Machine Learning

Project Introduction

This week we will use data on titanic passengers. We will try to use data about the passengers to predict whether they survived or not.

1.1) Data

`train.csv`: The training data for this weeks project.

`test.csv`: The test data for this weeks project. The dataset on which we do the final evaluation of our model quality.

`penguins_simple.csv`: Practice / Lecture Dataset. We will use the penguins data in the lectures to explain concepts.

All of the datasets are already on GitHub under `week2/data`

1.2) Goals

- Understand the concept of Machine Learning
- Understand the machine learning models Logistic Regression, Decision Trees and Random Forests
- Learn which Feature Engineering techniques exist and how to apply them to our data
- Learn how to evaluate a model
- Build the best model possible in terms of "score"/"accuracy"
    - Ok but fairly easy: 0.76
    - Good :> 0.77
    - Very Good > 0.78
    - Awesome: > 0.8
- Submit our results to kaggle)

2) Machine Learning

2.1) What is it?

- You give data to the computer and ask the computer to learn about the computer using certain method and tools.
- Model training itself; the more data you put in, the better the model gets
- Providing data, model learns from experience, model improves over time, model makes predictions

2.2) Types of Machine Learning

2.2.1) Supervised Learning

- Know the right answer (at least for a sample of the data)
- Existence of an output variable that we want to predict: **y**
- We use input features **X** to predict **y**

2.2.1.1) Regression

y is a numeric value

2.2.1.2) Classification

- y is a class/category - Survived or Deady can be binary as in our Titanic case or you can have multiple classes

2.2.2) Unsupervised Learning

- There is no **y**Unsupervised learning algorithms are finding patterns in the data: **X**

2.2.2.1) Clustering

- Eg. we have customer data from a supermarket in we want to cluster customers into different categories

2.3.2.2) Dimensionality Reduction