LONDON METROPOLITAN UNIVERSITY

Machine Learning CS7052 Lecture 3

Dr. Elaheh Homayounvala

week 3

LONDON METROPOLITAN UNIVERSITY

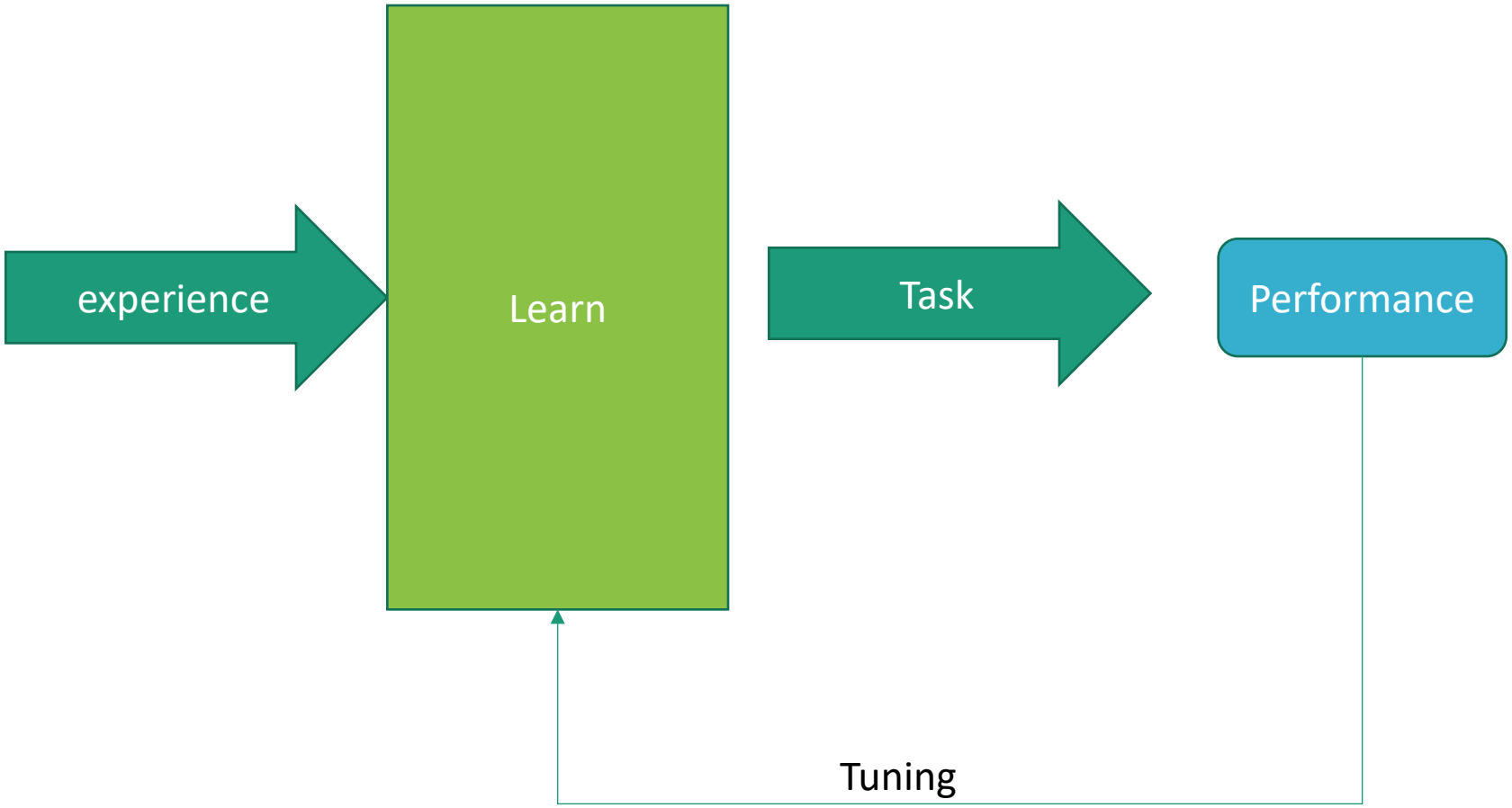
Outline of today's lecture

- Review last two weeks
  - What is Machine Learning (ML)?
  - Types of Learning
  - Understanding Data and Data analysis Process
  - A First Application, Iris species
- Student Rep
- Groups for coursework
- Supervised learning, K-Nearest Neighbours (k-NN)
  - KNN Classification
  - KNN Regression
  - Overfitting and underfitting

LONDON METROPOLITAN UNIVERSITY

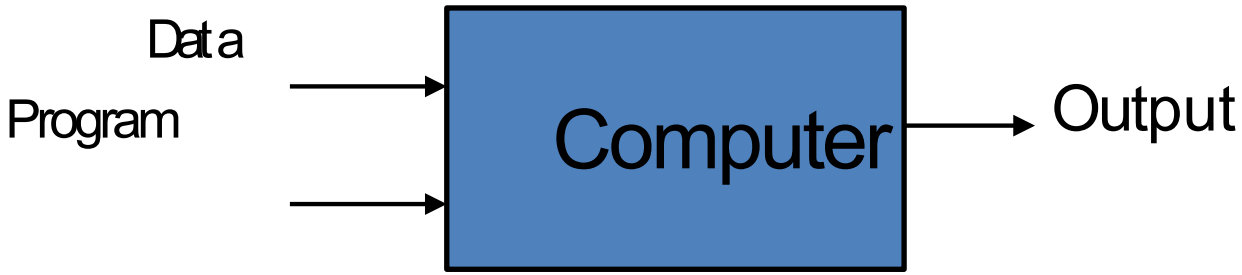
Review last two weeks

What is ML?

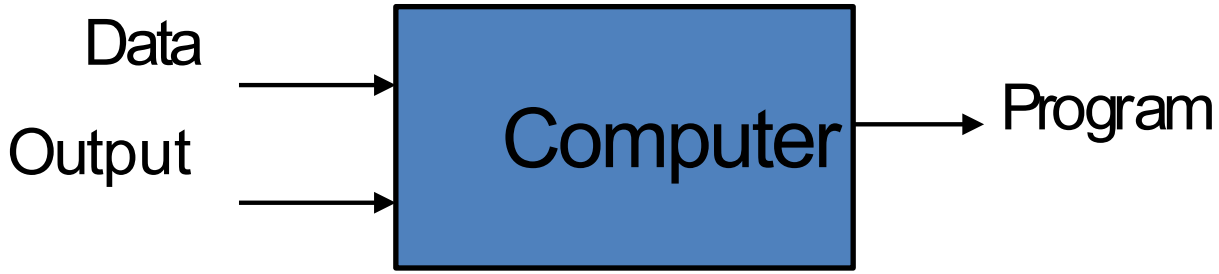


Traditional Programming vs. Machine Learning

Traditional Programming



Machine Learning



LONDON METROPOLITAN UNIVERSITY

Types of Learning

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
- Semi-supervised learning
- Reinforcement learning

LONDON METROPOLITAN UNIVERSITY

Supervised vs Unsupervised Learning

- Supervised learning
  - Classification, Target data/result/label is categorical
  - Regression, Target data is numerical

