Data Science Campus

Machine Learning Fundamentals

Alex Noyvirt

Claus Sthamer

Data Science Campus, ONS

26 January 2022

Data Science Campus

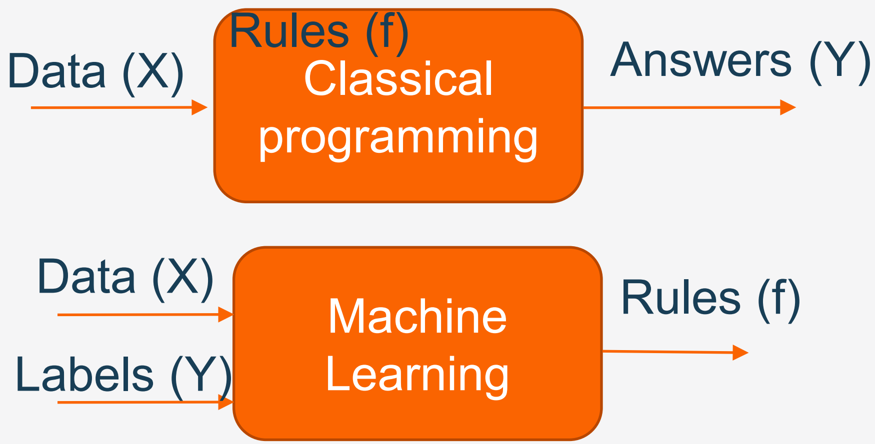
Dubai EXPO 2020

Outline

- What is Machine Learning
- Types of ML
- Supervised ML
  - Training & Test Data
  - Supervised ML Algorithms
  - Data Preparation
  - Over & under-fitting
  - Hyperparameters
- Unsupervised

What is Machine Learning

In, 1959 Arthur Samuel defined machine learning as a *“Field of study that gives computers the ability to learn without being explicitly programmed”*



$$Y=f(X)$$

Instead of humans formulating all the rules needed, we leave the ML algorithm to find all the rules needed to carry out the task.

Types of Machine Learning

- **Supervised – there are labels**
- **Unsupervised – no labels available**
- Semi-Supervised Learning
- Reinforcement Learning

