First Step to Practical Machine Learning

Knowledge Sharing for CPE/SKE students

Sirakorn Lamyai

Student, Kasetsart U.

October 30, 2018

Before we start...

Make sure these are installed on your computer.

This page is a guide for installing on Windows

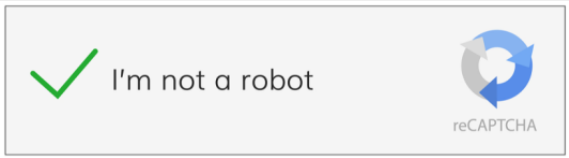
- Python 3.6: Download and install at `https://www.python.org`
- NumPy, Scipy, Matplotlib, Scikit-learn, MLxtend:
  Run `pip install numpy scipy matplotlib sklearn mlxtend`

Outline

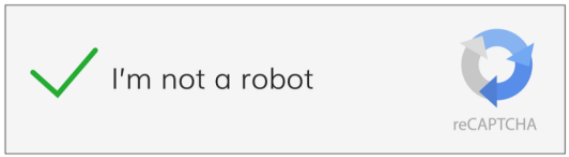
Introduction to Machine Learning

What is Machine Learning?

What is Machine Learning?

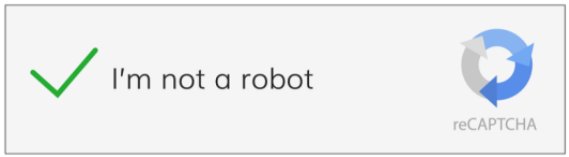


What is Machine Learning?



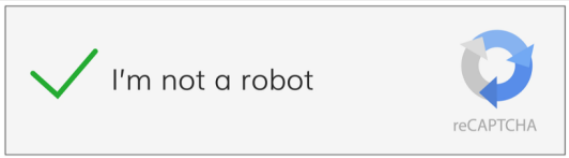
- This is Recaptcha.

What is Machine Learning?



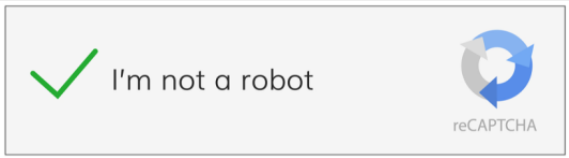
- This is Recaptcha.
  - Recaptcha helps stop millions of spam a day.

What is Machine Learning?



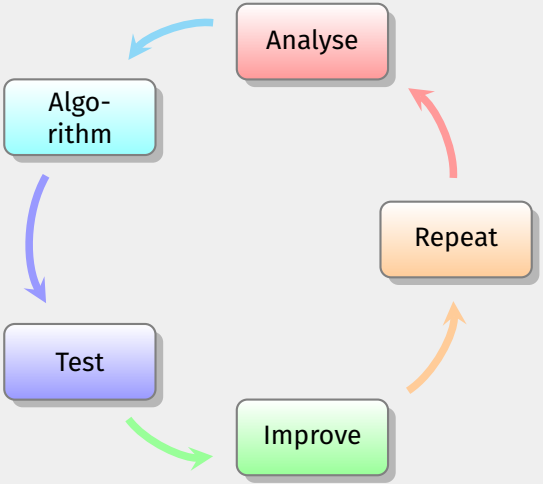
- This is Recaptcha.
  - Recaptcha helps stop millions of spam a day.
  - In some old days, we have to type Captcha texts to distinguish ourself from bots.

What is Machine Learning?