- Unsupervised learning
  - Clustering, Target data is not available/descriptive

Data Analysis Process

- Source: Nelli's book page 8

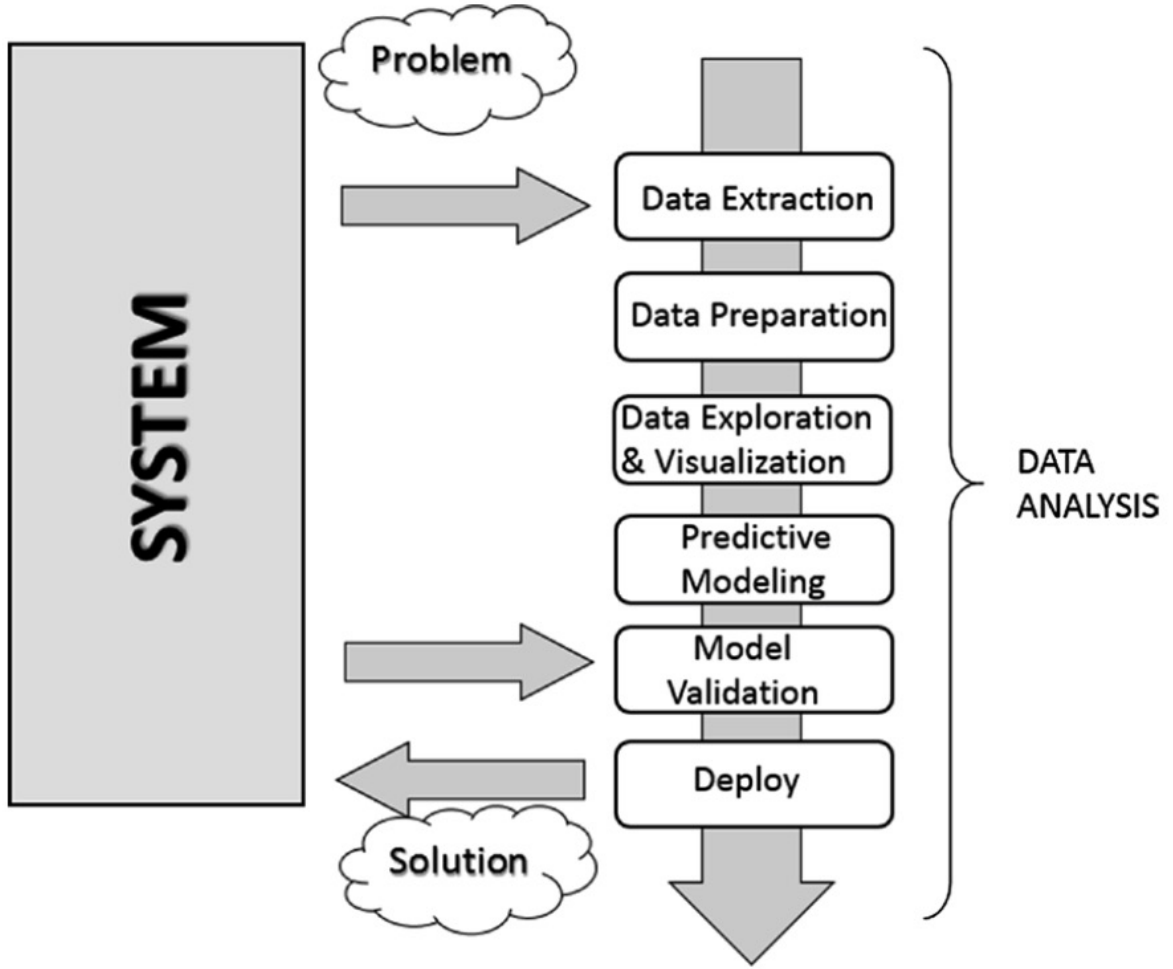
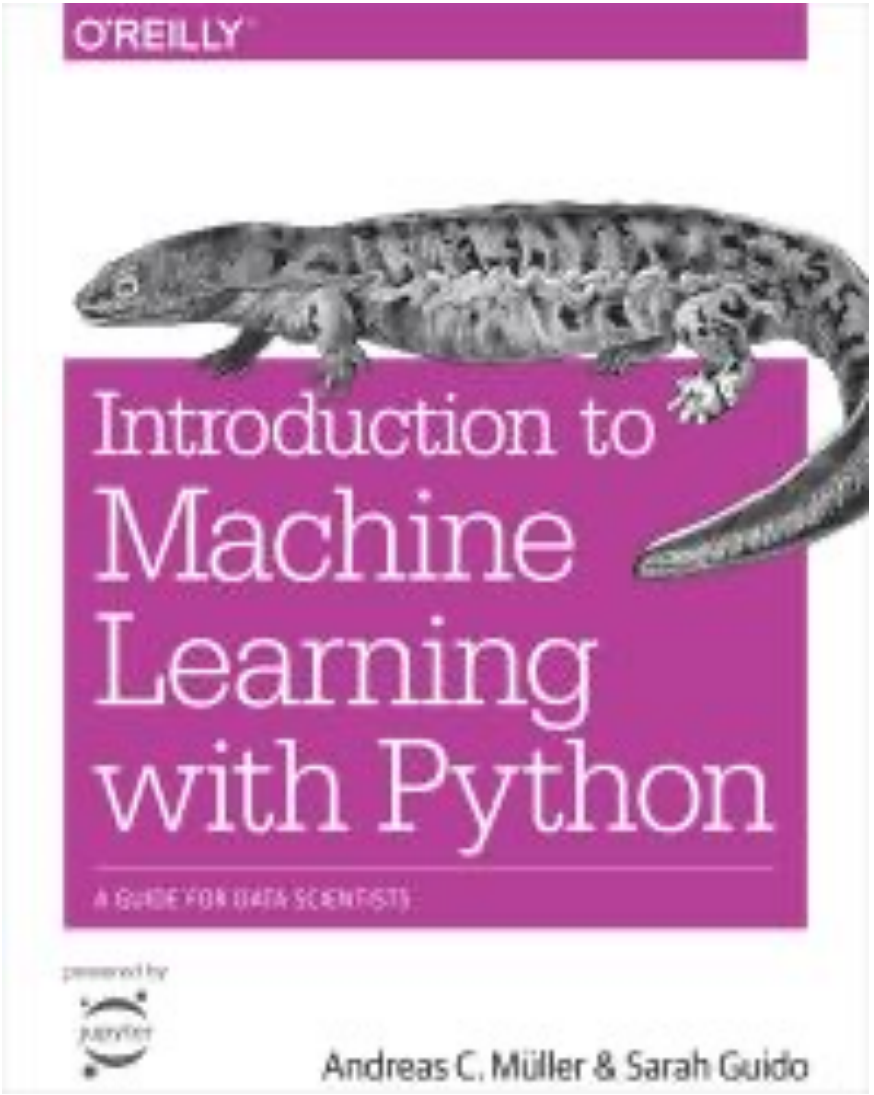


Figure 1-1. *The data analysis process*

A First Application, Classifying Iris Species

- Muller & Guido's book
- Chapter 1, pp. 13-23



A First Application, Iris species

Build a machine learning model

- that can learn from the measurements of these irises whose species is known
- so that we can predict the species for a new iris

LONDON METROPOLITAN UNIVERSITY

k-Nearest Neighbours

LONDON METROPOLITAN UNIVERSITY

k-Nearest Neighbours

- k-NN is arguably the simplest machine learning algorithm
- Building the model consists only of storing the training dataset
- To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset—its “nearest neighbours.”
- k-Nearest Neighbours algorithm is abbreviated as k-NN.

Muller and Guido’s book, Chapter 2, page 37-46

K-NN Classification

- Simplest version, K=1
- K-NN considers exactly **one** nearest neighbour
- Which is the **closest** training data point to the point we want to make a prediction for.

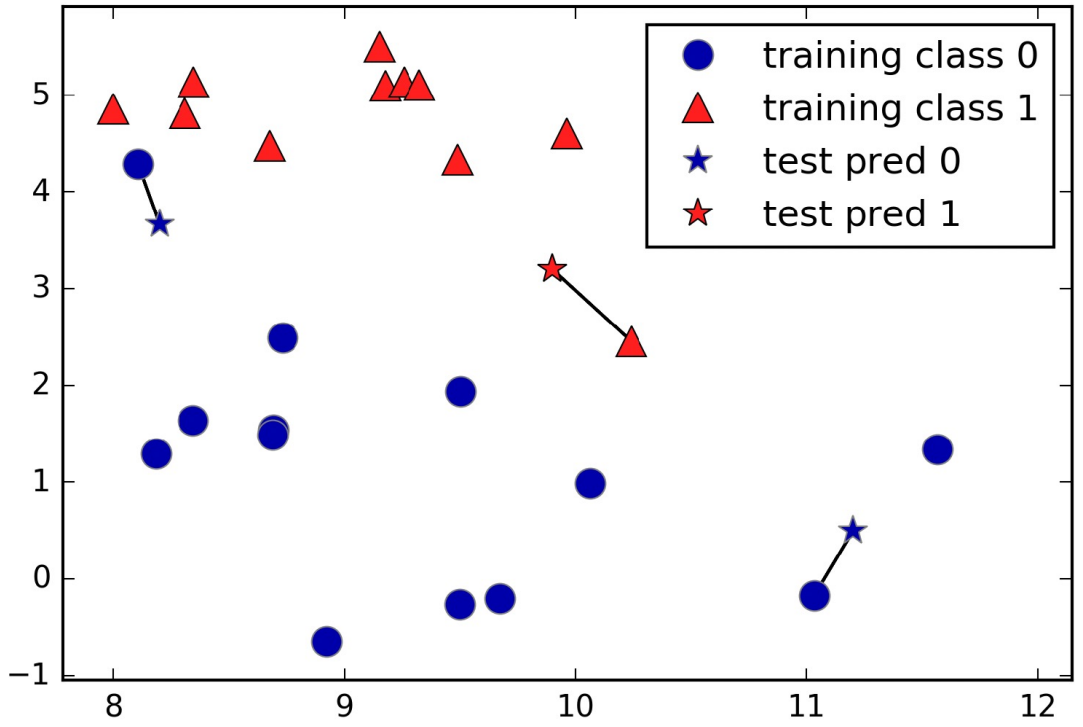


Figure 2-4. Predictions made by the one-nearest-neighbor model on the forge dataset

