University of Lancashire

Introduction to Machine Learning for Data Science

Karen O'Shea

Where opportunity creates success

Learning Objectives

University of Lancashire

Describe common examples of Machine Learning algorithms for Data Science

Apply algorithms to assist with research question(s) design

Semester 2: Next Steps…

- Build a Data Science Model…
  - Algorithms include: Linear/Logistic Regression; classification; predictive in nature; could be designed to seek new insights
- Train, test and validation datasets required
- Evaluate outputs and build stories

Why Machine Learning?

University of Lancashire

- With machine learning we can do the following much faster:
  - Determine whether images contain human faces – image recognition
  - Predict whether an ad is appealing or personal enough for a user to click on it – predictions
  - Create accurate YouTube video captions – speech recognition, speech-to-text translation
  - Whether a transaction is fraudulent
  - Whether an email is spam

Data Science, Data Analysis and ML

- Data Science/Data Analysis
  - generate insight and find patterns
- Data Science/Machine Learning
  - find patterns and use learned abilities on new data
- Testing/validity of dataset(s)
  - Evaluate and confirm

Predictive Model Design Steps

University of Lancashire

Data Analysis:

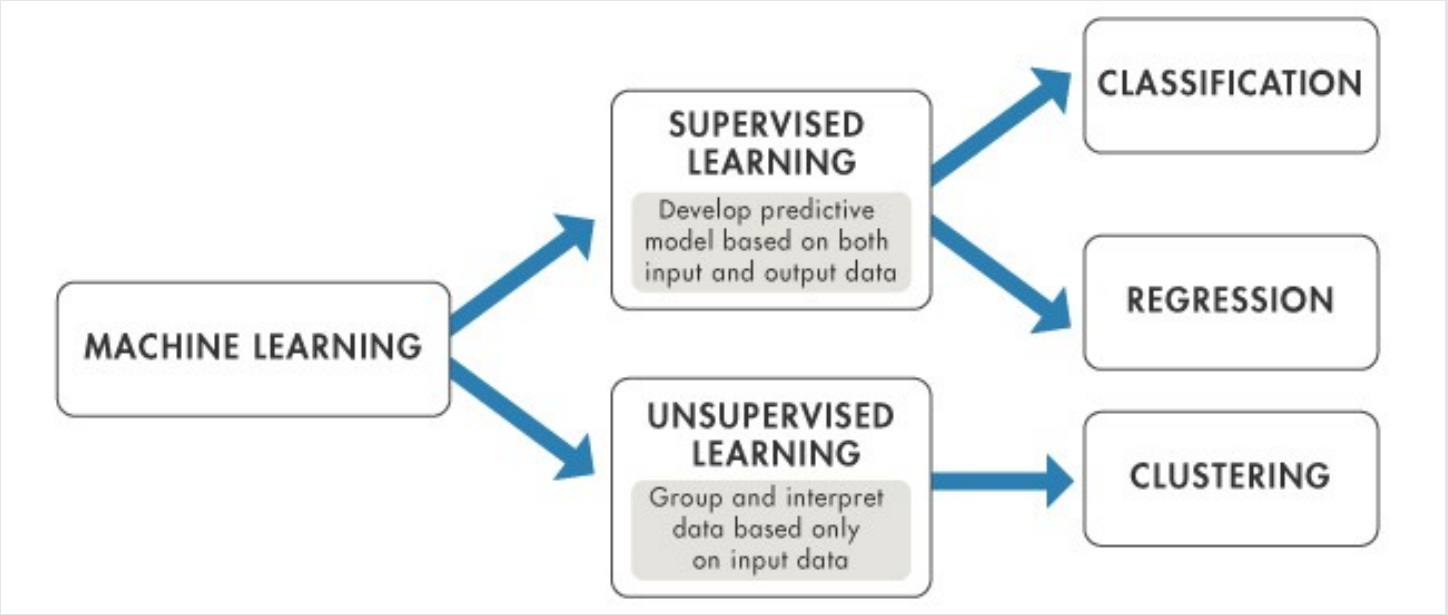
- Data review, data cleaning, including attending to missing data
- Convert inputs (categorical variables)

Data Science:

- Import required modules – Scikit-Learn
- Classification using chosen model
- Cross validation and accuracy scores
- Metrics