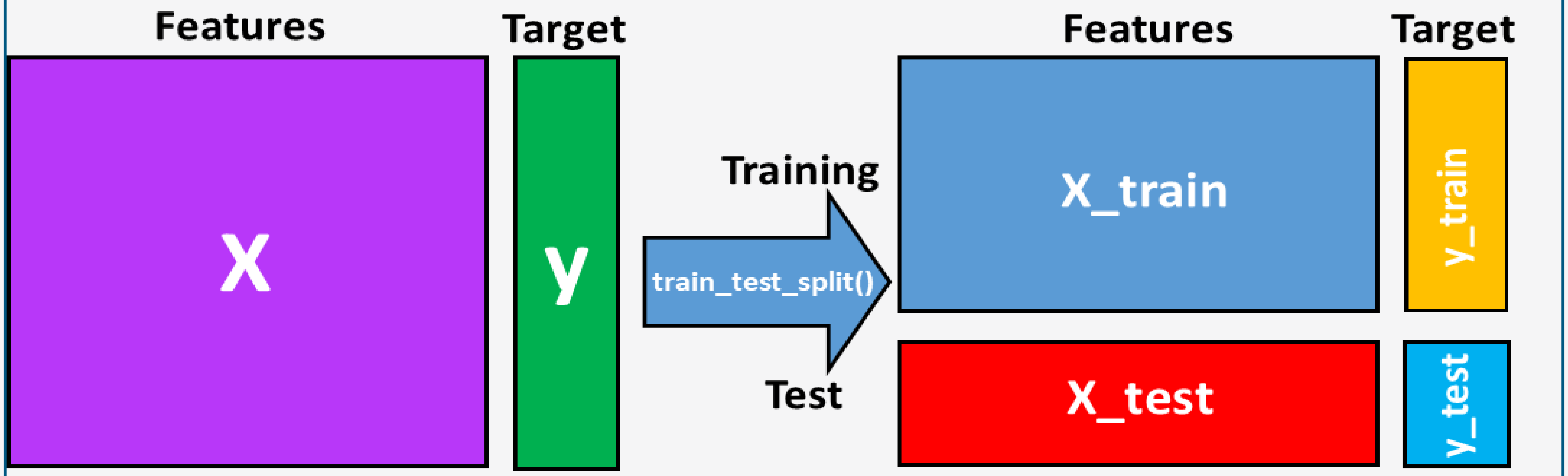
Supervised learning – some terminology

Target, Labels

Independent variables

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Dependant variable

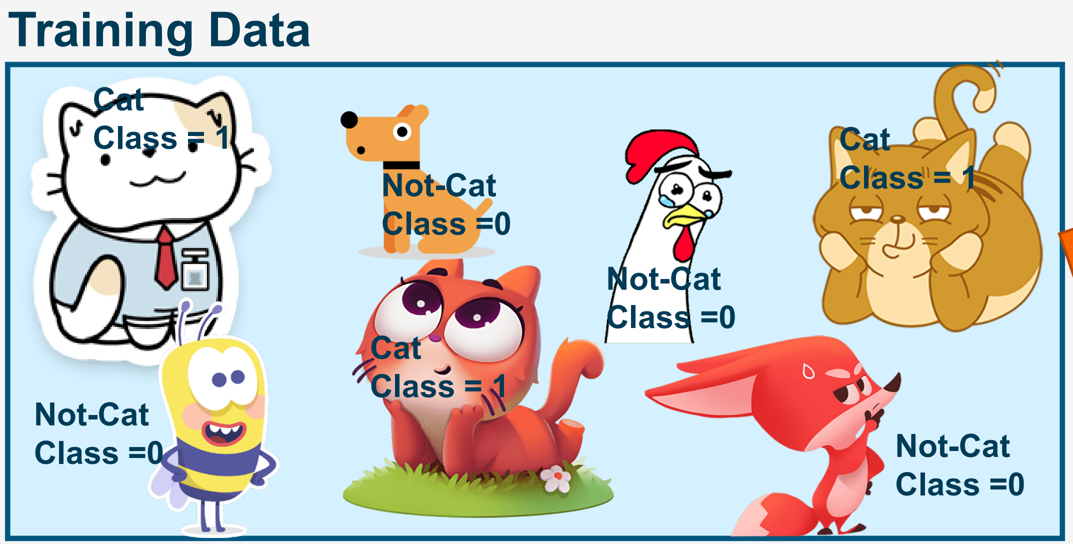
Supervised Learning – Training and Test Data

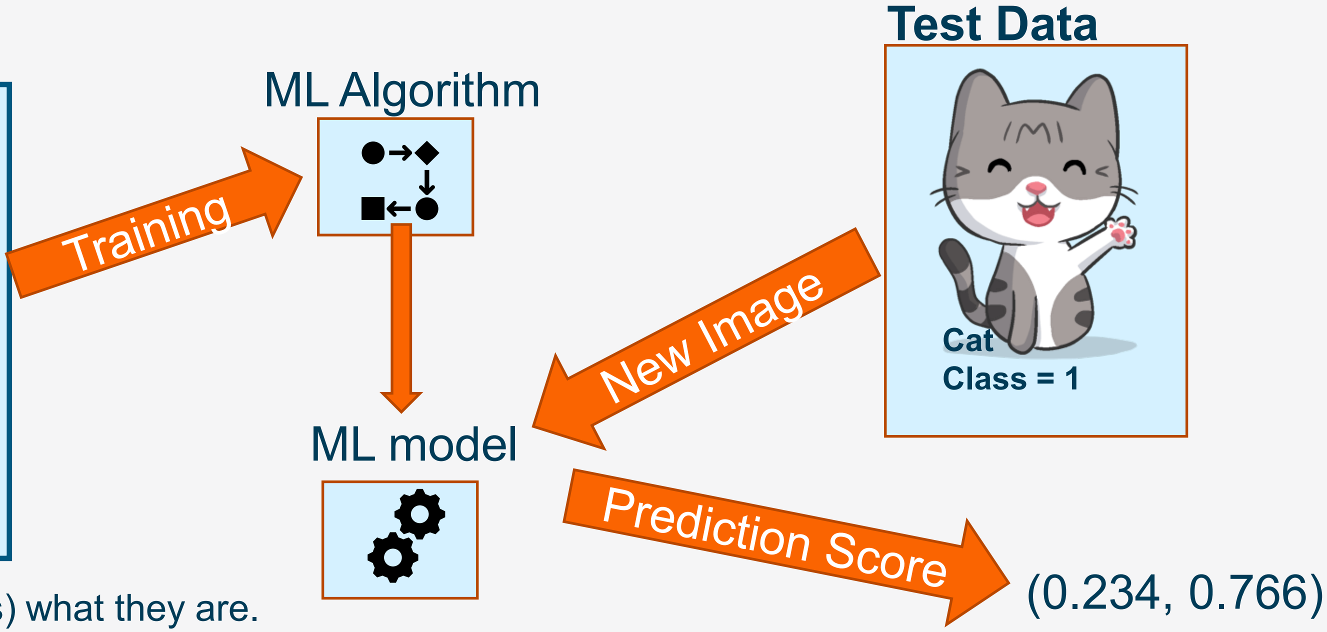



```
df_pre_edit_train, df_pre_edit_test = model_selection.train_test_split
    (
    df_pre_edit_change,                      # Input Features with Target
    test_size = 0.3,                         # Proportion of Test Data, Default = 0.25
    stratify=df_pre_edit_change['Change’]    # Stratify on the Change column (Target)
    )
```

Supervised Learning – an example

How to recognise a cat? But not just one cat but all cats? What are the rules?





