https://edin.ac/3WQHRYm

Inf2 - Foundations of Data Science
Introduction to supervised learning - Classification

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Announcements

- Classification with Nearest neighbours
  (Lecture notes chapter 6)
- k-Nearest neighbour classification,
  setting hyperparameters, and metrics
  (Lecture notes chapter 7)
- Lab: k-Nearest Neighbours with scikit-learn
- Workshop: Critical evaluation and multiple
  regression practice
  - Critical evaluation assessed in the exam!
- Schedule change - feedback on formative visualisation
  and interpretation moved to Monday Week 11.
  - No lecture on Wed of Week 10

This week – Nearest Neighbours and Evaluation

- Today
  - Classification as an example of a supervised learning task
  - Nearest neighbour classification
  - Evaluating classifier performance
- Wednesday
  - k-Nearest Neigbours classifiction
  - Evaluation – Metrics
  - Evaluation – Cross-validation

The data science process

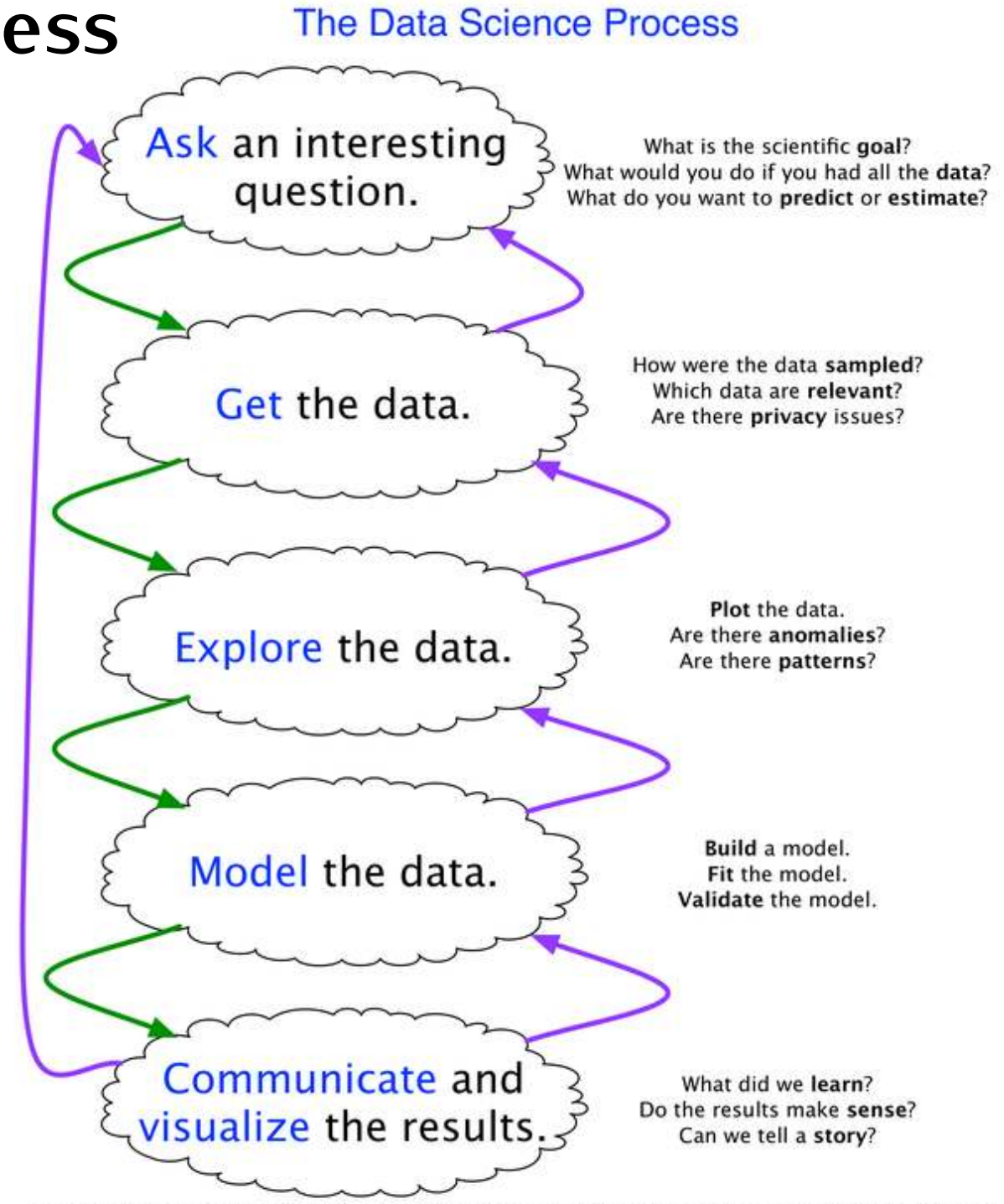
I. Data: ethics, collection, representation, wrangling, exploration, visualisation and descriptive stats

II. Linear models

III. Intro to Machine Learning

IV. Statistical inference

V. Regression and inference



The classification problem

Imagine you are a bank manager...

