CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Practical Machine Learning

Practical Machine Learning

Lecture: Introduction to Machine Learning

Ted Scully

Guidelines for Contacting Lecturers

- Please do not call into a lecturer's office without first making an appointment.

- The lectures and practical labs are a best forum for communicating with your lecturer about course contents. If you need clarification on certain concepts I recommend you ask for clarification during the lecture or during the practical lab.

- If the issue is urgent then please contact the lecturer via email to make an appointment.

Dates for Christmas Break

- For those of you wishing to book flights you can book flights on or before **Dec 14th**.

- Please note that you may still have some assignments due after this date but they can be submitted remotely through Canvas.

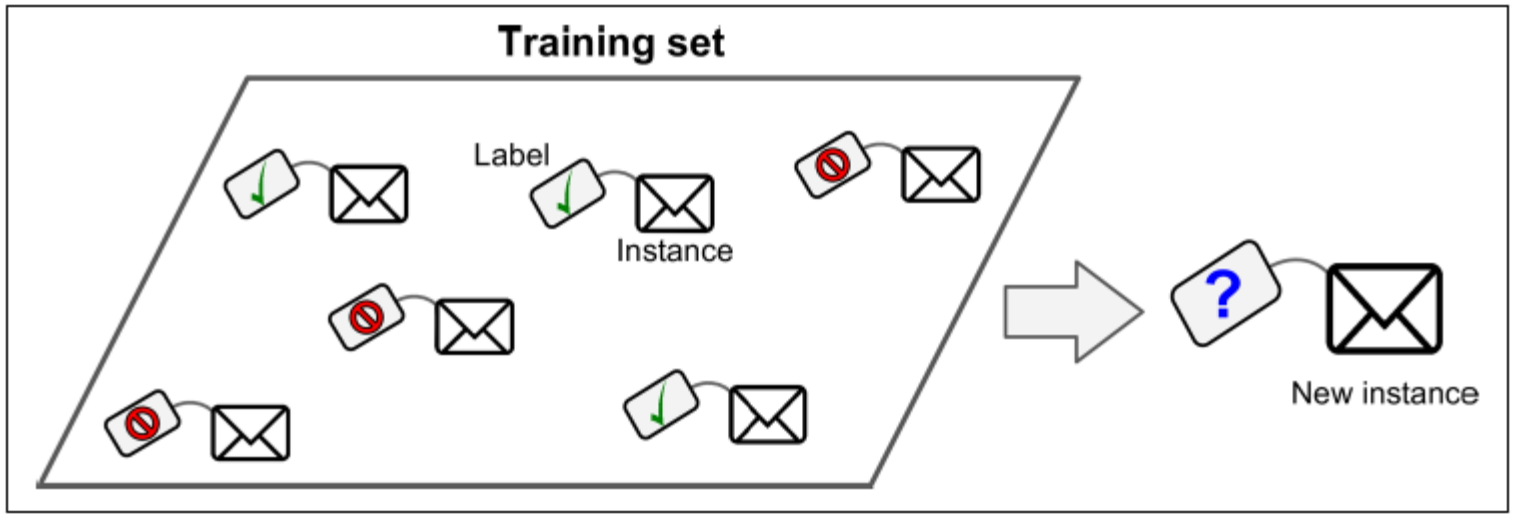
Categories of Machine Learning Algorithms

- Machine learning algorithms can be divided into four main categories
  - Supervised Learning Algorithms
  - Unsupervised Learning Algorithms
  - Semi Supervised Learning Algorithms
  - Reinforcement Learning Algorithms

Supervised Learning Algorithms

- Supervised learning algorithms used **labelled training data** to learn.

- In other words, the training data you feed to the algorithm includes the desired solutions, called labels.



Supervised Learning Algorithms

- Supervised learning can be subdivided into either classification or regression algorithms.