A ML algorithm can learn from pictures as long we tell it (Labels) what they are.

ML can make an inference of the class of new pictures, it gives a score for the most likely class

Prediction Threshold for Cat class = 0.55
Cat Class score > Threshold → Cat

We do not know how we recognise things, but we are very good at it
→ The most powerful ML methods are the least interpretable

Chapter 0 – ML Introduction

Please follow:

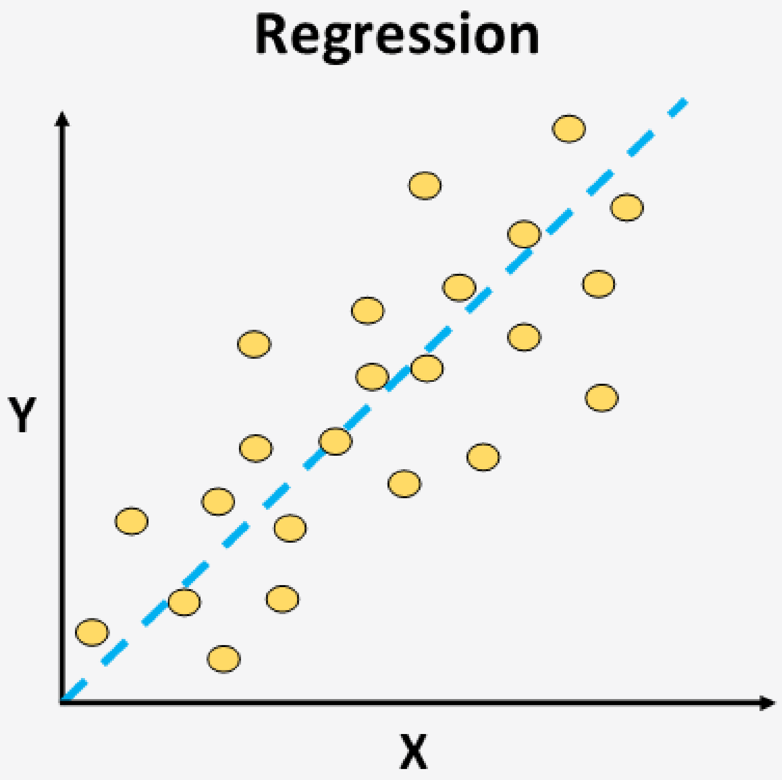
https://rb.gy/hmlc76

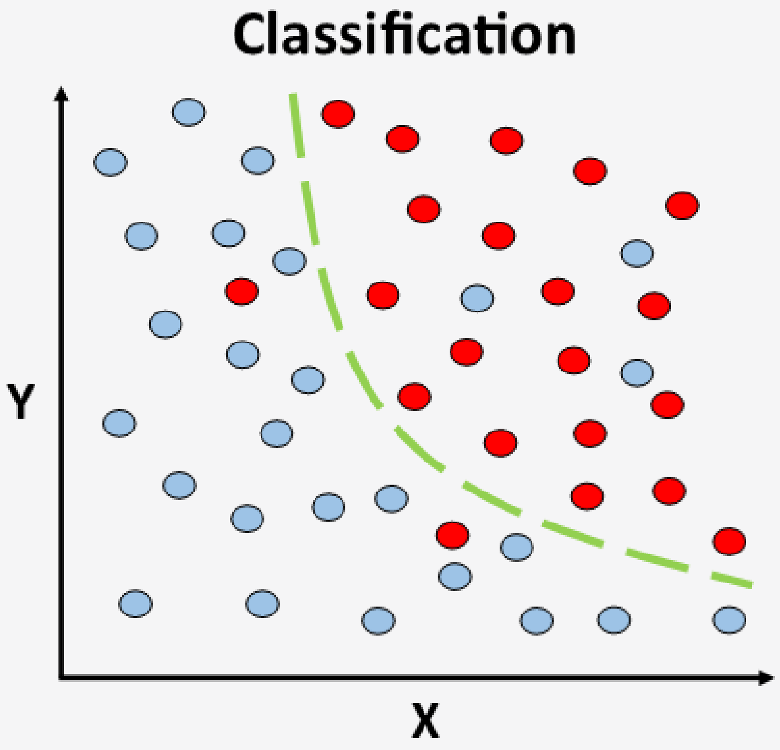
Supervised Learning - ML Algorithms

No Model is perfect → find the best fitting one → minimise error

- Regression & Classification
- Linear Regression
- Support Vector Regressor
- Logistic Regression
- SVM
- KNN
- Decision Trees