- Reference: Muller & Guido's book page 37

K = 3

- Three Nearest Neighbours

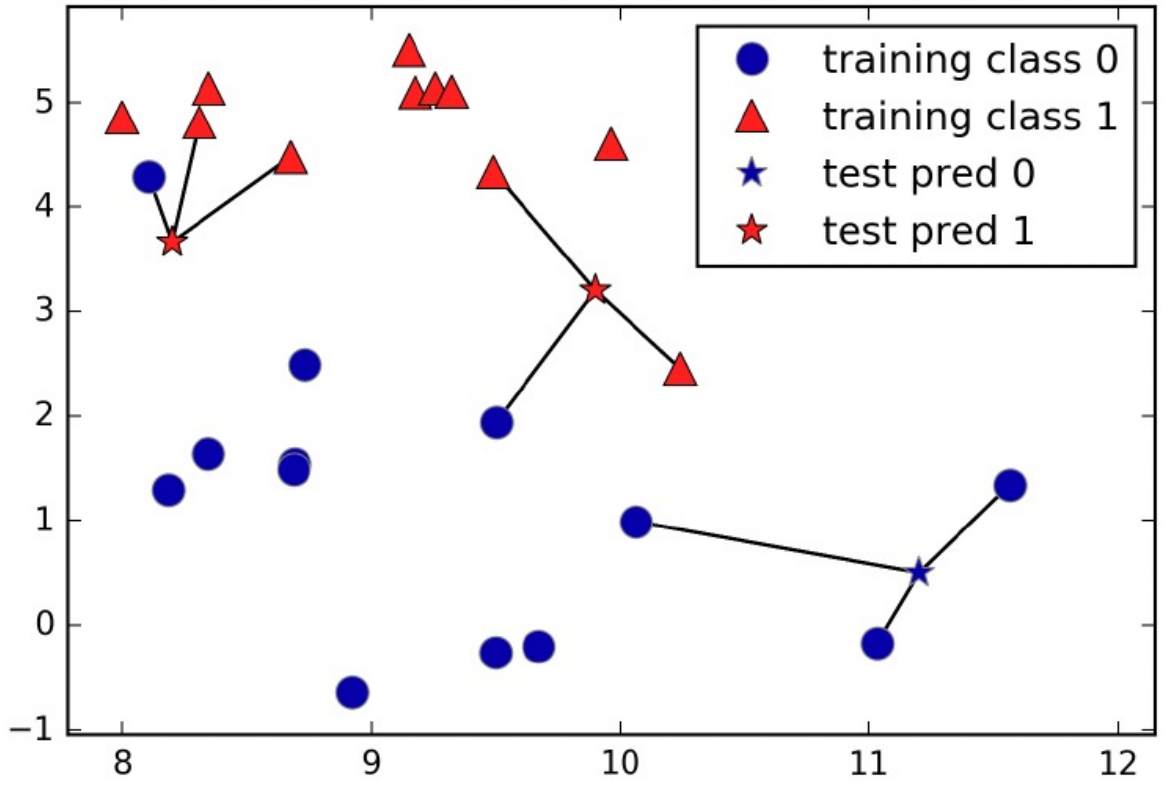
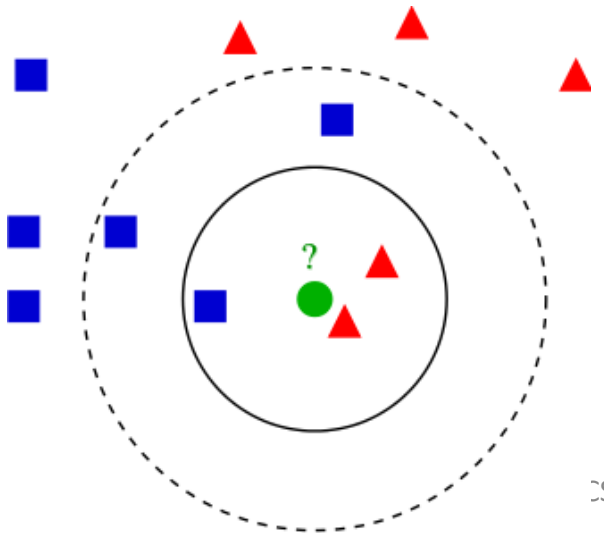


Figure 2-5. Predictions made by the three-nearest-neighbors model on the forge dataset

K-nearest Neighbours Algorithm, Classification

1. Find k closest objects to the predicted object x in the training set.

2. Associate x the most frequent class among its k neighbours.



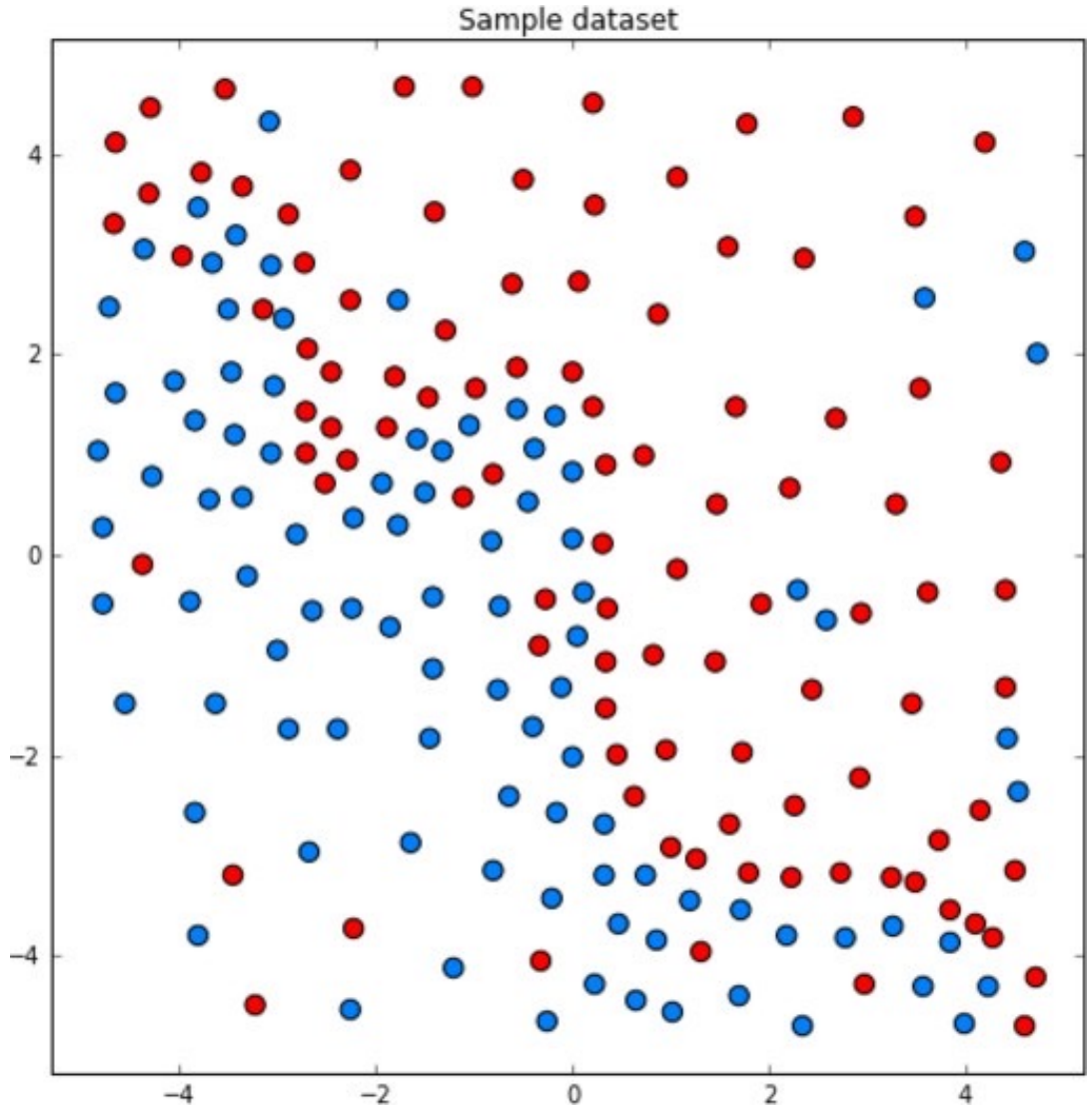
Comments

- k = 1: nearest neighbour algorithm[1]
- Base assumption of the method[2]:
  - similar objects yield similar outputs

[1] what will happen for $k = N$?
[2] what is simpler - to train k-NN model or to apply it?