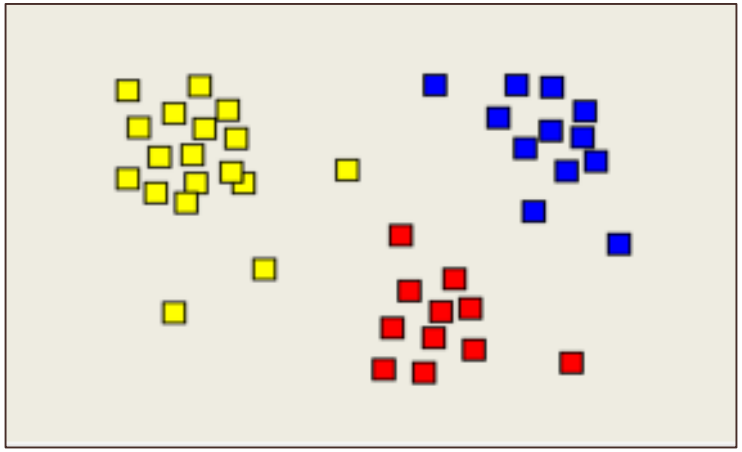
- In classification the objective is to correctly **predict the category** that new objects or cases belong to based on specific attributes they have. This decision is based on previous data that you have already observed.

- Regression is similar expect that rather than predicting a category we want to **predict a numerical value**. For example, predict the concentration of a drug based on a chemical analysis or predict a persons lifespan based on information about their health and lifestyle.

Unsupervised Learning Algorithms

- In unsupervised learning the algorithm is not provided with any labelled training data and **must learn patterns from the data**.

- Unsupervised algorithms seek out **similarity between pieces of data** in order to determine whether they can be characterized as forming a group.
  - These groups are termed clusters.
  - There are a broad range of clustering machine learning techniques.
  - Example- K Means Clustering (is told in advance how many clusters it should form -- a potentially difficulty)