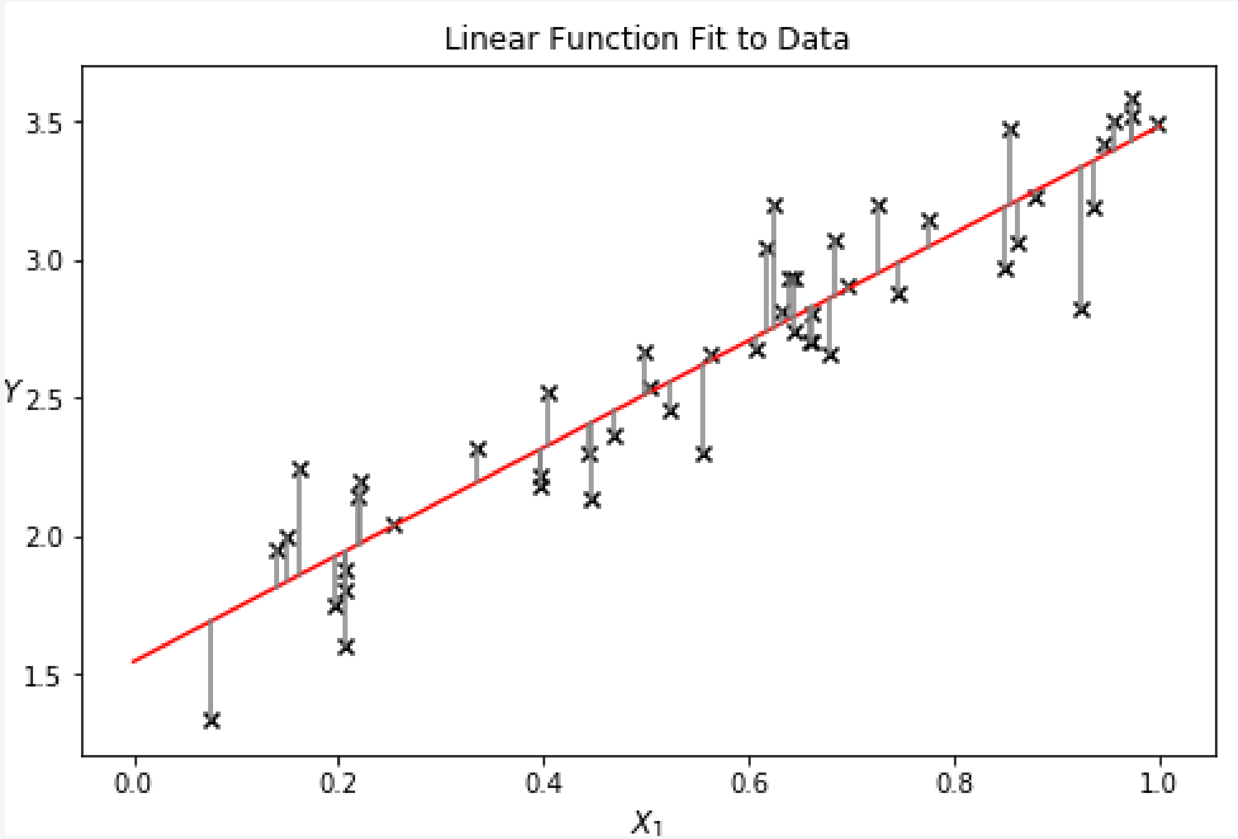
Regression and Classification





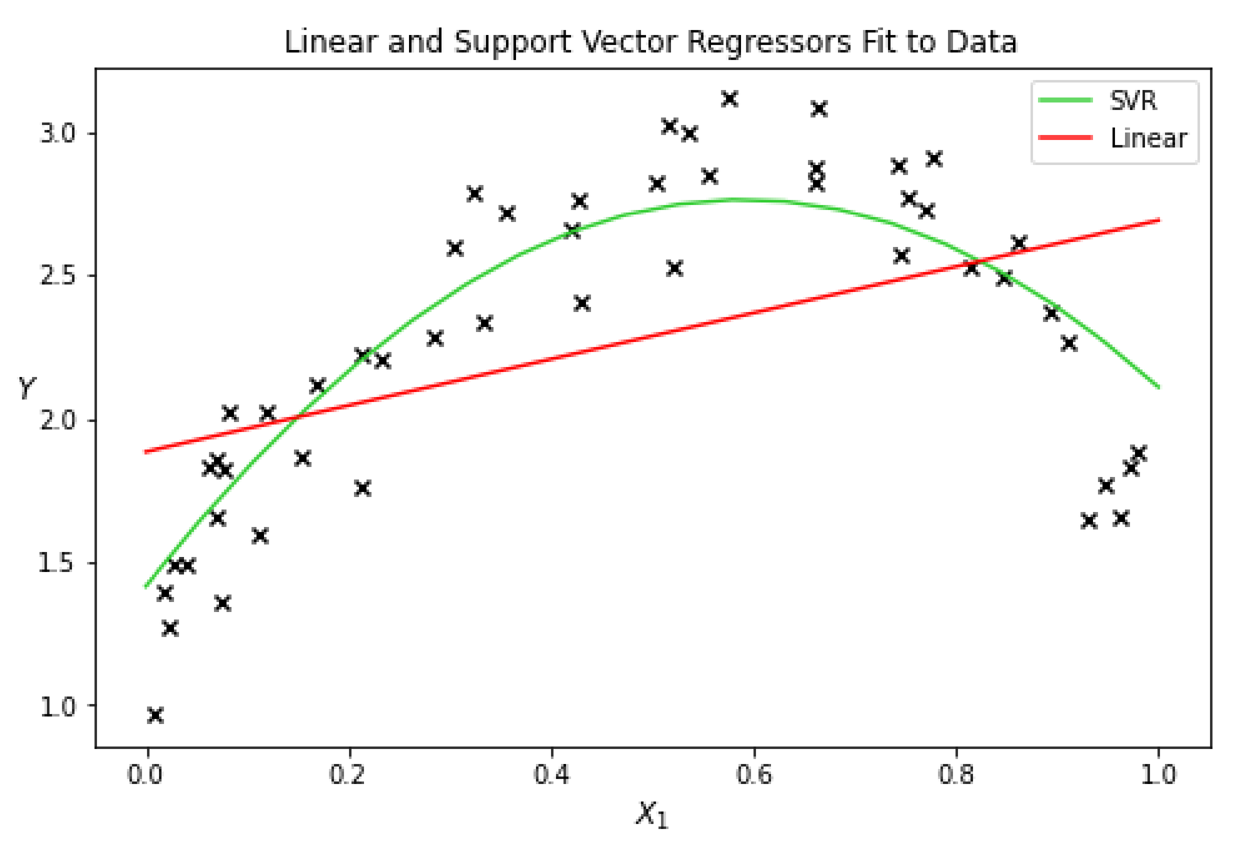
Linear Regression

- Regression analysis is used in statistical modelling for estimating relationships between independent variables and dependent variables
- Y = a + bX
- It is the simplest Machine Learning algorithm. It is used to predict values of a continuous variable, e.g. price, age or salary
- Find the best fit line.
- Least square estimation for estimation of accuracy
- Multi-Dimensional → Hyperplane