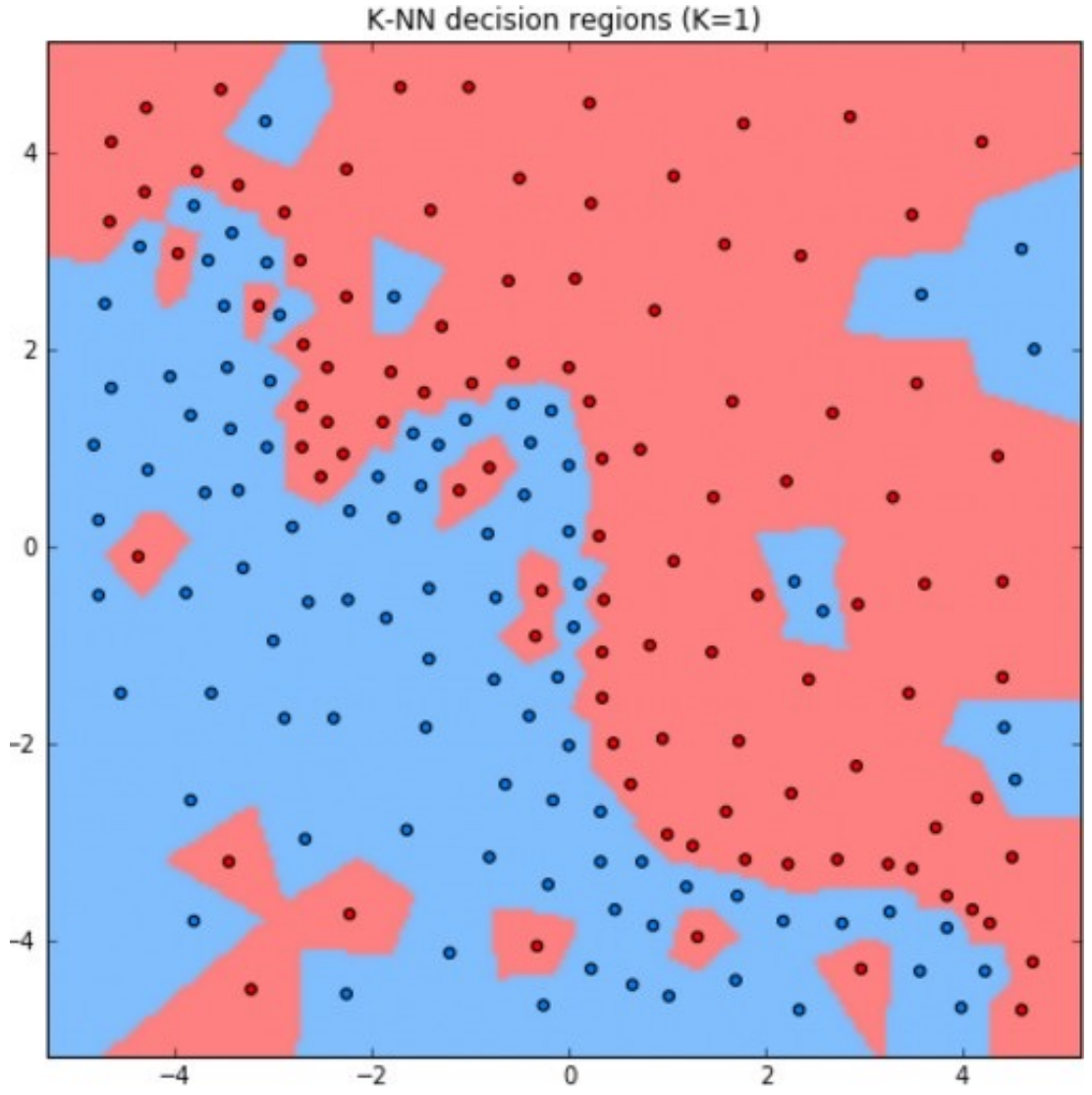
Sample Dataset

- How can we draw decision boundaries?