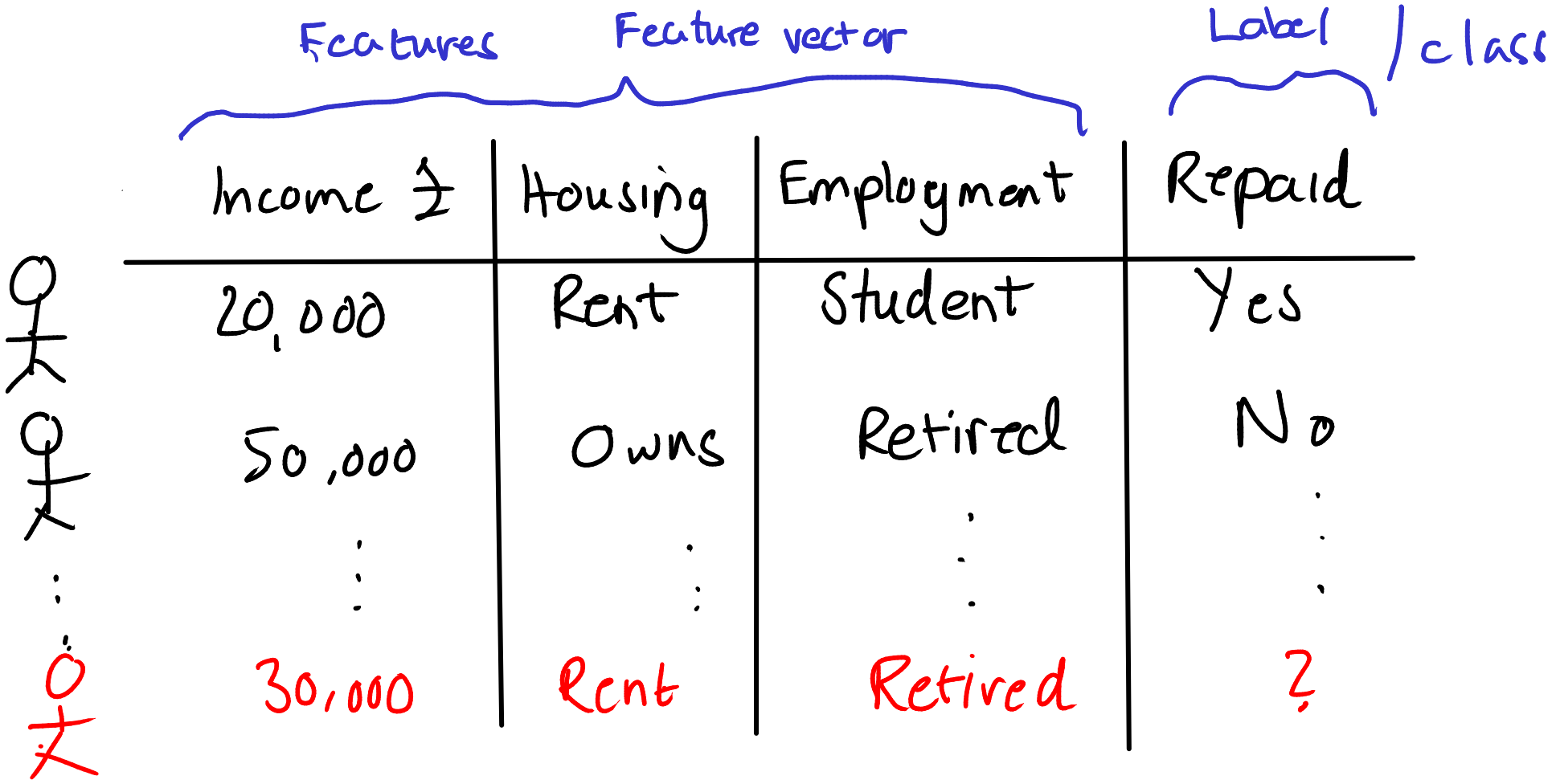
Income	Housing	Employment
£30 k	Rent	Retired

- Do you predict this customer would repay a loan of £10,000?
- How good is your prediction?
- Is your predicton method fair to the people it affects?

