**Simple Linear Regression**

So the independent variables can be measured on a continuous or categorical scale.

* Plot variables on a scatter plot
* Draw a line through the data (line of best fit)

The line of best fit is representative of the polynomial

Y hat is the dependent variable, x one is the independent variable

theta one represents the gradient of the line

theta zero is the intersect

they are the coefficients of the linear equation

theta 0 and theta 1 must be calculated to find the best fit for the data.

For example:

If y (y being an actual observation) = 250

And y-hat is 340

Then y – y-hat

250 – 340

= 90

The error is 90 or residual error

The residual error is the distance from the observation to the fitted regression line

The mean of all residual errors shows how well(poorly) the line fits to the dataset.

Mathematically this is represented by the equation mean squared error

The goal is to minimise the error to find the line of best fit.

Rearranging the equations theta 0 and theta one can easily be calculated.

First the mean of the independent and dependent columns must be calculated.

Maths note: Bars are means, hats are estimates!

**Model Evaluation Approaches**

The goal of a model is to accurately predict an unknown case.

As such the model needs to be evaluated.

There are 2 approaches:

* Train and test on the same data set
* Train/Test split

How can the accuracy of a model be trusted?

A selection of the data can be used for testing.

The whole dataset is used for training the model and then a subsection of that dataset.

A small portion of the data is not labelled.

We pass the feature set to the model to train then compare what values it predicts for the test set vs. the actual values.

**Regression evaluation Metrics**

“Error” is normally used to denote accuracy and is simply the average difference between the predicted values and the actual values.

**Training Accuracy vs. Out-of-sample accuracy.**

Training and testing on the same dataset normally produces a high training accuracy.

Training accuracy is the correct predictions the model makes when training the model.

It is not necessarily a good thing as often the means the model is ‘overfitted’ meaning the model is very specific to the data it has been given and possibly wont work in a generalised use case.

Out-of-sample accuracy is the percentage of correct predictions the model makes on data that the model has not been trained on.

Train and test on the same data set often produces low out-of-sample accuracy due to overfit.

Out-of-sample accuracy is the actual goal of the model.

TO improve this, a train/test split is used. In a train test split a portion of the data is not used for training and is then used to test the model for accuracy based on the training the model on the rest of the data.

Once you have tested the model for accuracy, use the testing set to train the model again because why waste data if it is accurate and useful.

The train test split is however highly dependent on the dataset on which it is trained.

**K-Fold Cross-validation**

In K-fold cross-validation the model is ‘folded’ or divided in K sections. Eg. 4 folds of 25%. That 25% is used for testing and the other 75% is used for training. The process is then repeated on the next fold.

The accuracy for each fold is calculated and the the mean av. Of all folds is the accuracy of the model.

**Evaluation Metrics for Model Evaluation**

MAE = Mean Absolute Error

MSE = Mean Squared Error

RMSE = Root Mean Squared Error

RAE = Relative Absolute Error

RSE = Relative Squared Error

R^2 = Isnt an error but shows accuracy of the model.

Error is the diff between the datapoint and trendline.

**Calculations**

R-Squared (Coefficient of Determination) quantifies the proportion of variance in the dependent variable that is explained by the independent variable in a regression model. This provides an indiciation of how well the regression model fits the data.

R-Squared value ranges for 0 – 1 where 1 indicates the models explains all variability and 0 means it does not explain any.

Variability is the spread of data points around the central value or the deviation from the mean.

Correlation refers to the strength of and direction of the linear relationship between two variables. In essence indicating how closely the data points cluster to the trend line correlation is often calculated using Pearsons correlations coefficient which ranges from -1 to 1 either a complete negative correlation of total correlation.