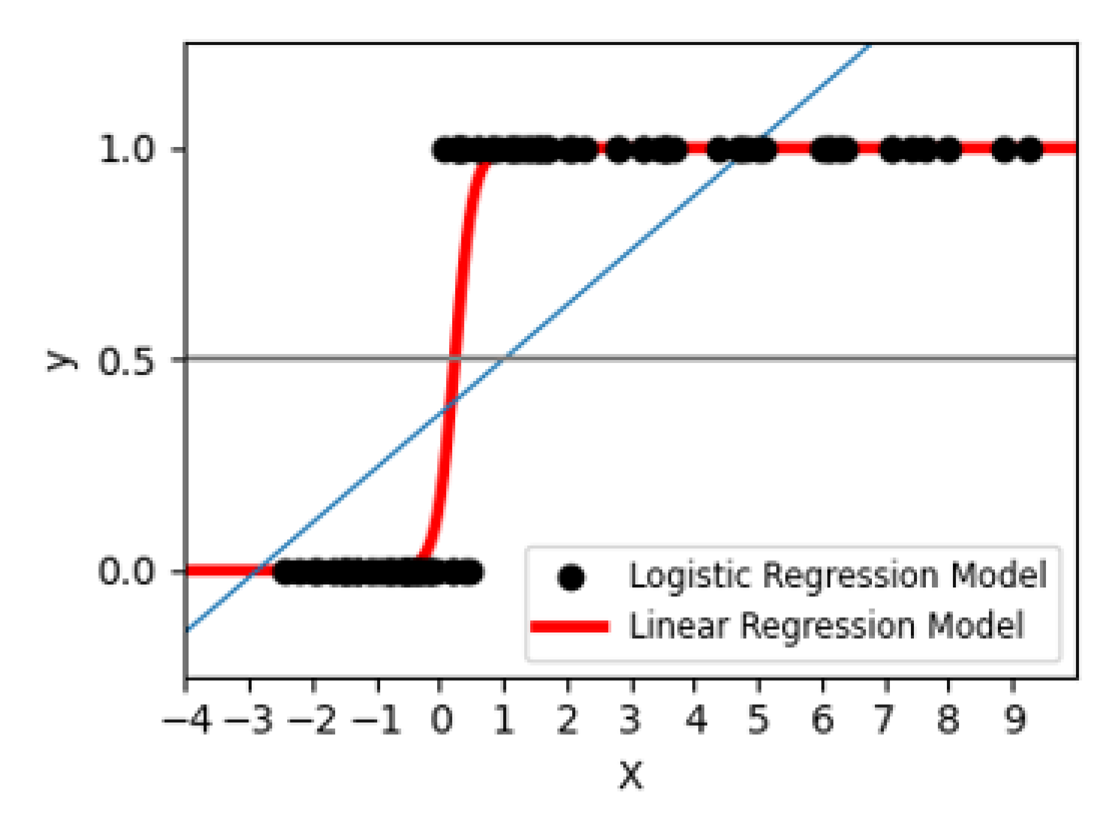
SVR - Support Vector Regressor

- Linear model is inflexible
- SVR is a better fit for these data points
- Non-linear fit
- Needs more training data



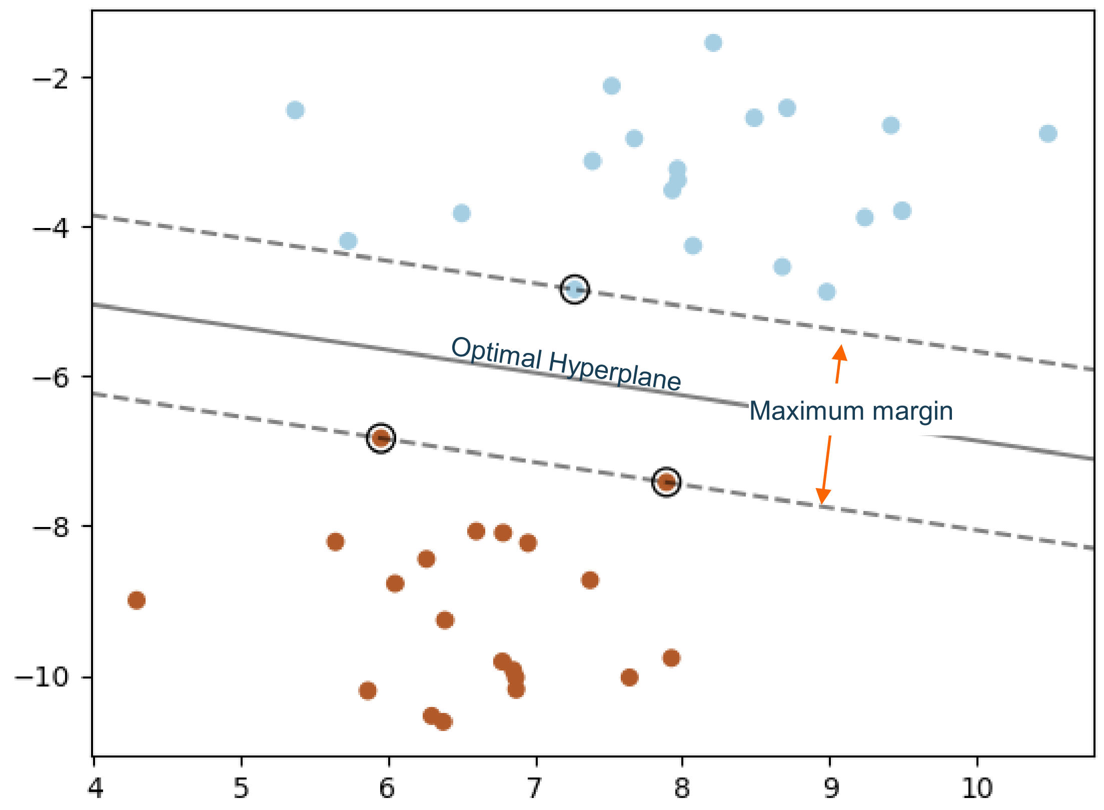
Logistic Regression

- A Linear function would be a bad model for these data points
- Logistic Regression mainly used for binary classification
- The output can only between 0 and 1, e.g. yes/no or 0/1, Cat / Not-Cat
- Find the S-curve (Sigmoid Function) to classify the sample
- Non-linear transformation of Linear Regression
- Linear Regression predicts the outcome
- Here we predict:
  - Ln[p/(1-p)] = a + BX   (Log of the odds ratio)
  - P is curve between 0 and 1
  - If p > 0.5 → Prediction is 1, if p < 0.5 → prediction is 0