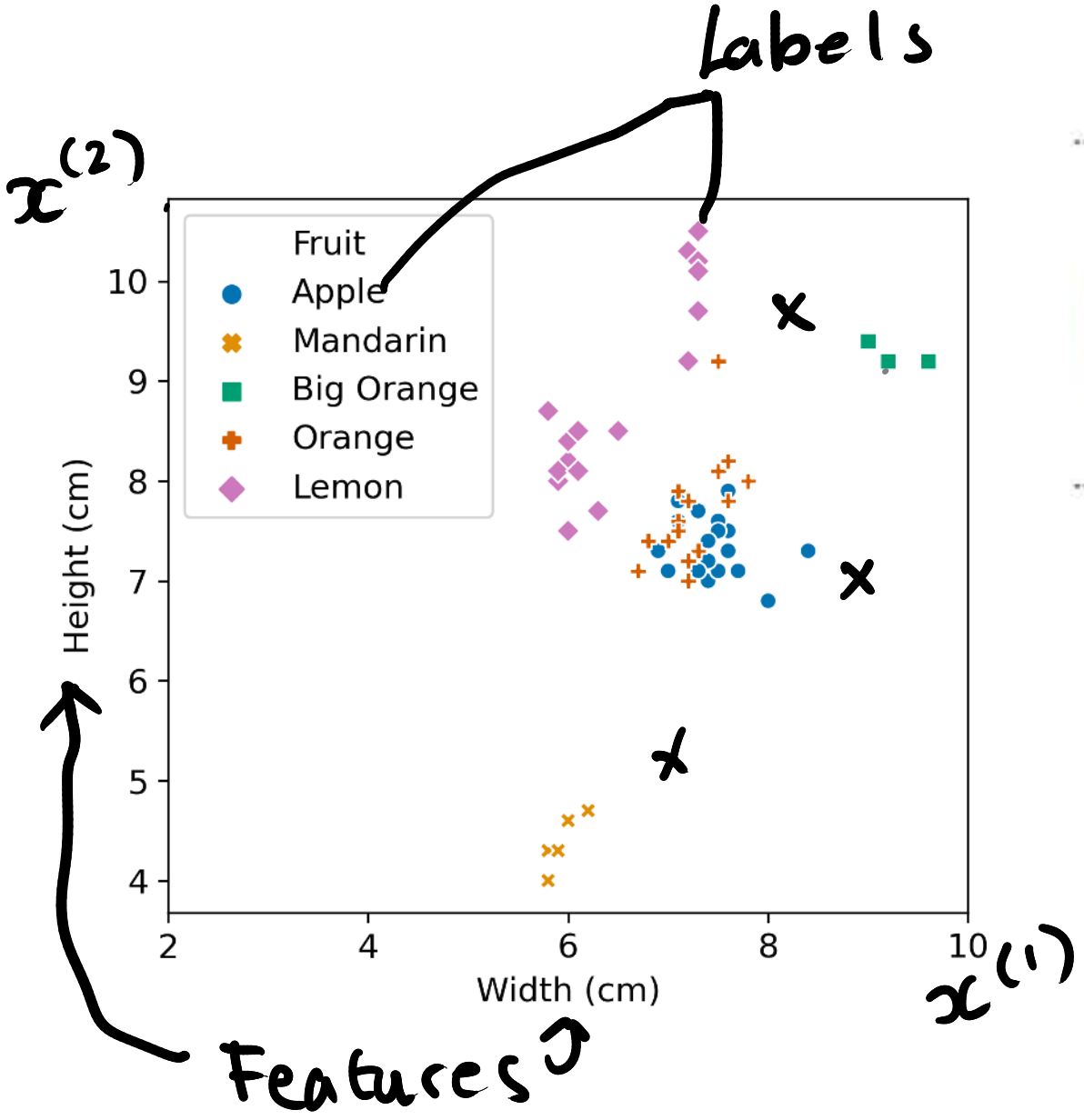
Overview

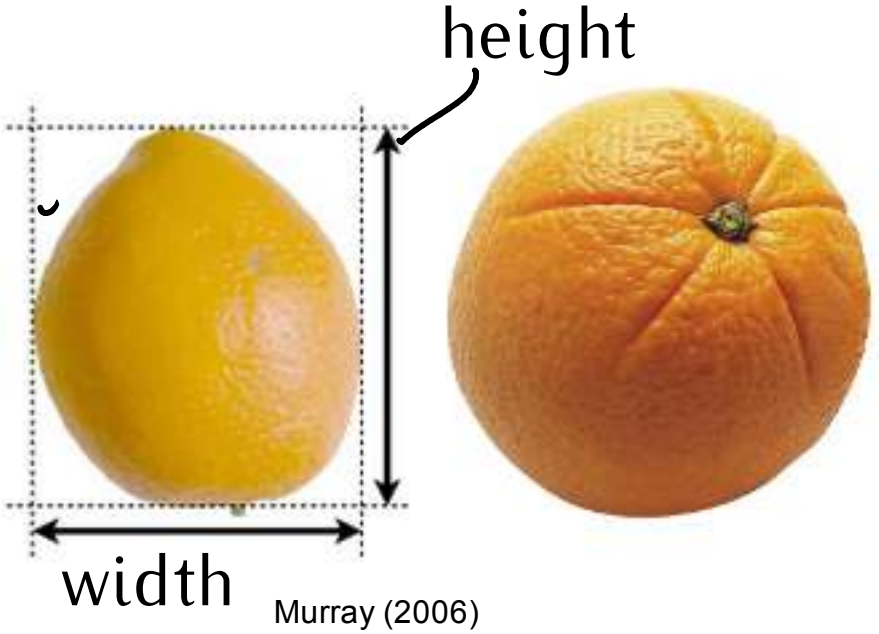
- The problem of classification
- The principle of the logistic regression classifier
- The 1-nearest neighbour classifier
- Evaluation

Supervised learning of classifications



Visualising the classification problem





Murray (2006)

Definition of a classifier

Feature vector $\underline{x} = (x^{(1)}, \ldots, x^{(D)})^T$

$C = \{$ Apple, Orange, Lemon, ... $\}$

Properties

1. Decision boundaries
2. Classification errors

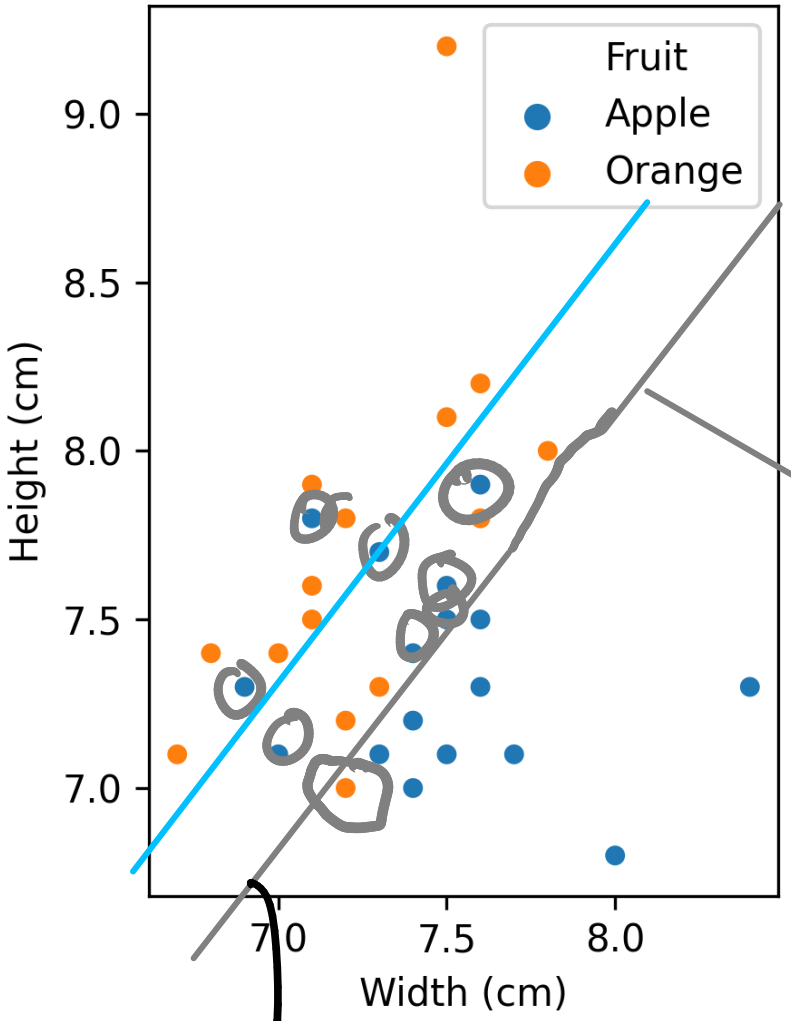
Inf2 - Foundations of Data Science: Intro to supervised learning - Principle of the logistic regression classifier