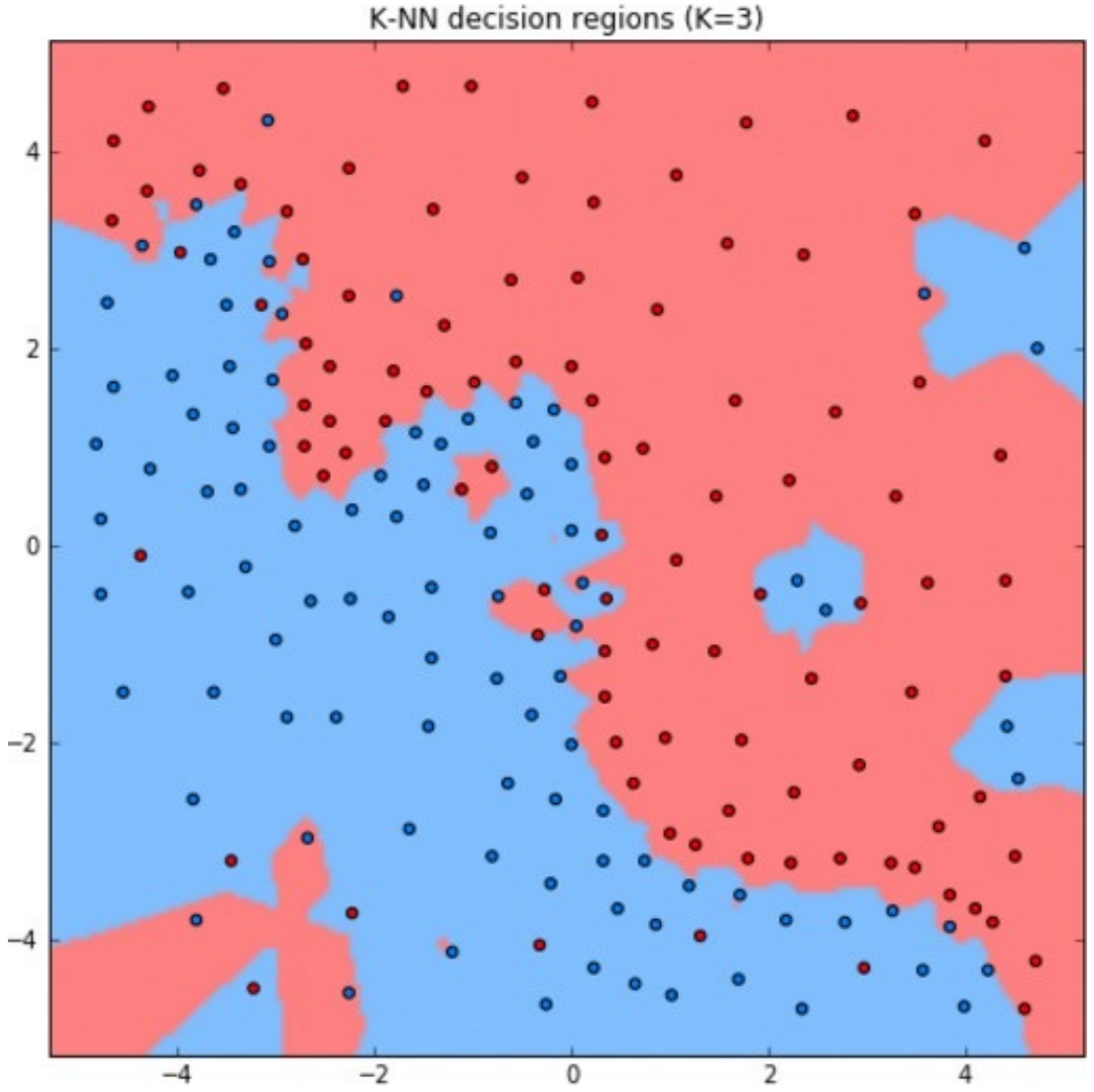
Example: K-NN classification

- Decision boundaries or Decision regions
- When k = 1



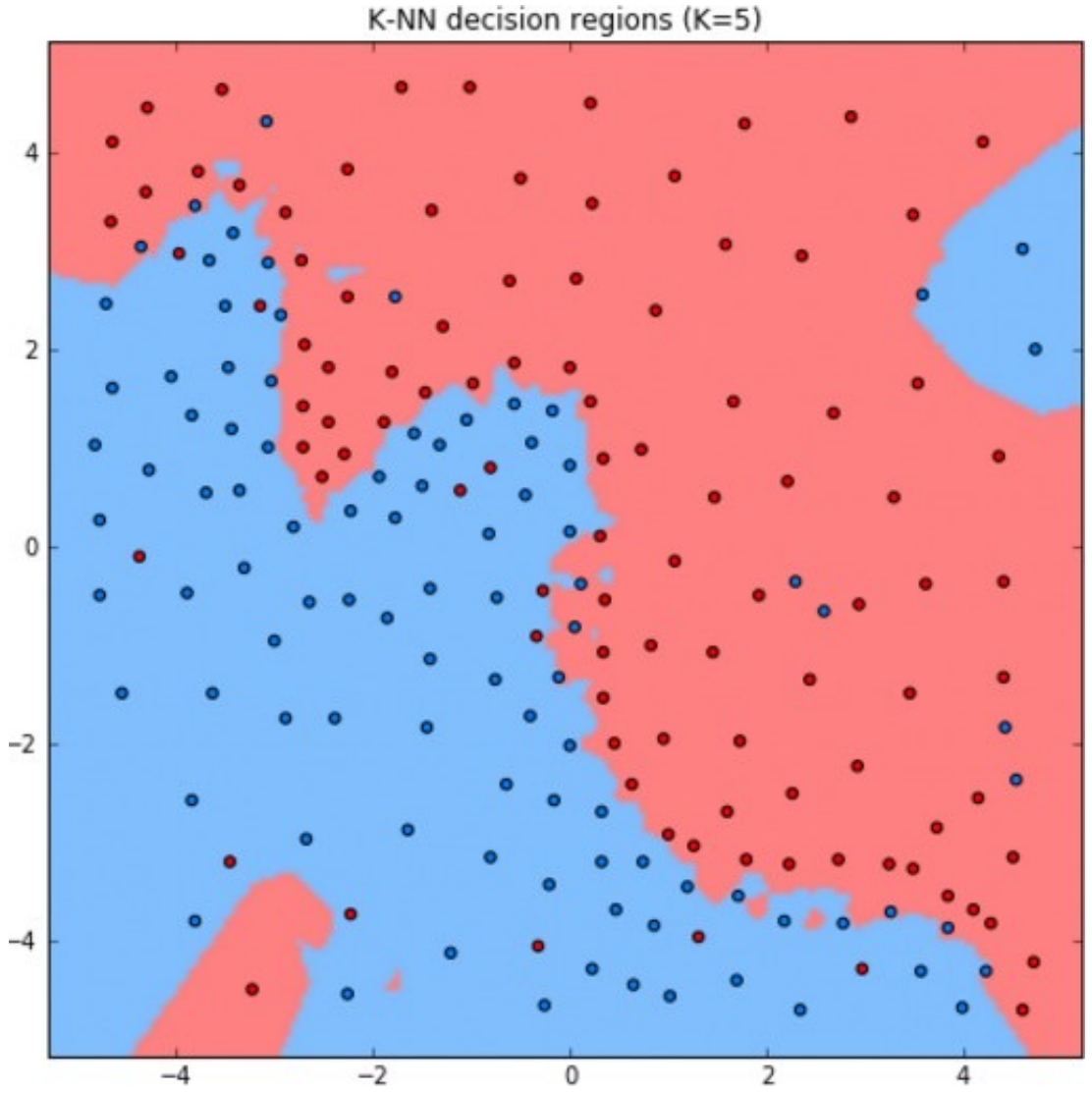
Example: K-NN classification

- Decision boundaries or Decision regions
- When k = 3



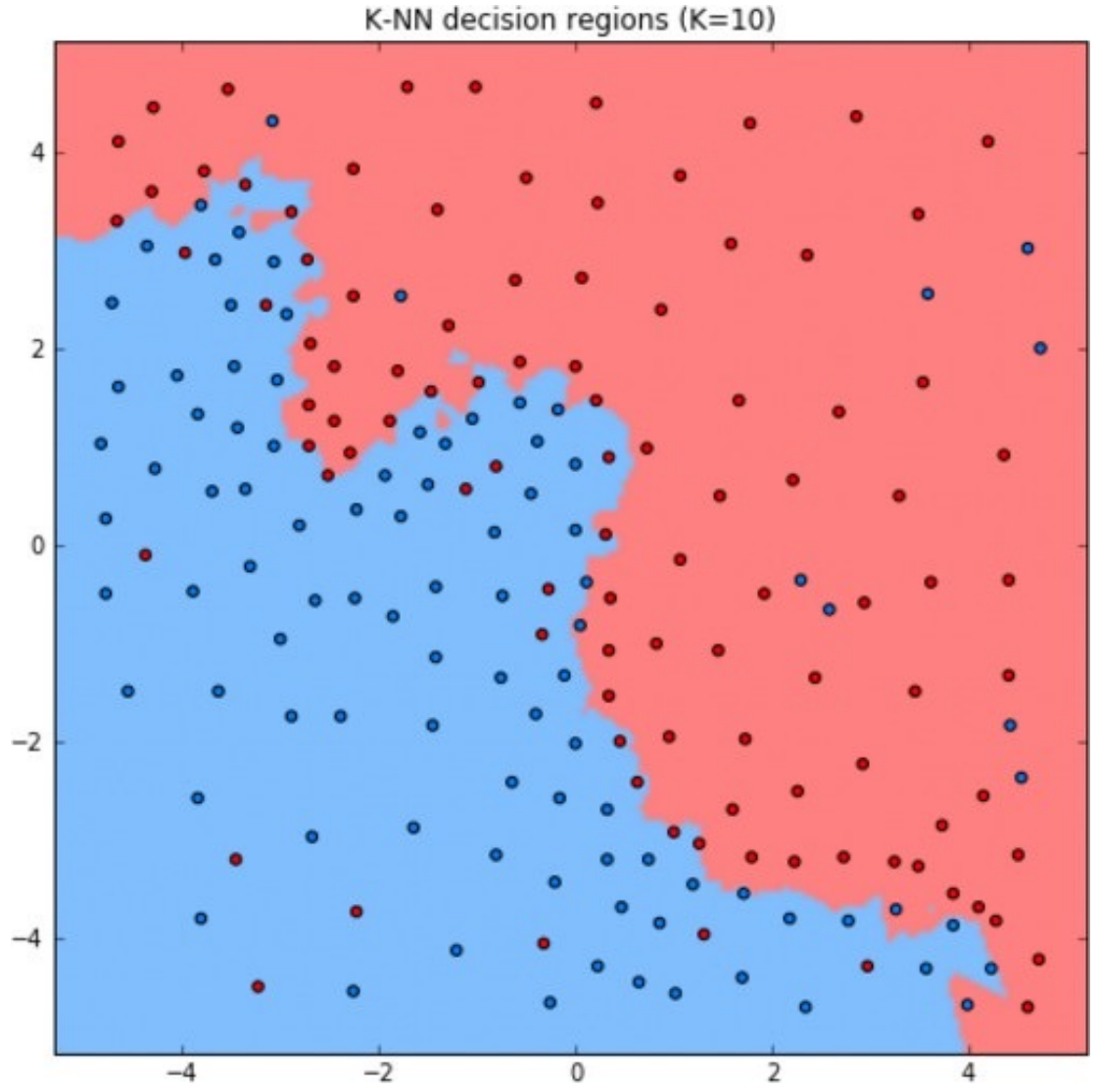
Example: K-NN classification

- Decision boundaries or Decision regions
- When k = 5



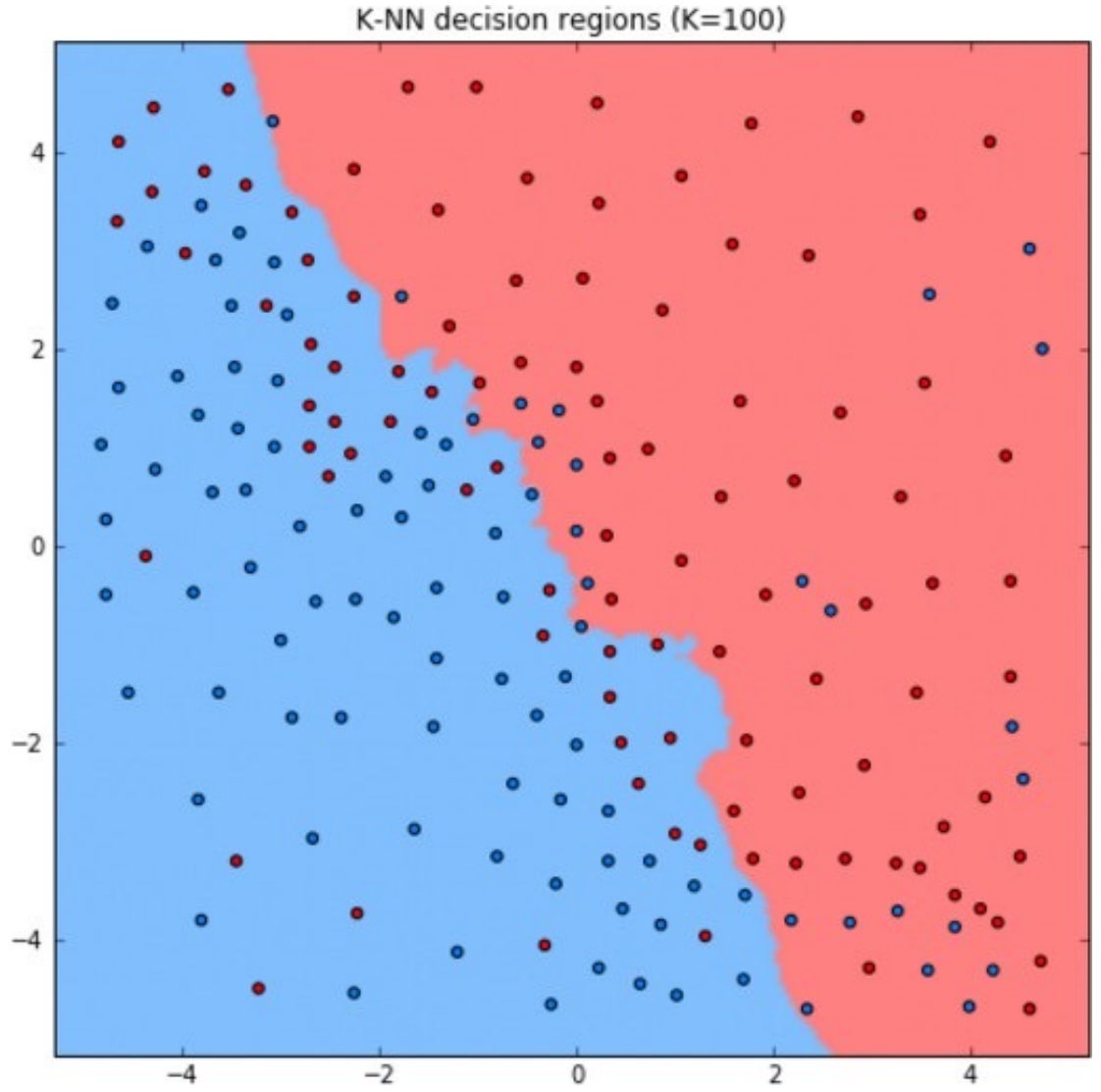
Example: K-NN classification

- Decision boundaries or Decision regions
- When k = 10



Example: K-NN classification

- Decision boundaries or Decision regions
- When k = 100