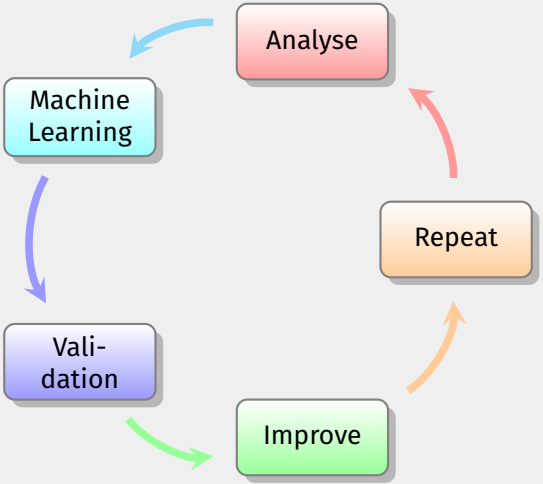


- This is Recaptcha.
  - Recaptcha helps stop millions of spam a day.
  - In some old days, we have to type Captcha texts to distinguish ourself from bots.
  - How is it possible that with a single click, an automated system can distinguish bots from humans?

Traditional programming approach



Machine learning approach



In other words...

Machine Learning

In other words...

Machine Learning

= Data + Data analysis algorithm

In other words...

Machine Learning

= Data + Data analysis algorithm

= Adapt to change

Machine Learning Problems

Types of Machine Learning problems

Types of Machine Learning problems

1. Supervised learning

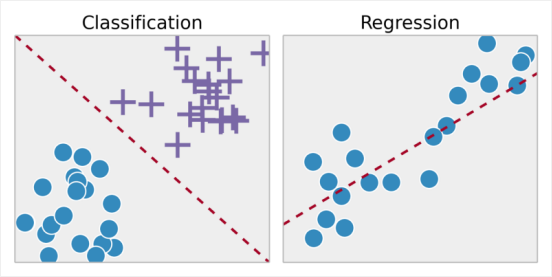
Types of Machine Learning problems

1. Supervised learning
2. Unsupervised learning

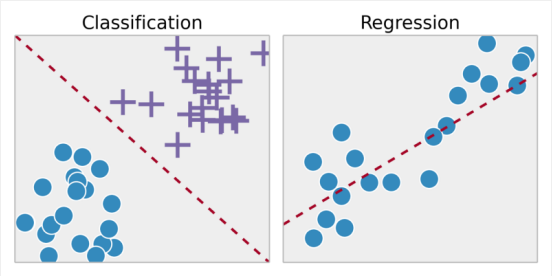
Types of Machine Learning problems

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Supervised learning

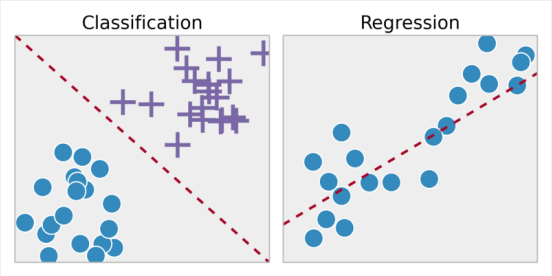


Supervised learning



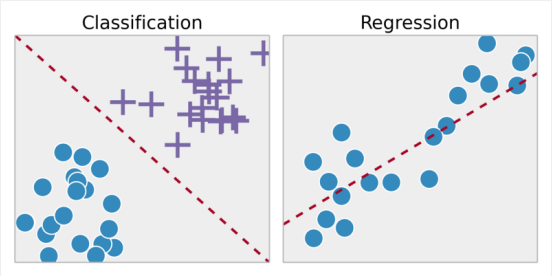
- Given a **training set** for the data, find a **model** to **generalise** well to **unseen** data.