THE UNIVERSITY *of* EDINBURGH

informatics

FOUNDATIONS OF DATA SCIENCE

Constructing a classifier by hand



w_1 w_0

$x^{(2)} > \beta_1 x^{(1)} + \beta_0$

$\Rightarrow$ Orange

Otherwise

$\Rightarrow$ Apple

$x^{(2)} = \beta_1 x^{(1)} + \beta_0$

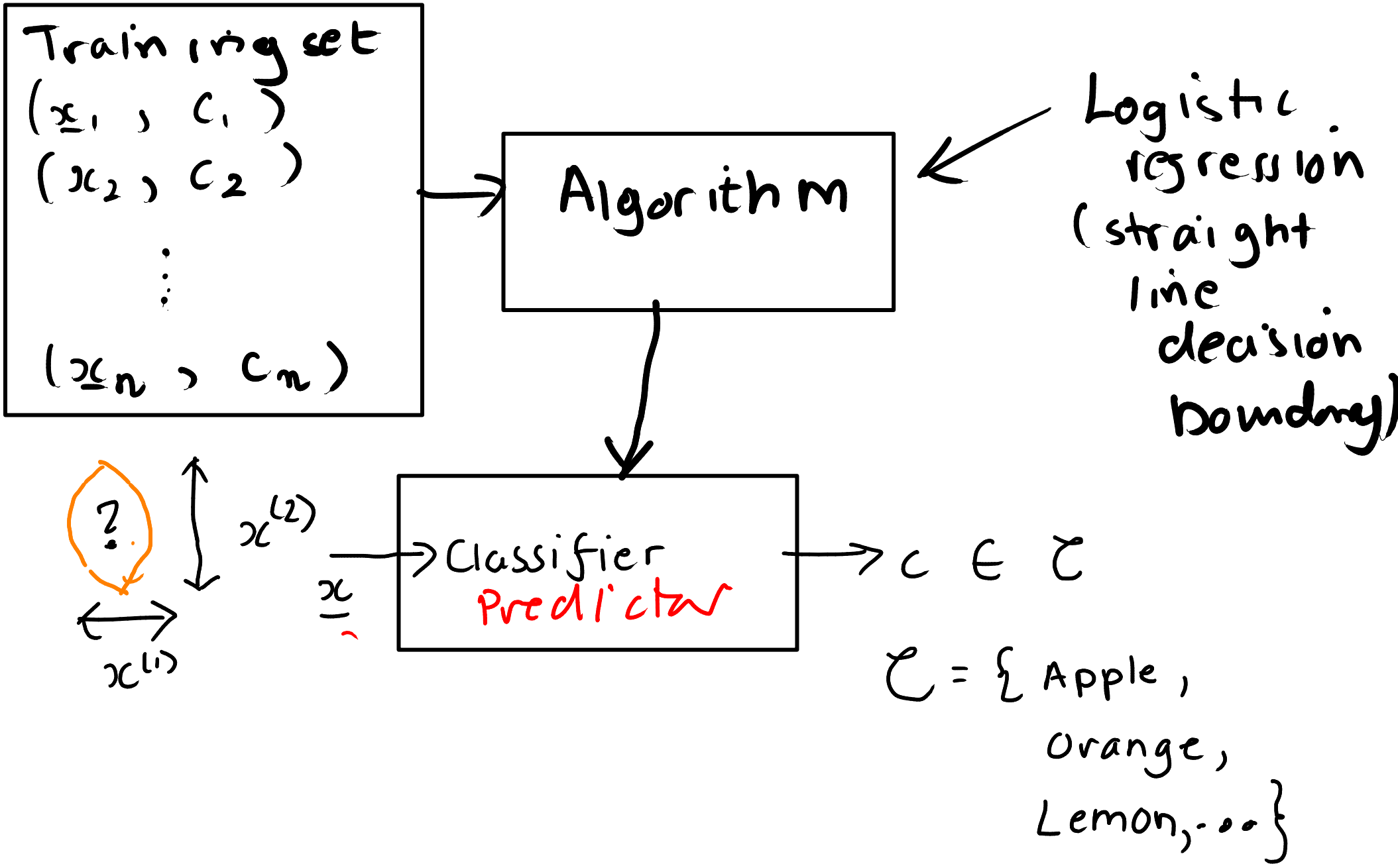
misclassification.

$1 + 8 = 9$ misclass.

$2 + 5 = 7$ "