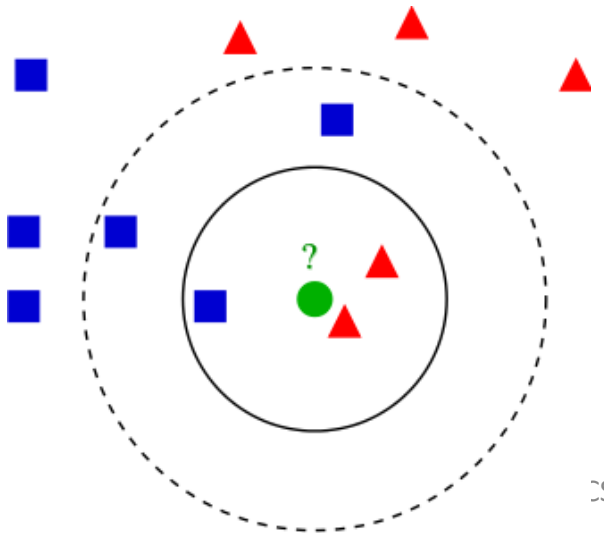
LONDON METROPOLITAN UNIVERSITY

k-Neighbours Regression

k-Nearest Neighbours Algorithm

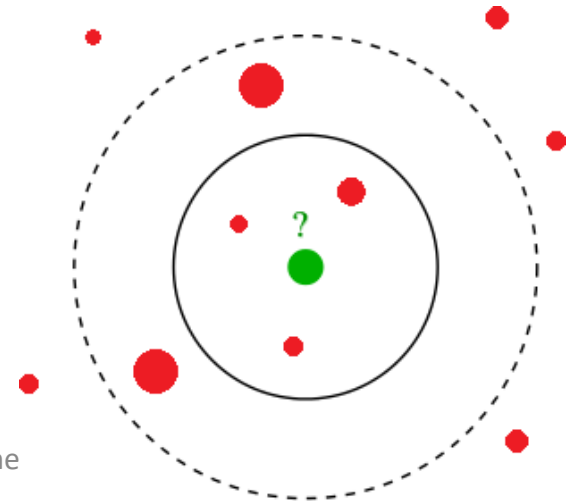
Classification:

1. Find k closest objects to the predicted object x in the training set.
2. Associate x the most frequent class among its k neighbours.



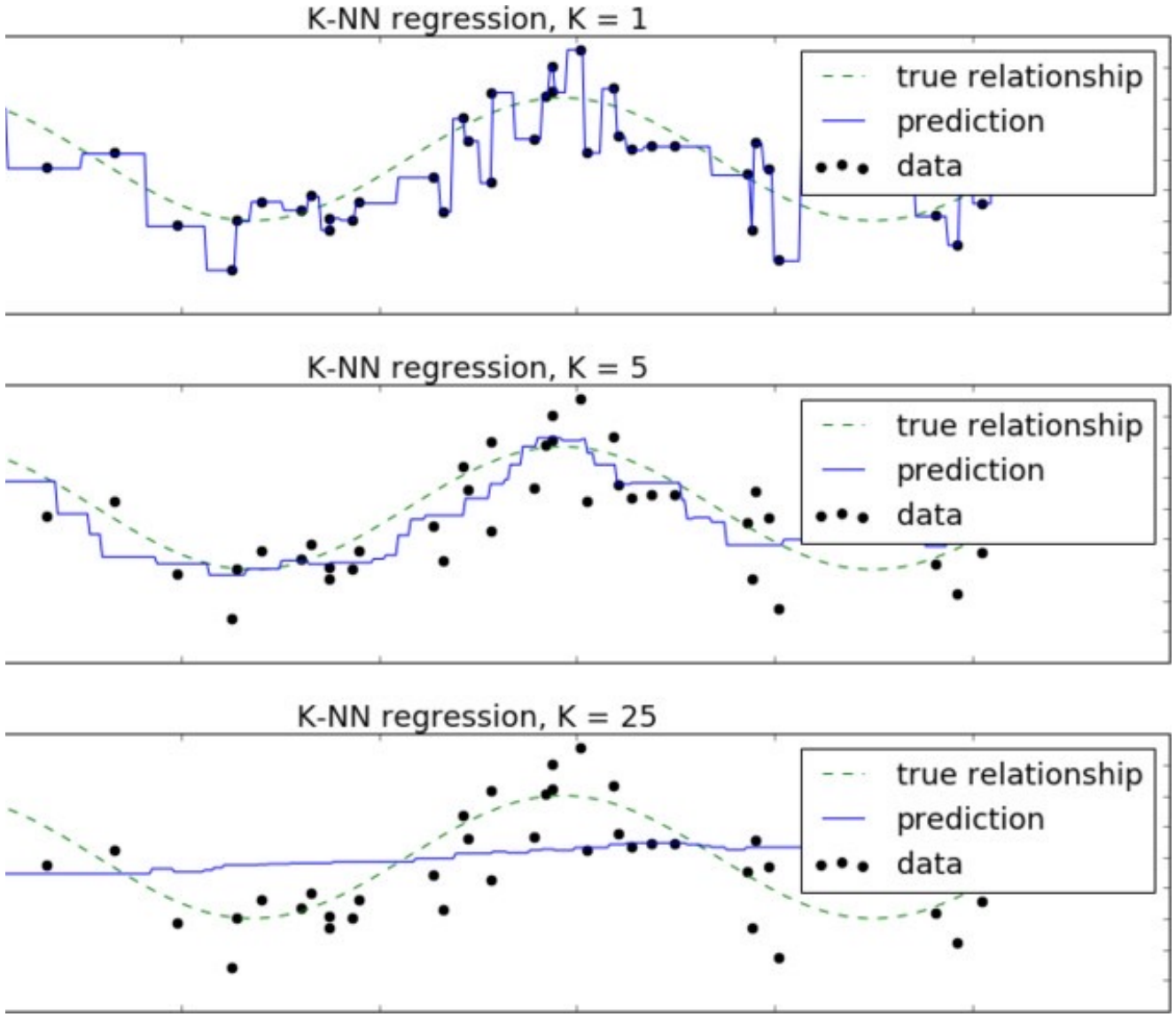
Regression:

1. Find k closest objects to the predicted object x in the training set.
2. Associate x average output of its k neighbours.



K-NN Regression

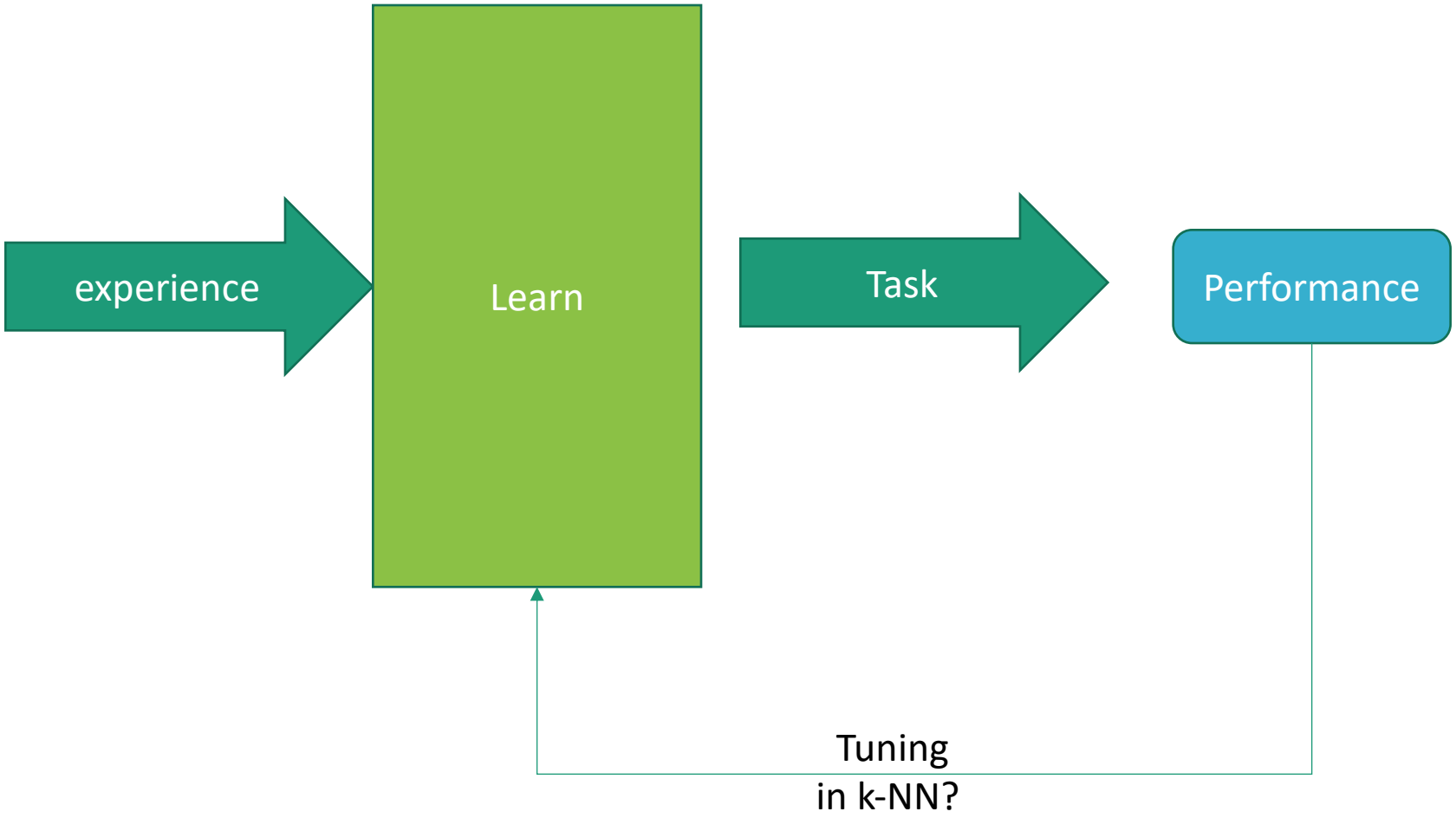
- Can we predict numerical values using k-NN Algorithm?