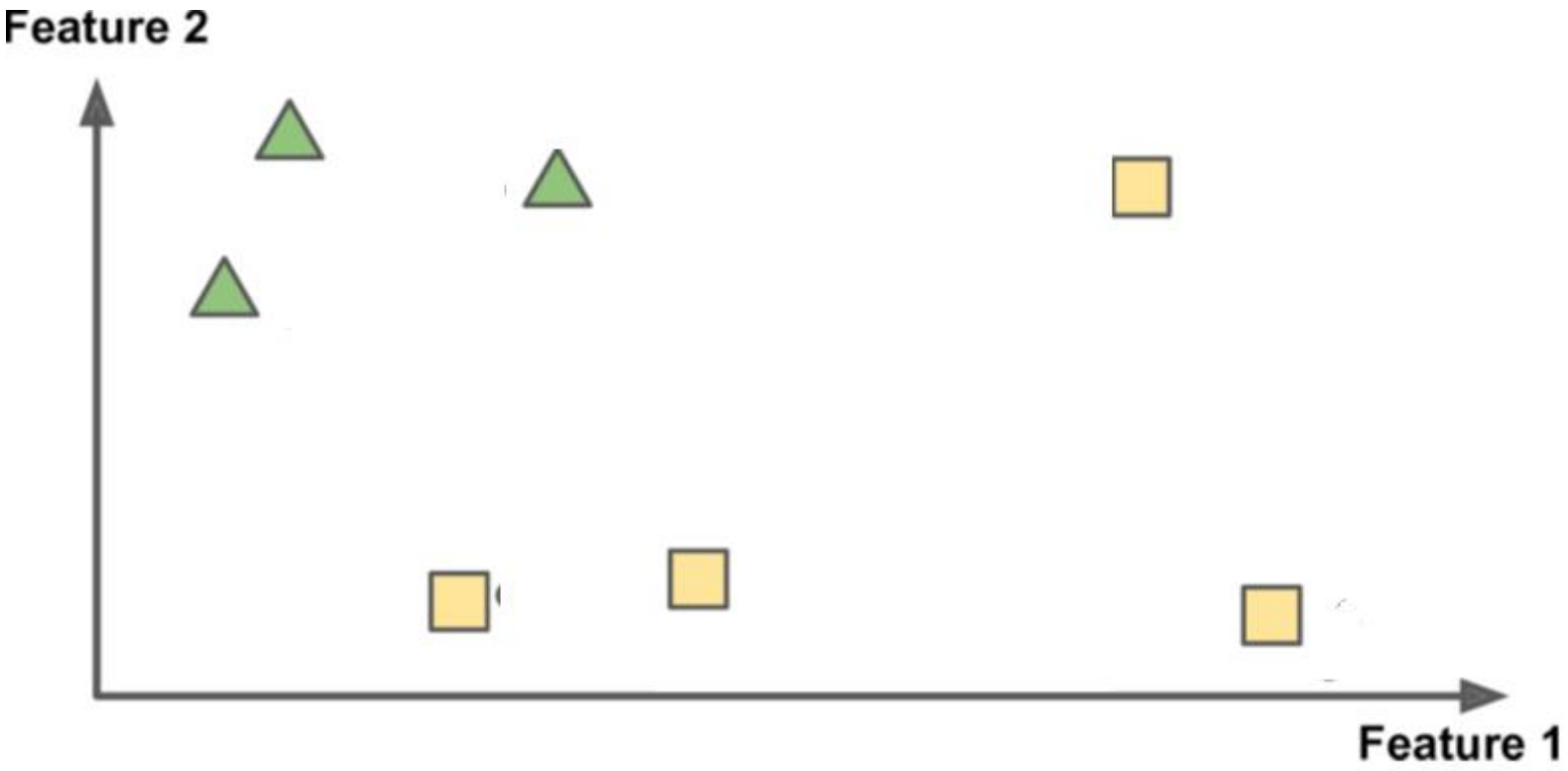


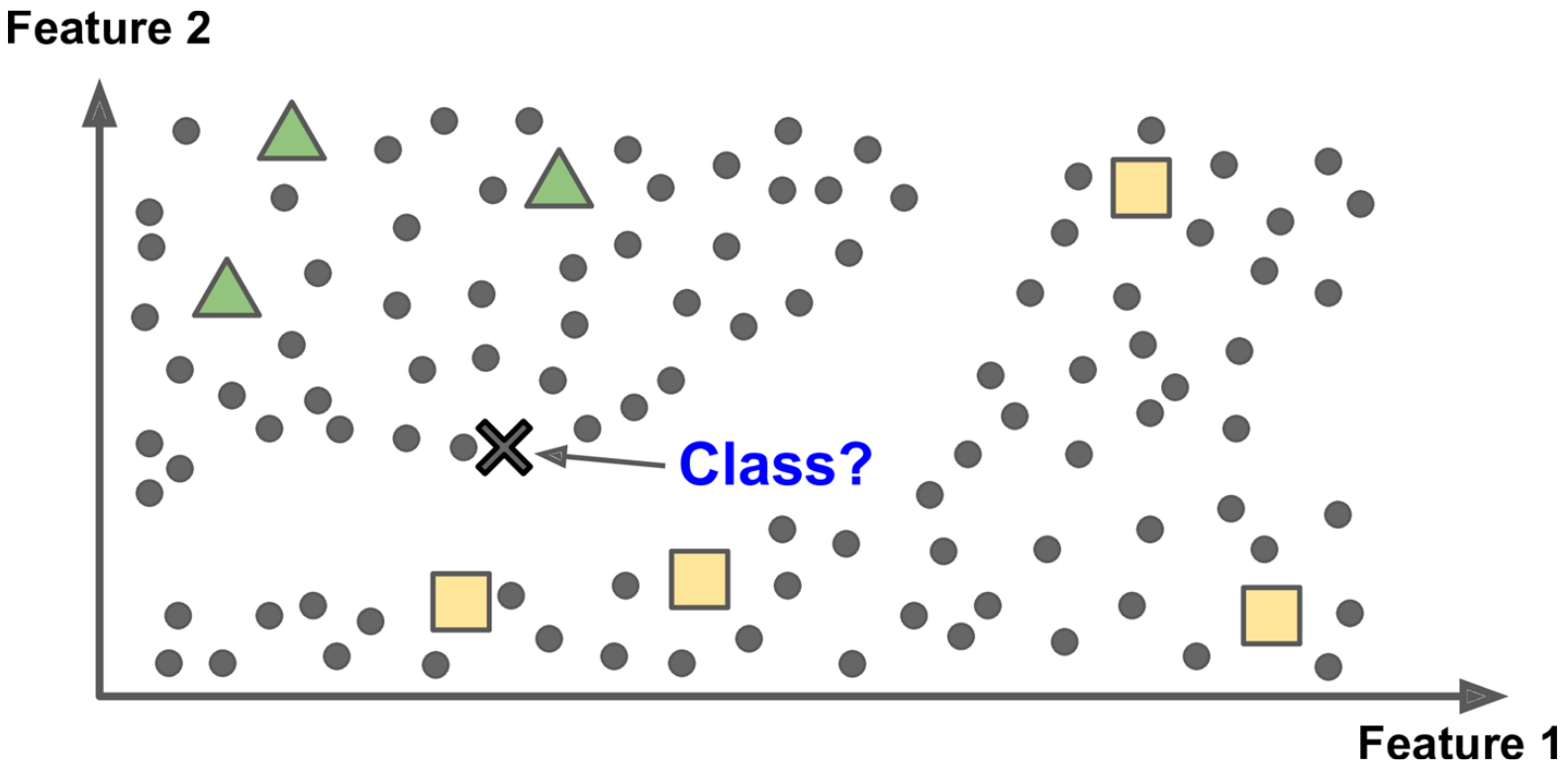
Semi-Supervised Learning

- The semi-supervised learning approach to machine learning **combines** the supervised and unsupervised learning techniques.

- Remember supervised learning uses a labelled training set, while unsupervised learning techniques use unlabelled data.

- A semi-supervised approach **utilises both labelled and unlabelled data** for training.