Decision boundary

Constructing a classifer by supervised learning



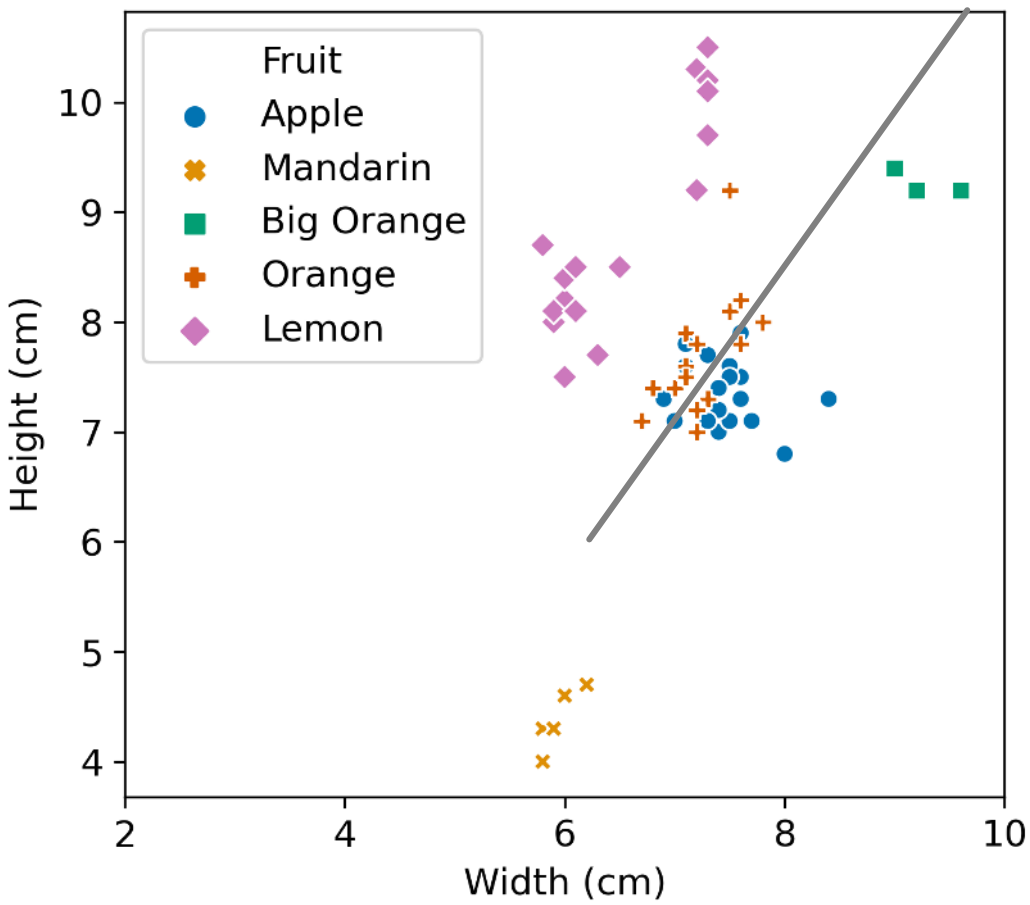
Inf2 - Foundations of Data Science: Intro to supervised learning - Nearest neighbour classification