Supervised and Unsupervised Learning

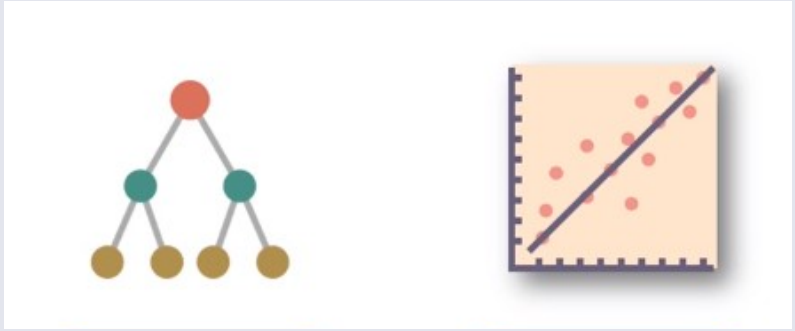
University of Lancashire



Machine Learning Algorithms

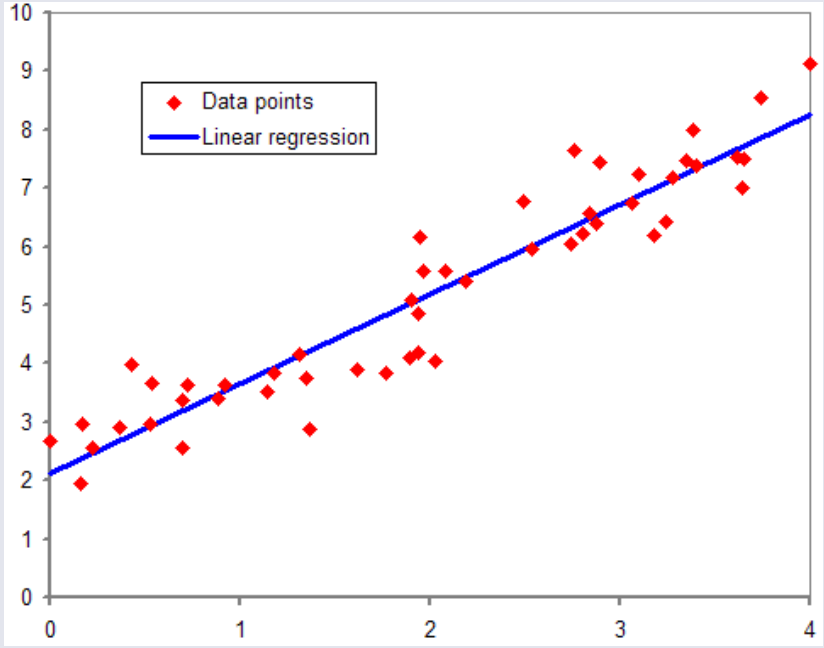
Question:

- Linear Regression
- Logistic Regression (Classification)
- Decision Tree (Random Forests)
- Unsupervised learning (Clustering)



Discuss: What do we know so far from our research?

Linear Regression – Supervised Learning



Equation of a line:

y = Mx + C

M = gradient
C = Intercept

Line of ‘best’ fit

Linear Regression Example

University of Lancashire

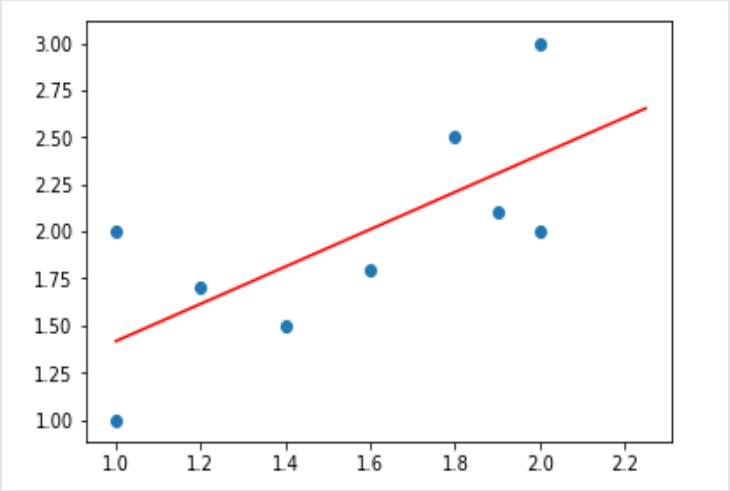
```
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
x = np.array([[1, 1], [1, 2], [2, 2], [2, 3],[1.2,1.7],
              [1.4,1.5],[1.6,1.8],[1.8,2.5],[1.9,2.1]])

x_values = x[:,0].reshape(-1,1)
y_values = x[:,1].reshape(-1,1)

#fit x, y data to a linear regression model
model = LinearRegression().fit(x_values,y_values)

#generate some random number for x and use the model to predict y
random_number = np.arange(1.0,2.3,0.05).reshape(-1,1)
y_pred = model.predict(random_number)

plt.scatter(x[:,0],x[:,1])
plt.plot(random_number, y_pred,color="red")
plt.show()
```



SCIKIT-LEARN

University of Lancashire

- Foundational ML package
- Many useful algorithms
- Classification, regression, and unsupervised learning – predictive modelling
- Limited functionality for new advancements (deep learning)

Assignment 2: Next Steps…

Keep asking the questions:

- Does my data make sense?
- Is the data consistent?
- What do you make of the data's distribution? Does it change over time…is this to be expected?
- Use visualisations here to help – do you need to normalise data first?
- Is the data complete…missing values or anomalies?
- Do you understand the features…any data transformations required (consider ordinal and nominal data) ?
- Balanced or unbalanced data? Require a 50/50 split. Consider 'undersampling' or 'oversampling'.
- Any additional data that might be beneficial?

Assignment 2: Future steps…

Consider:

- Metric for evaluation – what is meant by a ‘good’ model?

- Splitting data – training and testing sets (import train_test_split). What percentage of your dataset would be appropriate for training?

Learning Objectives

University of Lancashire

Describe common examples of Machine Learning algorithms for Data Science

Apply algorithms to assist with research question(s) design