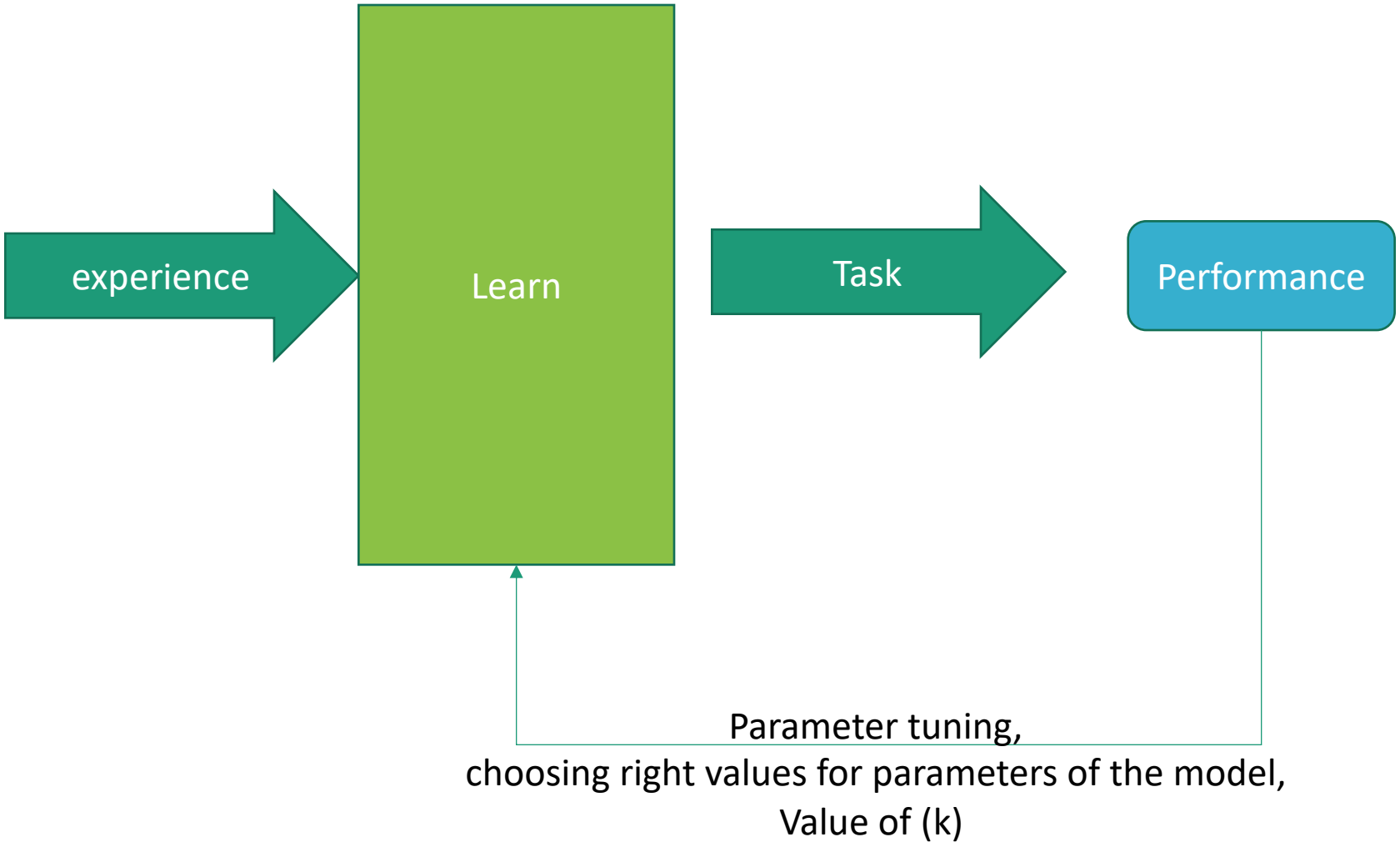
Parameters

- There are two parameters in k-NN:
  - The number of neighbours
  - How do you measure distance between data points?
    - Euclidian distance
    - Other?

Optimisation in ML?



Optimisation in ML?



Strengths and weaknesses of k-NN

- Easy to understand and interpret
- Building the model is fast, easy to implement
- Does not need training, may be applied in online scenarios
- Prediction can be slow when training data is very large
  - Number of features (hundreds or more)
  - Number of samples
- Accuracy deteriorates with the increase of feature space dimensionality

LONDON METROPOLITAN UNIVERSITY

Overfitting and Underfitting

Overfitting and Underfitting

- What is overfitting?
- What is underfitting?

Generalisation

In supervised learning

- we want to **build a model** on the training data and then
- Be able to make accurate **predictions** on **new, unseen data** that has the same characteristics as the training set that we used.

- If a model is able to make **accurate** predictions on unseen data
- we say it is able to **generalise** from the training set to the test set.

Accuracy of a model

- We want to build a model that is able to **generalise** as **accurately** as possible.

- We build a model that is accurate on training data set and then
- We hope that it is accurate on test set

- Accurate on
  - Train data
  - Test data

Build a model that is accurate on train data

- Target data: Bought a boat
- Feature: Age, …. , Owns a dog

Table 2-1. Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Build a model for me that is accurate on train data

- Everybody who owns a house buys a boat".

- But what about accuracy on test data?

Table 2-1. Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Build a model for me that is accurate on train data

- Everybody who owns a house buys a boat".
- Anyone over 52? 100% accurate on train data

Table 2-1. Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Build a model for me that is accurate on train data

- Everybody who owns a house buys a boat".
- Anyone over 52? 100% accurate on train data
- "If the customer is older than 45 and has less than 3 children or is not divorced, then they want to buy a boat."
- But what about accuracy on test data?

Table 2-1. Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

Overfitting and Underfitting

- Choosing too simple a model is called underfitting.
  - It is not even good on train data

- Overfitting occurs when you fit a model too closely to the particularities of the training set
  - High accuracy on train data but not on test data