THE UNIVERSITY *of* EDINBURGH

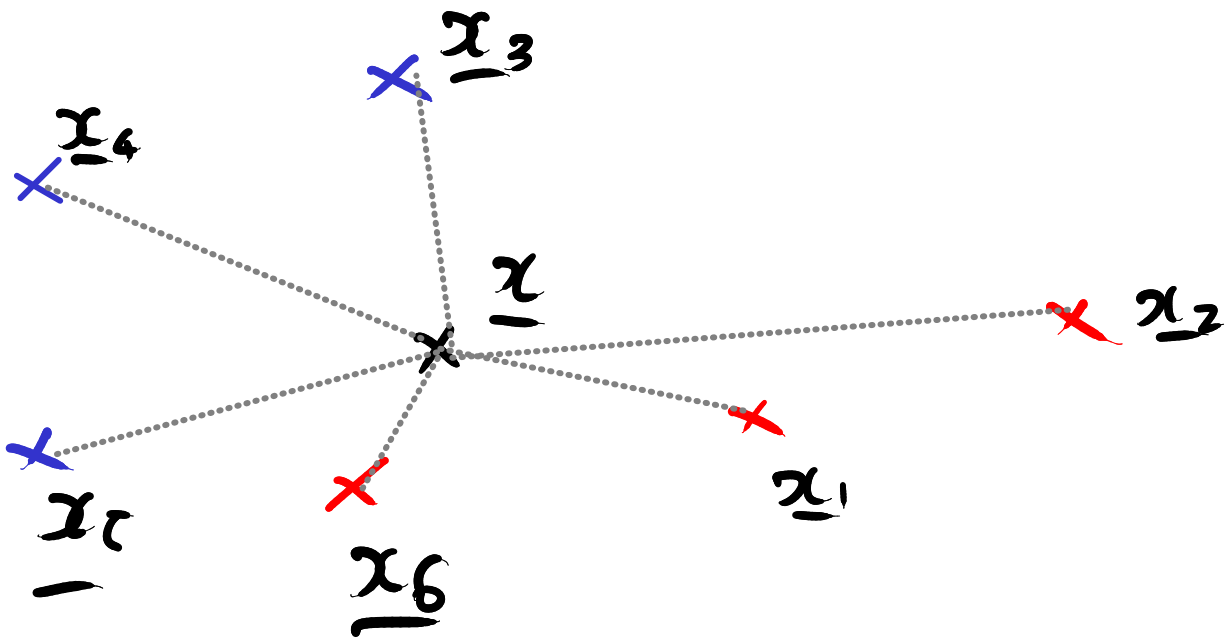
informatics

FOUNDATIONS OF DATA SCIENCE

Straight decision boundaries don't fit the data



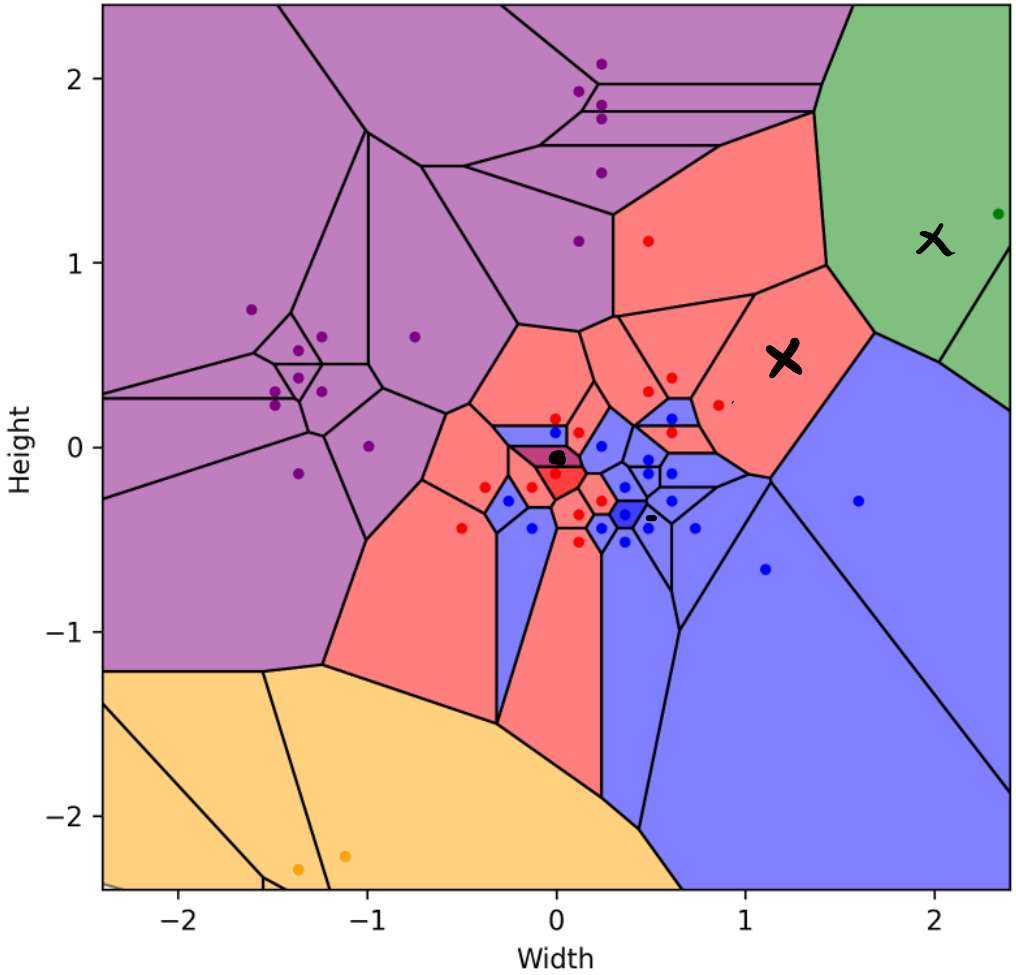
Principle of nearest neighbour classification