Supervised learning



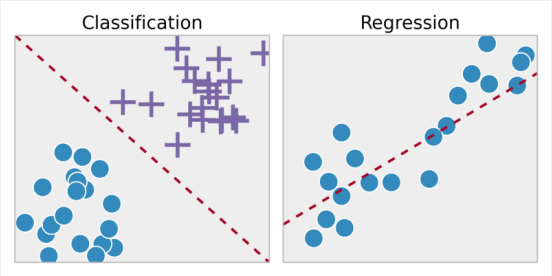
- Given a **training set** for the data, find a **model** to **generalise** well to **unseen** data.
- Two main supervised learning problems

Supervised learning



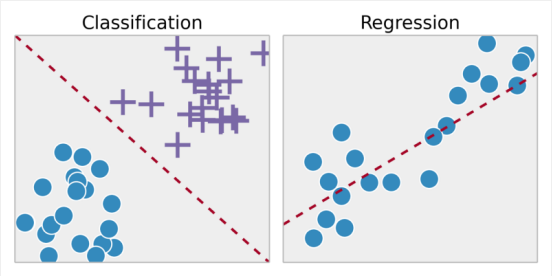
- Given a **training set** for the data, find a **model** to **generalise** well to **unseen** data.
- Two main supervised learning problems
  - Classification: On the discrete data

Supervised learning



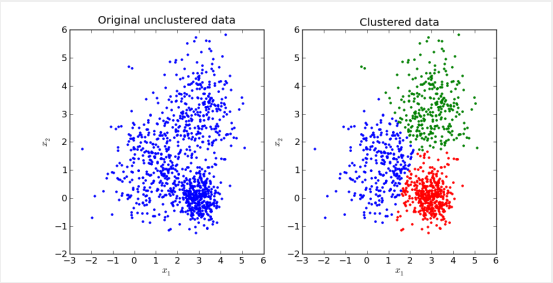
- Given a **training set** for the data, find a **model** to **generalise** well to **unseen** data.
- Two main supervised learning problems
  - Classification: On the discrete data
  - Regression: On the continuous data

Supervised learning

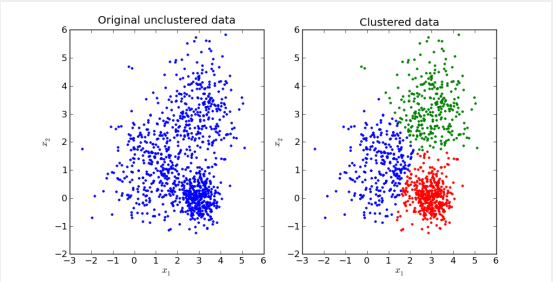


- Given a **training set** for the data, find a **model** to **generalise** well to **unseen** data.
- Two main supervised learning problems
  - Classification: On the discrete data
  - Regression: On the continuous data
- Example problems: Spam E-mail detection, Facial recognition

Unsupervised learning

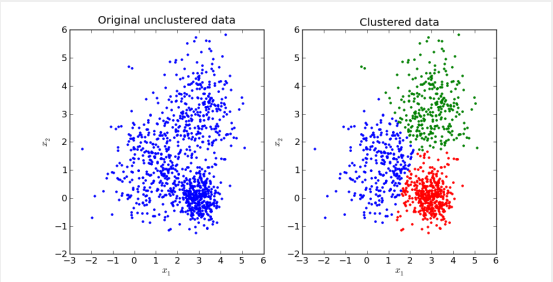


Unsupervised learning



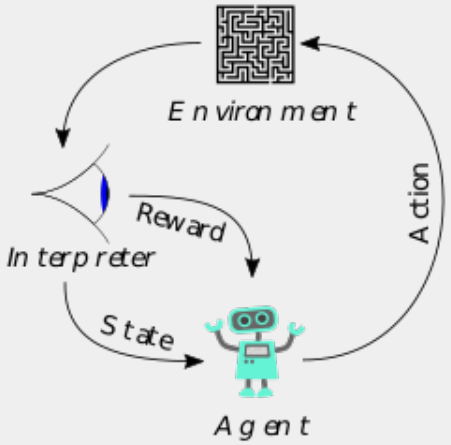
- Discover **hidden** structure in **non-labelled** data.

Unsupervised learning



- Discover **hidden** structure in **non-labelled** data.
- Example: Clustering, Generative models

Reinforcement learning



Model

Model

Model

- A result of the combination between...