  - Normally a small amount of labelled data is used along with a large amount of unlabelled data

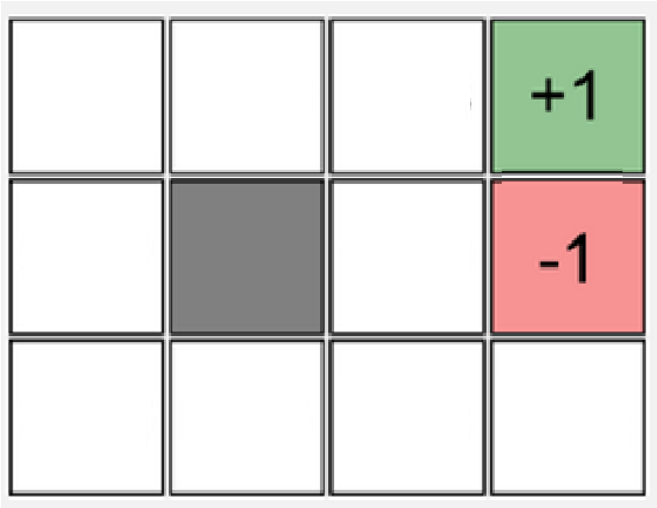

Feature 2
Feature 1


Feature 2
Class?
Feature 1

Reinforcement Learning Algorithms

- The objective of Reinforcement learning algorithms is to use observed rewards to **learn an optimal (or near-optimal) policy** for a given environment.