Euclidean distance

$$d(x, x_i) = \sqrt{(x_i^{(1)} - x^{(1)})^2 + (x_i^{(2)} - x^{(2)})^2}$$

Nearest neighbour classification of fruit



- Standardised variables
- Often better for distance-based algorithms

Inf2 - Foundations of Data Science: Intro to supervised learning - Evaluation

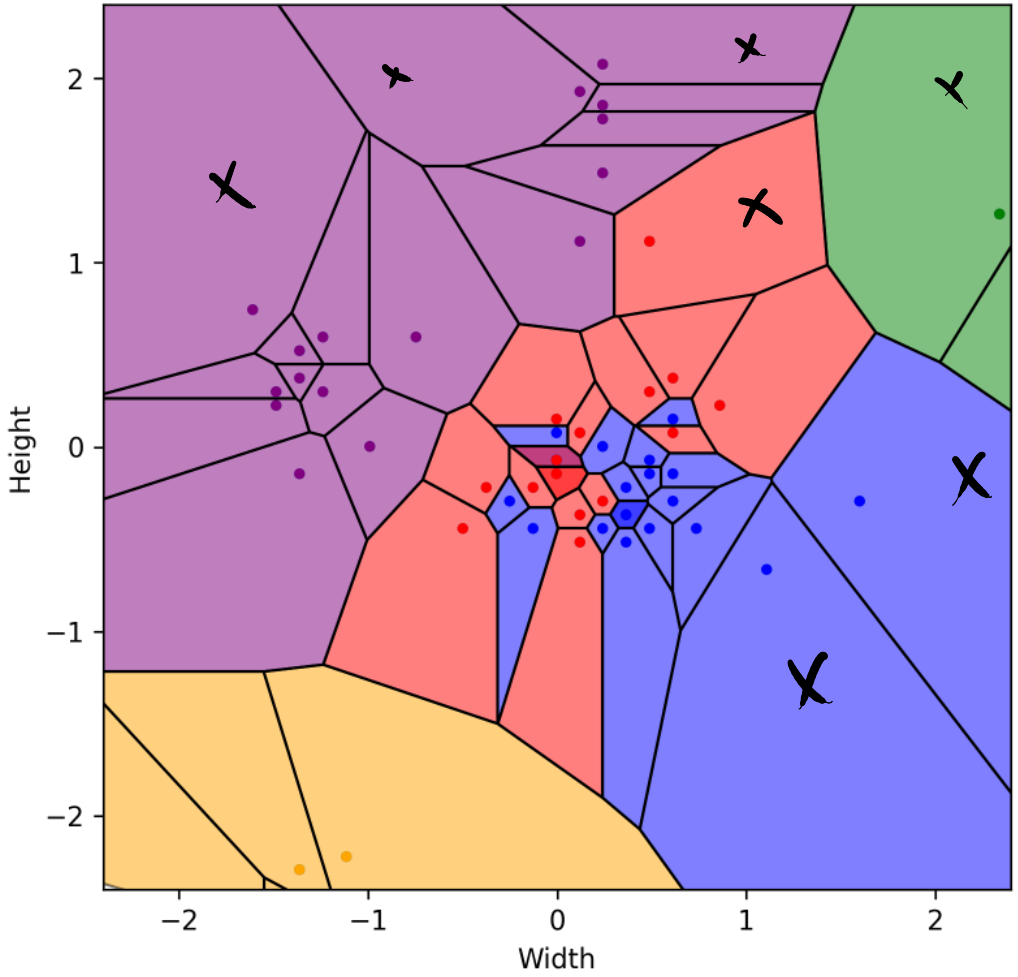
THE UNIVERSITY *of* EDINBURGH

informatics

FOUNDATIONS OF DATA SCIENCE

One metric: classifiction error rate

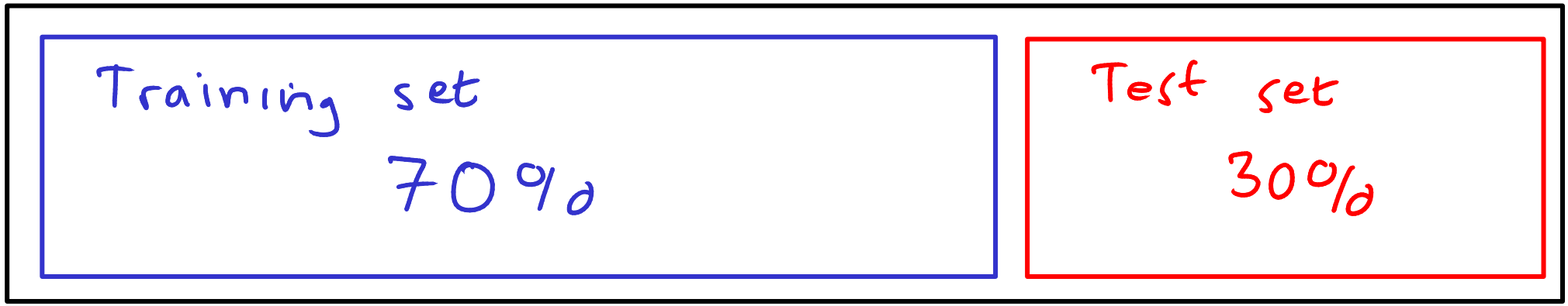
$$\text{Classification error rate} = \frac{\#\text{ misclassifications}}{\#\text{ examples}}$$



What is the classification error rate on the training set?

What error rate be on a randomly selected piece of friut not in the training set?

Evaluating generalisation to unseen data



1-NN

Training set error $\simeq 0$

Test set error > 0

Never use the test set to train the classifier and then quote the error based on the test set
The estimate of the error will be too optimistic.

Inf2 - Foundations of Data Science: Intro to supervised learning - k-Nearest Neighbours

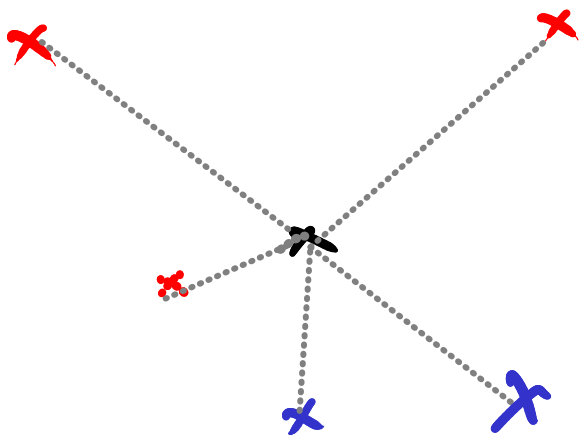
THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Principle of k-NN

To improve generalisation look at...