Model

- A result of the combination between...
  - a **method** to recognise the data, and

Model

- A result of the combination between...
  - a **method** to recognise the data, and
  - **sample datas** for such the method

Model

- A result of the combination between...
  - a **method** to recognise the data, and
  - **sample datas** for such the method

Data

Model

- A result of the combination between...
  - a **method** to recognise the data, and
  - **sample datas** for such the method

Data

Determine which group should the purple dot be in (red/green/blue) by **checking the colour of its nearest dot.**

Method

Beginning with our first model

Beginning with our first model

- We're going to write our **first own** machine learning algorithm called *k*-**Nearest Neighbour** (*k*-NN)

Beginning with our first model

- We're going to write our **first own** machine learning algorithm called ***k*-Nearest Neighbour** (*k*-NN)
  - *k*-NN is known to be very simple, with its concept as

Beginning with our first model

- We're going to write our **first own** machine learning algorithm called ***k*-Nearest Neighbour** (*k*-NN)
  - *k*-NN is known to be very simple, with its concept as

_k_-NN algorithm

To classify label of a data point, get *k* nearest data points to the data point, and select the major label among those data points.

Coding time!

Machine Learning Process

Machine Learning Process

- Train
- Test

(There'll be more of this, trust me.)

Choosing the parameter for k-NN algorithm

What is the bad way to choose k?

- What if we choose k = # of all points?
  - What will happen if our dataset's got 3 labels of A, B, C with 10, 20, and 30 data points of each?
  - Answer: Our model will always answer the labels with the highest data point count.
- What if we choose k = 1?
  - Let's try!

Coding time!

Training and Testing set

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**
  - Most of the time, the testing set will be around 10-25% of the entire dataset

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**
  - Most of the time, the testing set will be around 10-25% of the entire dataset
  - What will happen if we train on the testing set?

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**
  - Most of the time, the testing set will be around 10-25% of the entire dataset
  - What will happen if we train on the testing set?
  - What will happen if we test on the training set?

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**
  - Most of the time, the testing set will be around 10-25% of the entire dataset
  - What will happen if we train on the testing set?
  - What will happen if we test on the training set?
    - **Cheating!** Like letting the model *remembers* the answer instead of **generalising** the data pattern.

Training and Testing set

- We separate our dataset into 2 parts: the **training set** and **testing set**
  - Most of the time, the testing set will be around 10-25% of the entire dataset
  - What will happen if we train on the testing set?
  - What will happen if we test on the training set?
    - **Cheating!** Like letting the model *remembers* the answer instead of **generalising** the data pattern.
  - In other words, **don't test and train model on the same set of data.**

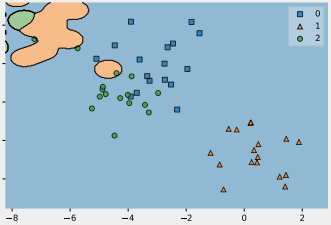
Choosing the best k

- **Train** with the training set, to let our model know how will the data looks like.
- **Test** with the testing set, to see on how our model performs.

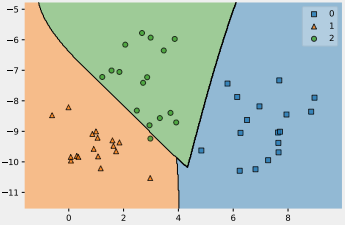
Warning! This is a simplified Machine Learning model training process, there are more to concerns!

Overfitting and underfitting

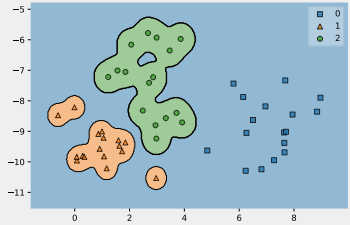
Which decision region is good?



Underfit: The model fails to recognise data pattern



Good fit: The model recognises data pattern **generally**



Overfit: The model **remembers** data pattern instead of generalising.

Good model must **generalise**