- The learning system, an observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards). It must then learn by itself what is the best strategy, called a policy, to get the most reward over time. **A policy defines what action the agent should choose when it is in a given situation.**



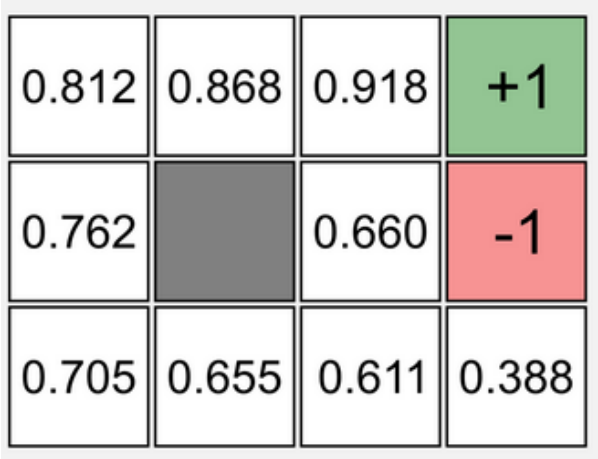
The intended direction of movement occurs with a probability of 0.8. With a 0.2 probability you will make a move at right angle to the intended direction. All other states have a reward of -0.04. The terminal states have rewards of +1 and -1 respectively.

Reinforcement Learning Algorithms

- The objective of Reinforcement learning algorithms is to use observed rewards to **learn an optimal (or near-optimal) policy** for a given environment.

- The learning system, an observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards). It must then learn by itself what is the best strategy, called a policy, to get the most reward over time. **A policy defines what action the agent should choose when it is in a given situation.**

- Unlike most other forms of machine learning the learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.



Basic Terminology Machine Learning Problem

- **Features (often referred to as attributes or variables)** below are Outlook, Temp, Humidity and Windy
- The **Class (Label)** is Play (for regression often referred to as regression target)
- **We refer to an instance as one row from the dataset.**
- **Inference –** Model takes in unseen feature vector and produces a classification.

Tennis Dataset

ID	Outlook	Temp	Humidity	Windy	Play?
A	sunny	hot	high	false	no
B	sunny	hot	high	true	no
C	overcast	hot	high	false	yes
D	rainy	mild	high	false	yes
E	rainy	cool	normal	false	yes
F	rainy	cool	normal	true	no
G	overcast	cool	normal	true	yes
H	sunny	mild	high	false	no
I	sunny	cool	normal	false	yes
J	rainy	mild	normal	false	yes
K	sunny	mild	normal	true	yes
L	overcast	mild	high	true	yes
M	overcast	hot	normal	false	yes
N	rainy	mild	high	true	no

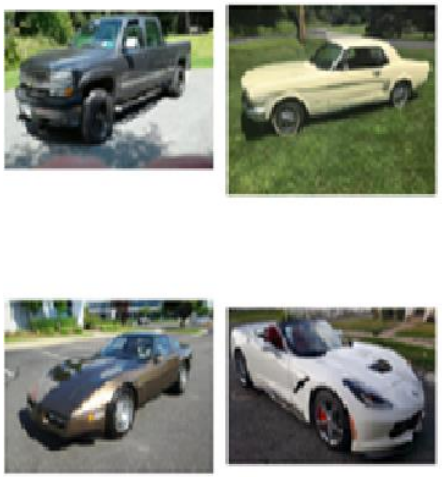
Challenges in Machine Learning

- There are a range of challenges that you may encounter when attempt to build a machine learning model and these can be largely categorized into either data-based issue or model based issues.

- **Insufficient Amount of Training Data**
  - To work well ML algorithms commonly need quite a lot of data.
  - Even for very simple problems you may often need many hundreds of training examples, and for complex problems such as image or speech recognition you may need millions of examples (unless you can reuse parts of an existing model).
  - It is also important to be able to **diagnose** you ML model to determine if a lack of data may improve it's overall level of accuracy.

Challenges in Machine Learning

- **Non-representative Training Data**
  - In order to generalize well, it is crucial that your training data be representative of the new cases you want to generalize to.
  - By using a non-representative training set, we will train a model that is unlikely to make accurate predictions.
  - This is often harder than it sounds: if the sample is too small, you will have sampling noise (i.e., non-representative data as a result of chance), but even very large samples can be non-representative if the sampling method is flawed. This is often referred to as sampling bias.