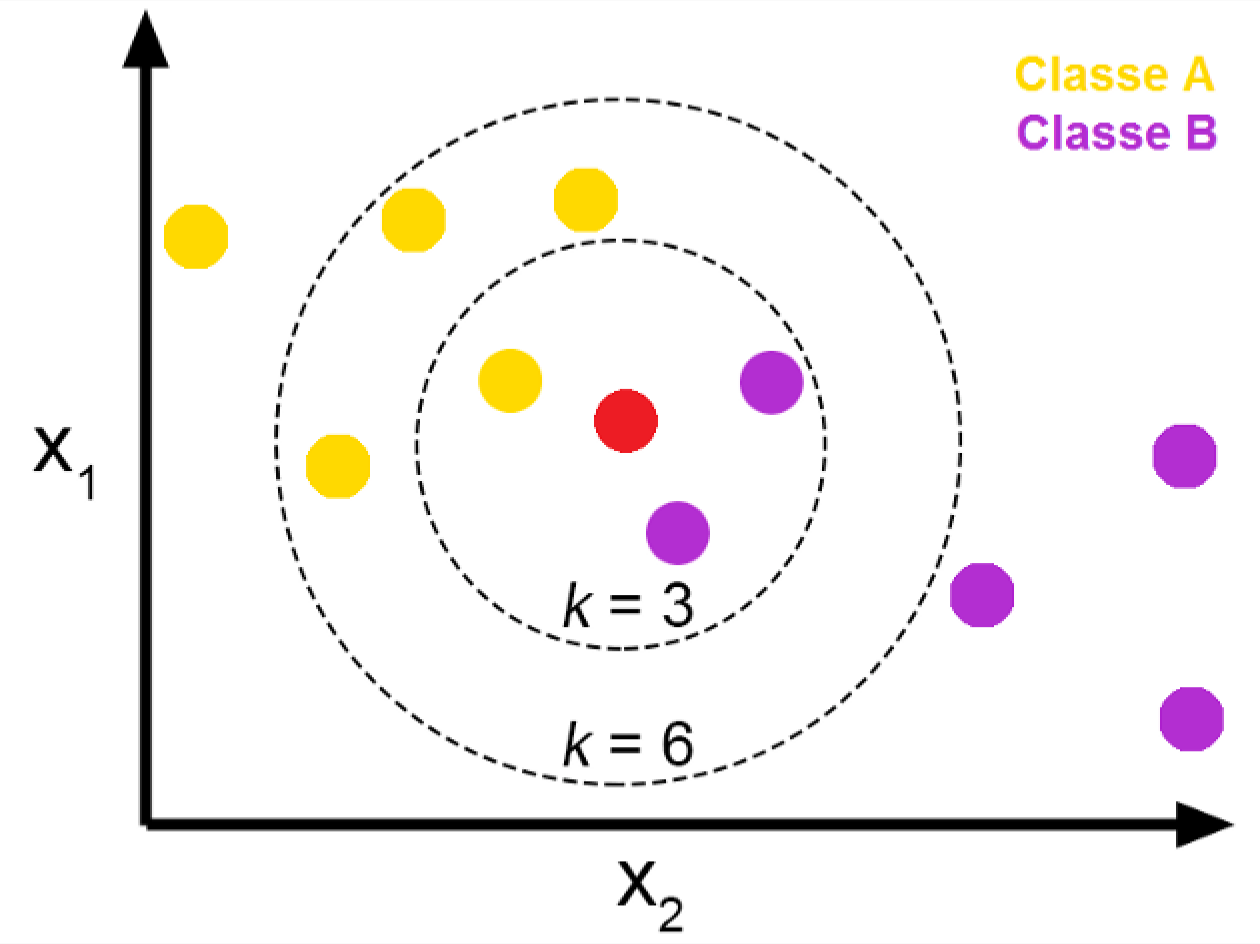
SVM – Support Vector Machines

- Linear model for classification
- Creates line or hyperplane in an N-dimensional space to separates data into classes
- N is the number of features
- Find Hyperplane with maximum distance between data points of both classes
- Maximizing the margin ➔ future data points can be classified with more confidence
- Data points on the maximum margin are called: Support vectors
- Support vectors influence position & orientation of the Hyperplane
- Hyperplane = Decision Boundary