Model Complexity vs. Accuracy

- Underfitting
- Overfitting

- The sweet spot

- Reference: Muller & Guido's book page 31

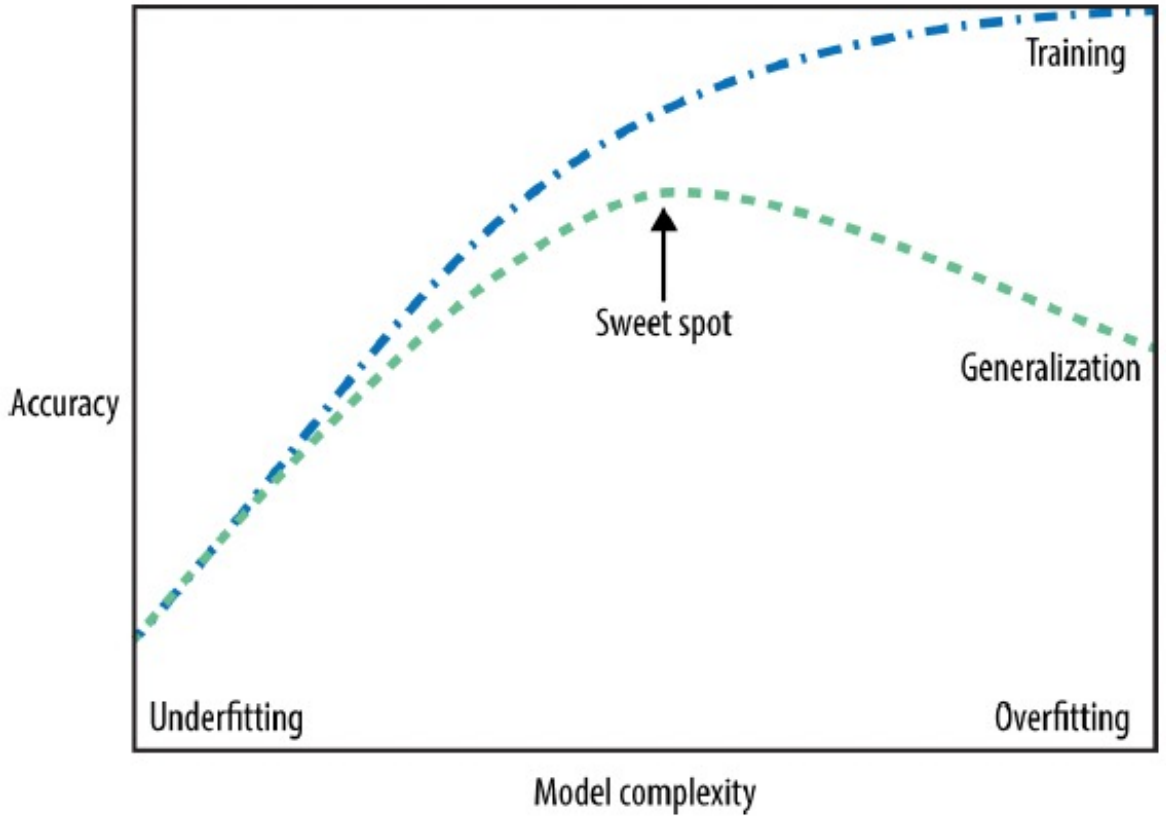


Figure 2-1. Trade-off of model complexity against training and test accuracy

The sweet spot

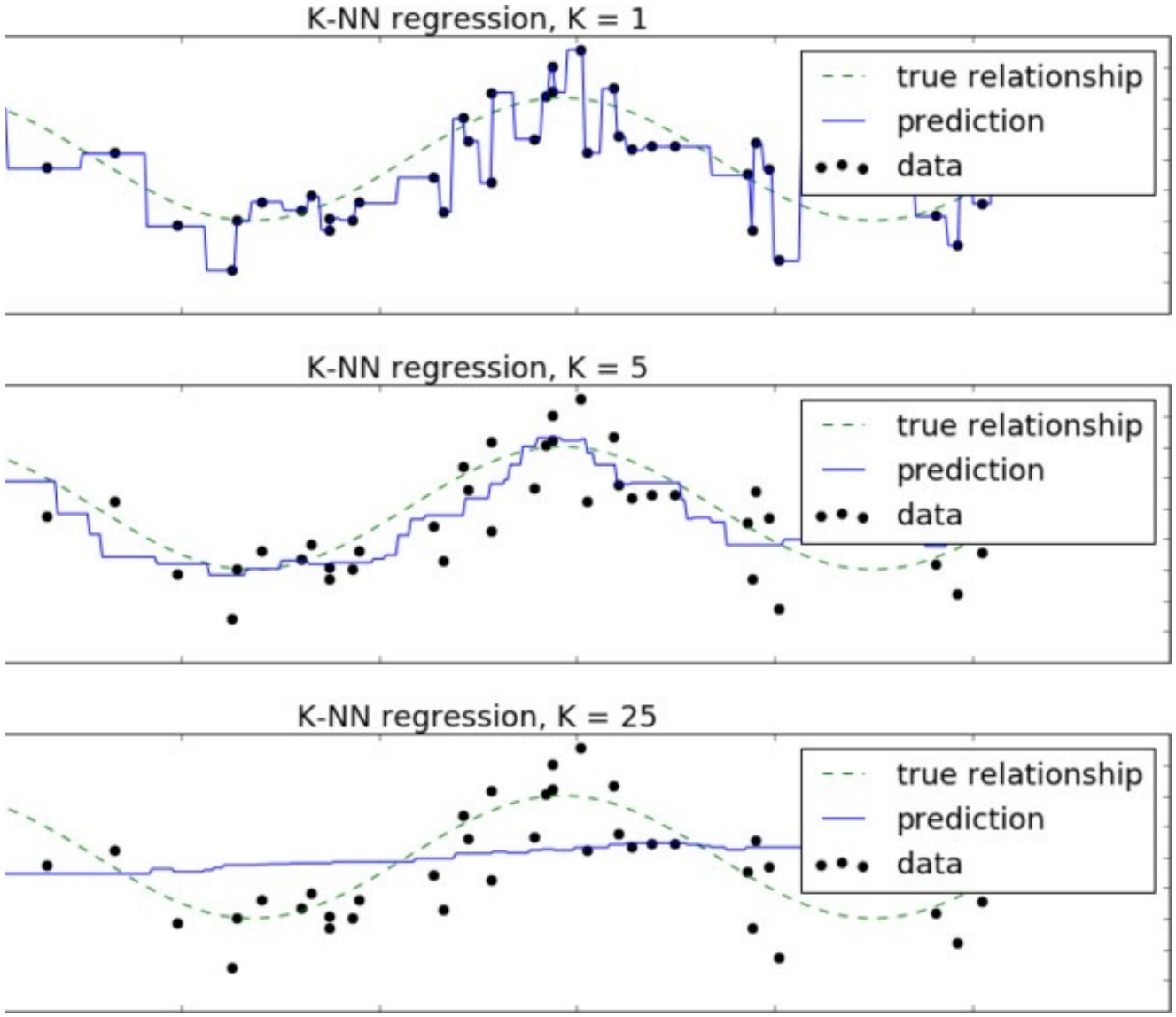
- There is a sweet spot in between that will yield the best generalization performance.

- This is the model we want to find.

K-NN Regression

- Can you identify which model is overfitting and which one is underfitting?

- Where is the sweet spot here?



Last weeks workshops

- Workshop 1:
  - Numpy
  - Pandas
  - Exploring Titanic dataset with Pandas
- Workshop 2
  - Iris species
  - More on pandas, reading and writing data, csv files, UK inflation data 1989-2022
  - Pandas in depth, data manipulation (string manipulation)

Iris dataset

- source: Muller and Guido's book, page 20

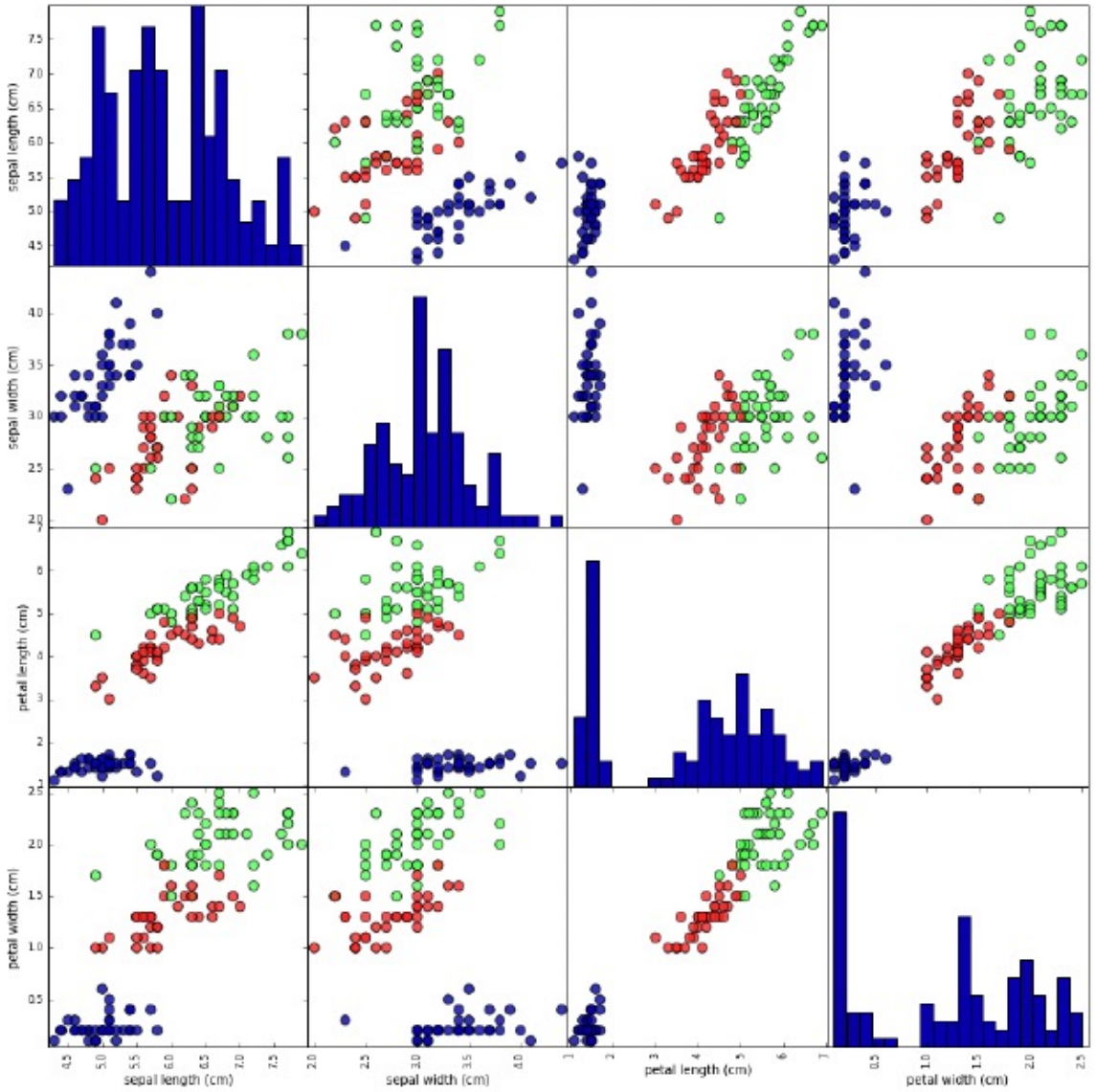


Figure 1-3. Pair plot of the Iris dataset, colored by class label

LONDON METROPOLITAN UNIVERSITY

Textbook chapters covered so far

- Nelli’s book, Chapters 1, 2, 3, 4 , 5 and 6
- Muller and Guido’s book Chapter 1 and 2 (partly)

LONDON METROPOLITAN UNIVERSITY

Summary

- Supervised learning, K-Nearest Neighbours (k-NN)
  - KNN Classification
  - KNN Regression
  - Overfitting and underfitting

Workshop today, workshop 3

- KNN, Chapter 2 Muller and Guido's book
- Complete workshop 2, task 2 (chapter 5 and 6 of Nelli's book)