Challenges in Machine Learning

- Issues with your Data
  - If your training data is full of errors, **outliers, and noise** (e.g., due to poor quality measurements, faulty sensors, etc), it will make it harder for the system to detect the underlying patterns.
  - It is very common to spend time cleaning up your training data.
  - For example, if some instances are clearly **outliers**, it may help to simply discard them.
  - If some instances are **missing** a few features (e.g., 5% of your customers did not specify their age), you must decide whether you want to ignore this feature altogether, ignore these specific instances with missing value, fill in the missing values.
  - A critical part of the success of a Machine Learning project is coming up with a good set of features to train on. **Feature selection**: selecting the most useful features to train on among existing features. **Feature extraction**: combining existing features to produce a more useful one.

Model Challenges in ML

> All generalizations are dangerous, even this one.
>
> - Alexandre Dumas

- When you build a machine learning model you hope that it generalizes well. That is, you hope your algorithm can take the training data and build a model, which in turn can take unseen instances and classify them correctly.

- Unfortunately models do not always generalize well.

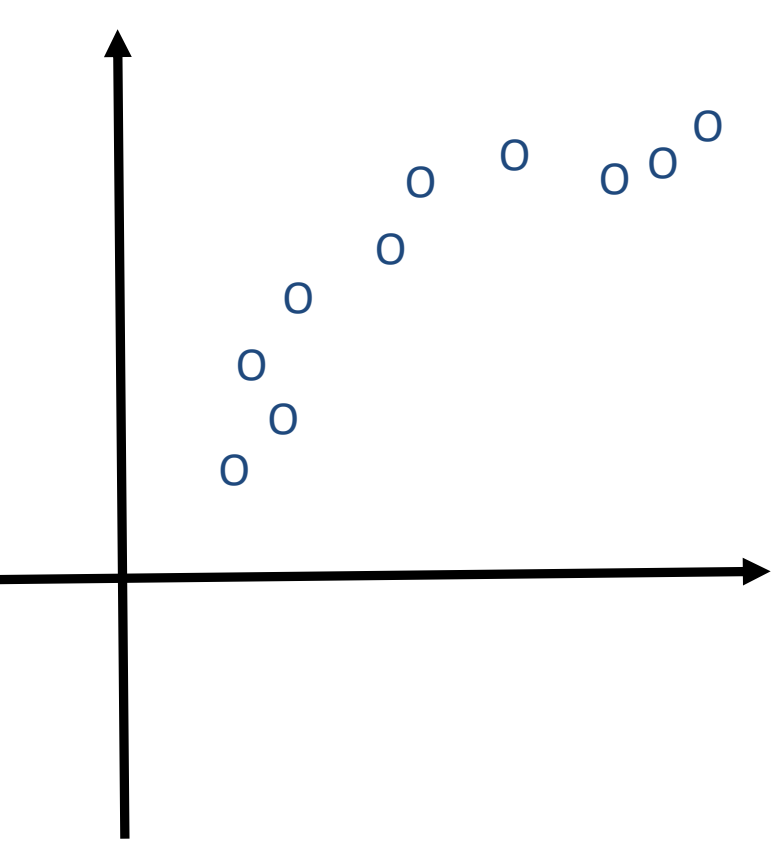
- Clearly we need strong evaluation methods to measure model performance. When models are underperforming we also need to be able to **diagnose** if the source of this problem is due to overfitting or underfitting.