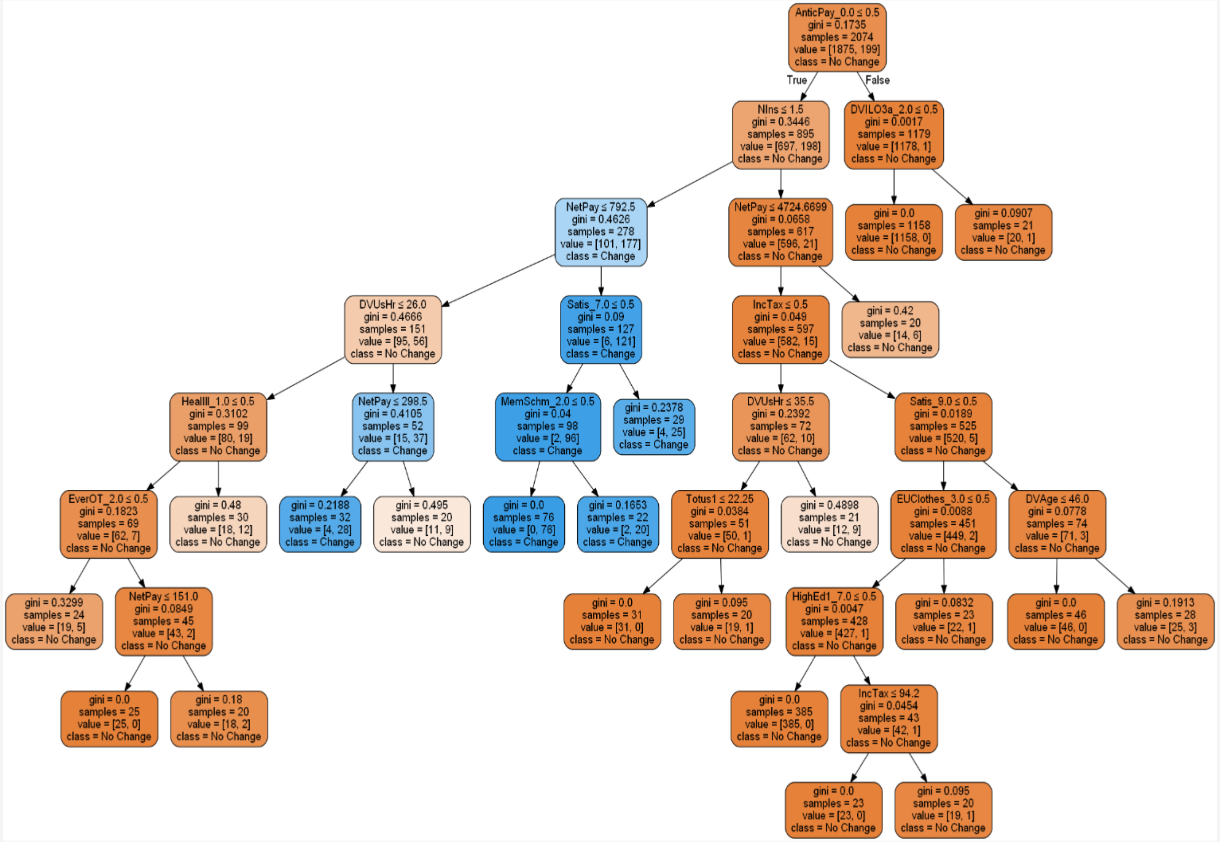
KNN – Lazy Learner

- K-Nearest Neighbours
- Finds distance to K closest data points
- Smaller distance → more similar
- Classification - votes for most frequent label of K neighbours
- Regression - average the labels
- Calculates distance between every data pair
- High calculation costs in higher dimensional space and sample size – square of sample size



Decision Trees

- Classification and Regression
- Tree-like model of decisions
- Nodes, branches and leaves
- Learning of simple decision rules: if-then-else
- The deeper the tree the more complex rules
- Splits are based on reducing Classification error



Random Forest

- Many decision trees
- Random sub-samples, random features
- Voting score

Unsupervised Learning

- **Clustering**: splitting or partitioning data into groups according to similarity.
- **Latent variable models**: discovering 'hidden' constructs based on observed data.
- **Dimension reduction**: reducing the number of features in a dataset, while retaining as much information as possible.
- **Outlier detection**: finding unusual data values.

Session Summary