K- nearest neighbours

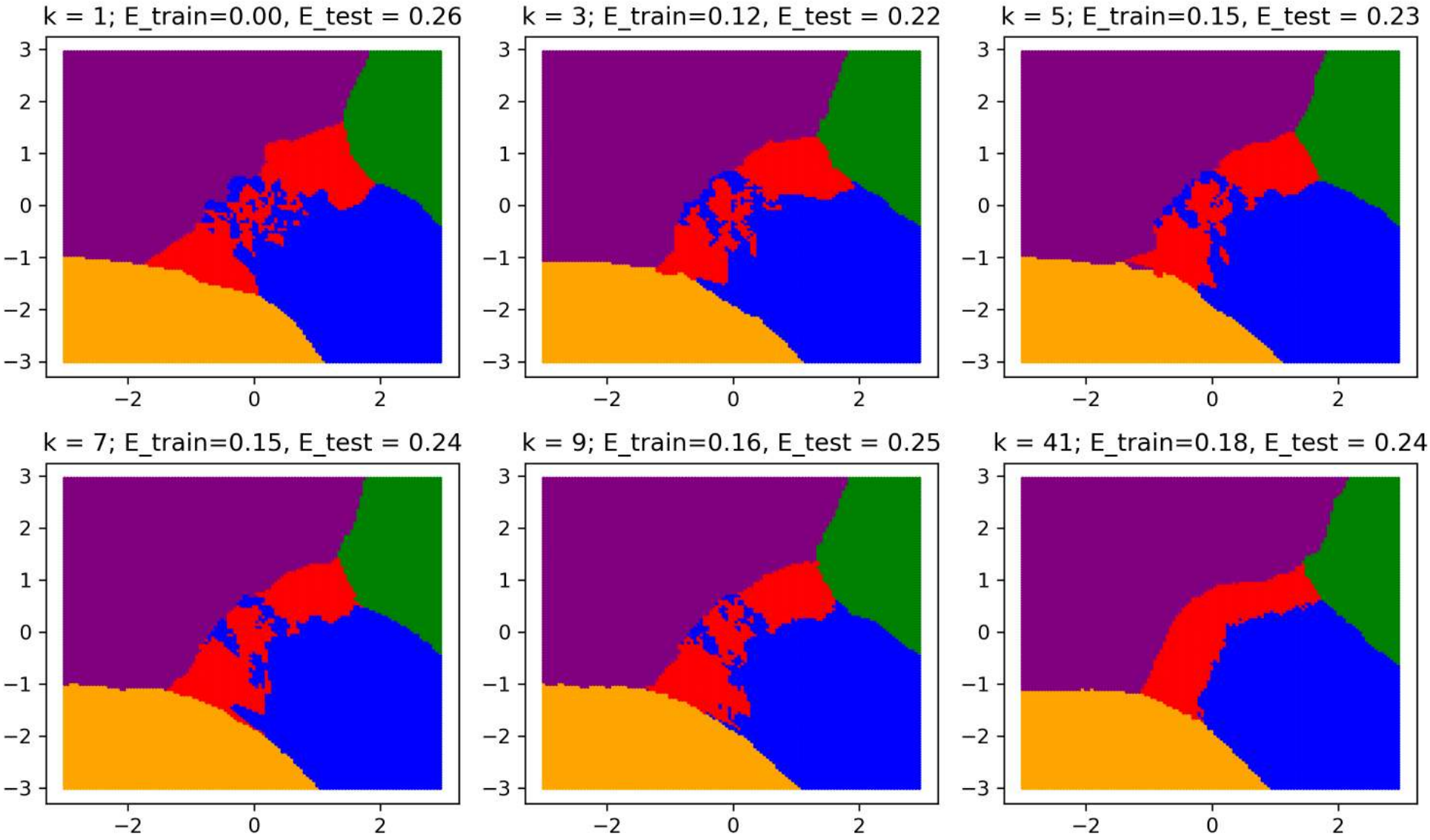


k	R	B	Winner
1	1	0	R
2	1	1	B/R
3	1	2	B
4	2	2	B/R
5	3	2	R

Ties: random resolution or weighting

Odd numbers of k

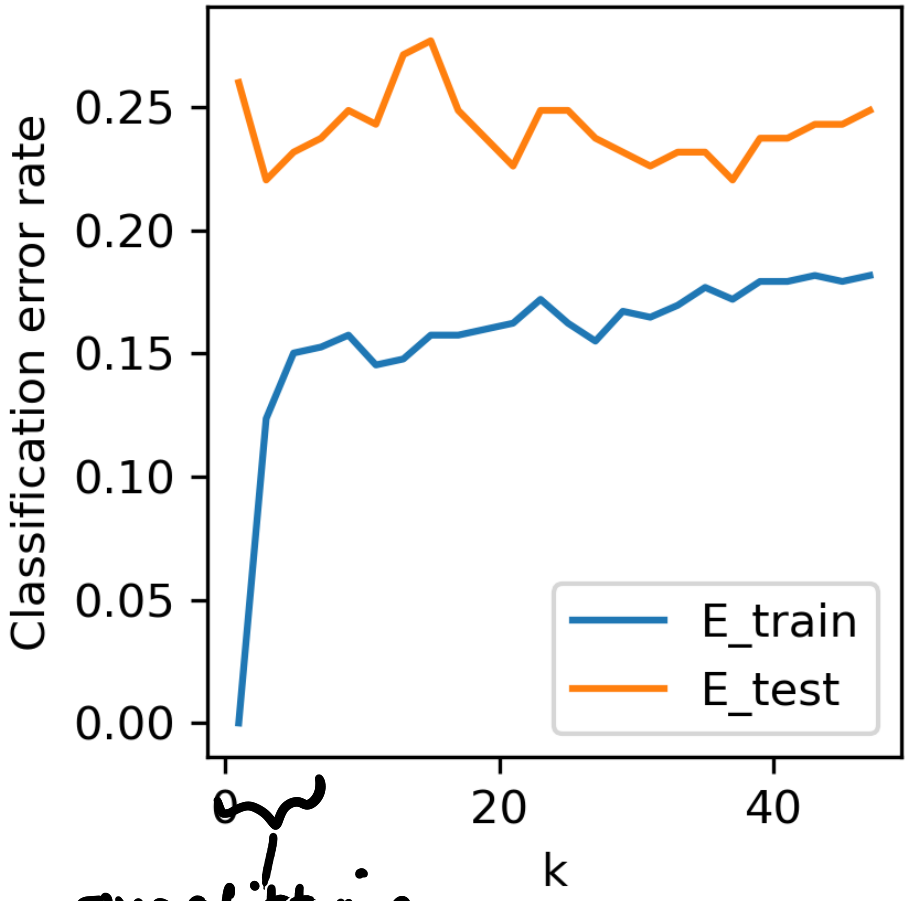
Effect of k-NN



How to pick k?

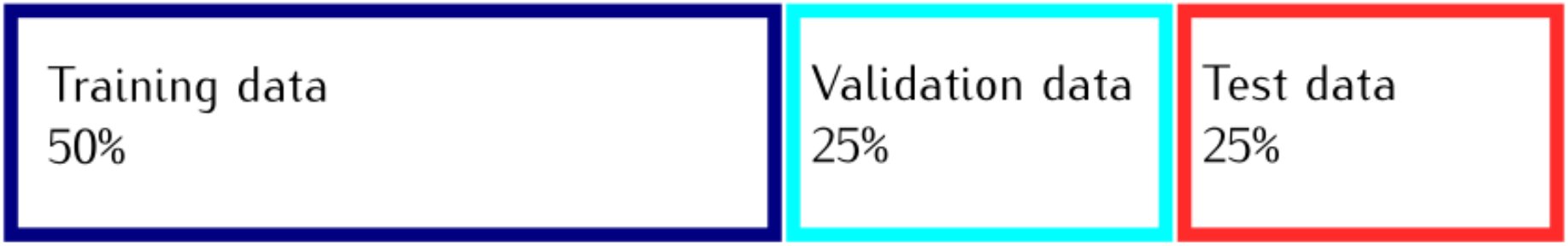
k is a hyperparameter

under fitting
over-generalising



overfitting
under-generalising

Training, validation and testing data



↑

Use to choose hyperparameters

Computational efficiency

- k-NN is efficient at training time

- Simple implementation is slow when classifying an unseen point - distances to k points have to be computed

- Various methods of speeding up k-NNs, such as k-d trees

Summary

- The classification problem is the problem of predicting the class of an unseen item based on its features, and the features and labels of previously seen data (the training data)
- A classifier has a decision boundary and items may be classified correctly or misclassified
- We construct a classifier using a supervised learning algorithm and a training set (and hyperparameters)
- 1-Nearest Neighbours is a simple classification algorithm
- Our evaluation of how well a classifier performs based on training data will be more optimistic than on an unseen test set