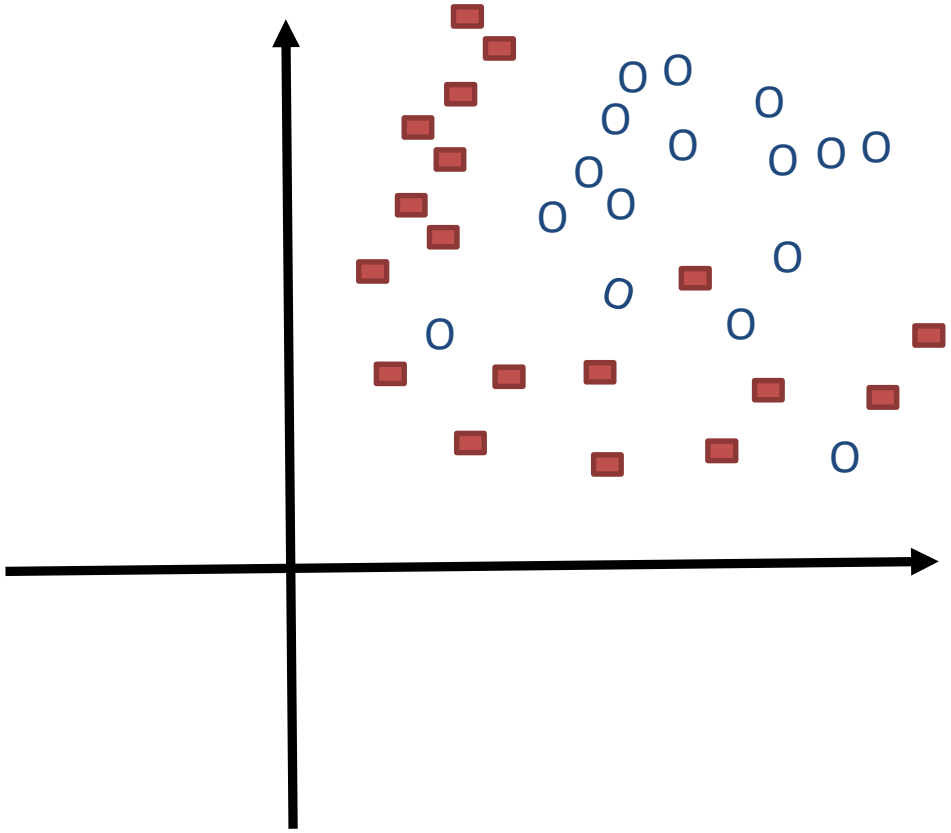
Overfitting

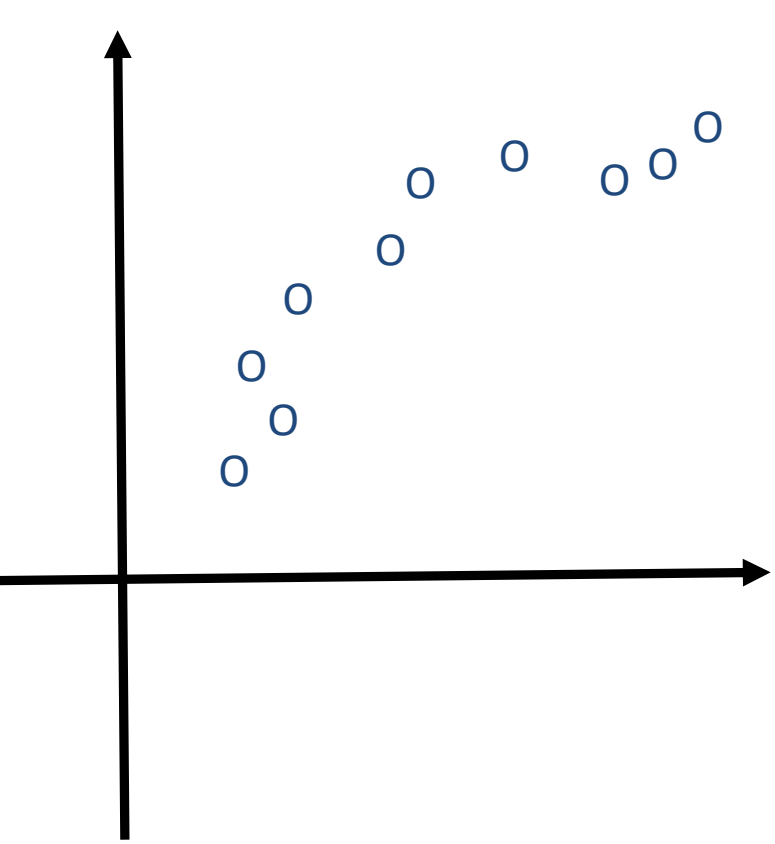
- Overfitting generally occurs when a model/**function is excessively complex**, such as having too many parameters relative to the number of labelled training data.

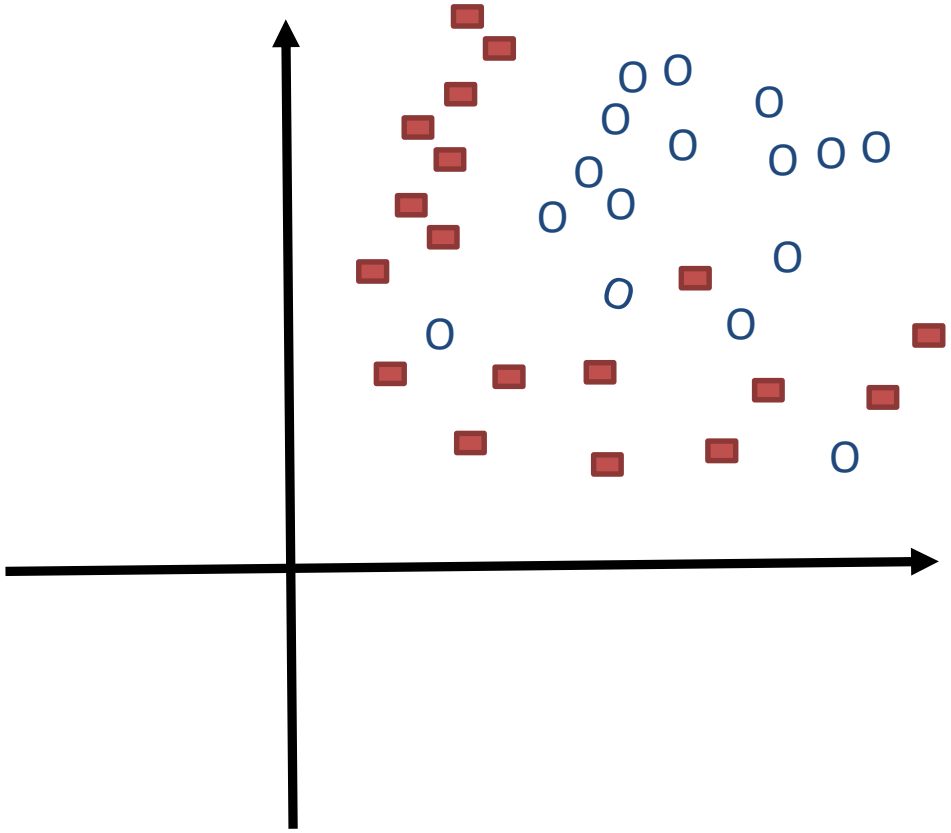
- A model/**function which has been overfit will generally have poor predictive performance on unseen data (it doesn't generalize well to unseen examples)**, as it can exaggerate minor fluctuations in the data.
  - A model is typically **trained** by maximizing its performance on some set of training data.
  - However, its overall performance is determined not by its performance on the training data but by its ability to perform well on **unseen data**.

- You can think of this as the difference between memorizing the data and generalizing from the data.









Overfitting

- Common methods used to address overfitting are:
  - **Simplify the model** by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of features in the training data or by constraining the model
  - To gather **more training data**
  - To **reduce the noise** in the training data

Underfitting

- As you might guess, underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data.

- The main options to fix this problem are:
  - Selecting a more **complex model**, with more parameters
  - Feeding **better features** to the learning algorithm (feature engineering)