

CEILIDH NIGHT

WED 12th NOV, 19:00



TICKETS £5

📍 **Informatics Forum**
No dress code!



THE UNIVERSITY OF EDINBURGH
informatics



<https://edin.ac/3WQHRYm>

Inf2 - Foundations of Data Science

Introduction to supervised learning - Classification



THE UNIVERSITY *of* EDINBURGH
informatics

F O U N D A T I O N S
O F
D A T A
S C I E N C E

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

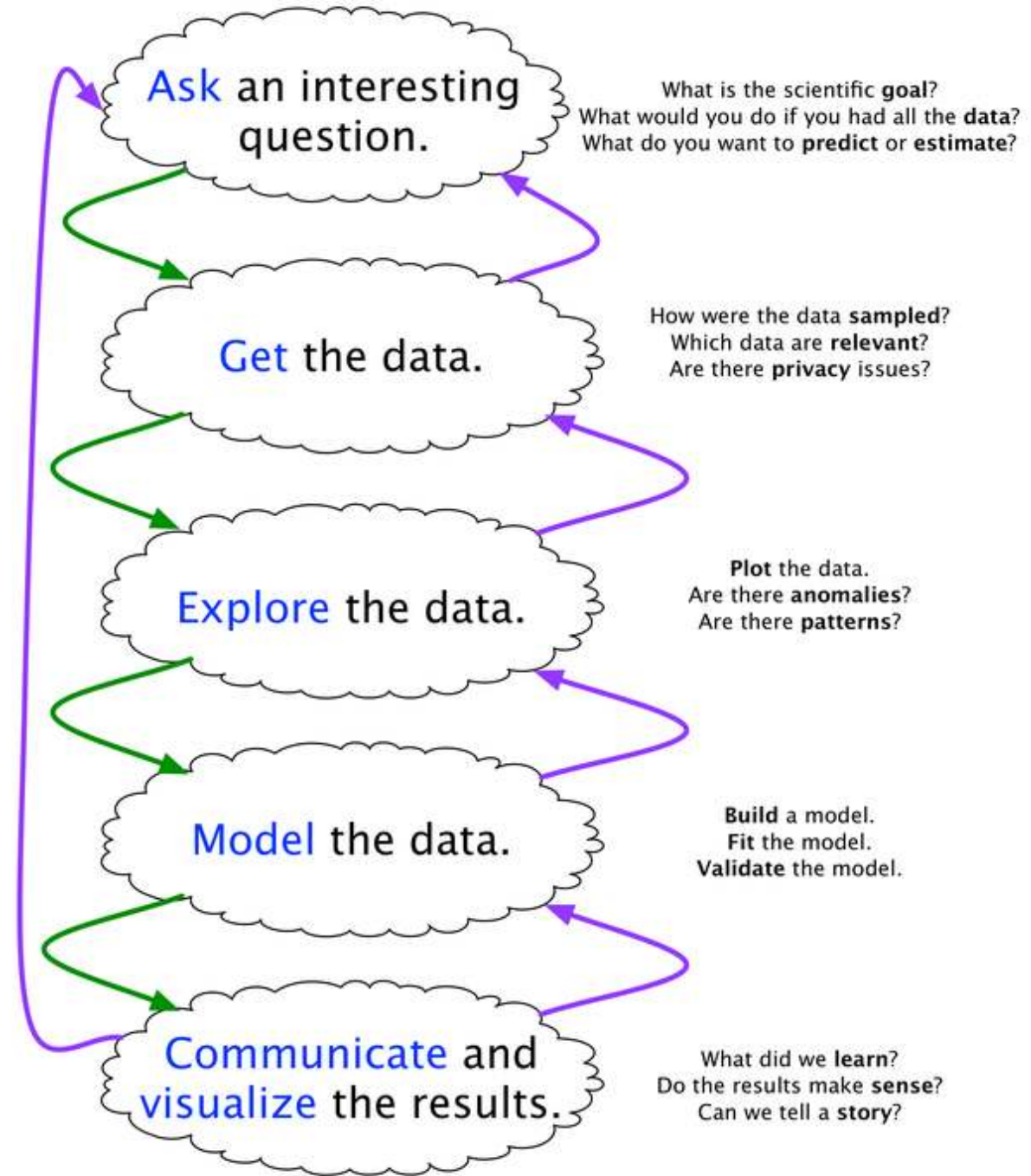
This week – Nearest Neighbours and Evaluation

- Today
 - Classification as an example of a supervised learning task
 - Nearest neighbour classification
 - Evaluating classifier performance
- Wednesday
 - k-Nearest Neighbours classification
 - 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 Data Science Process



The classification problem

Imagine you are a bank manager...



Income
£30k

Housing
Rent

Employment
Retired

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

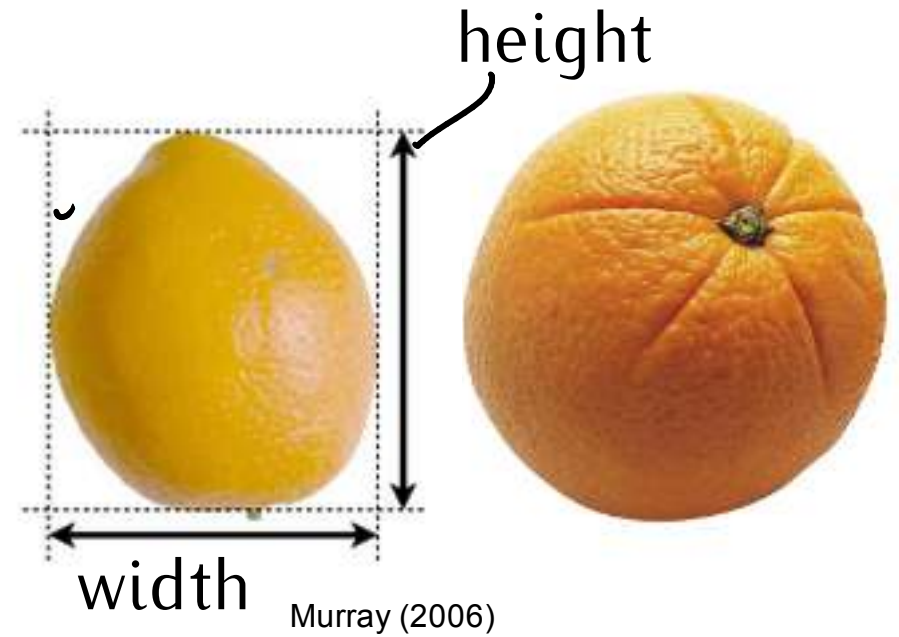
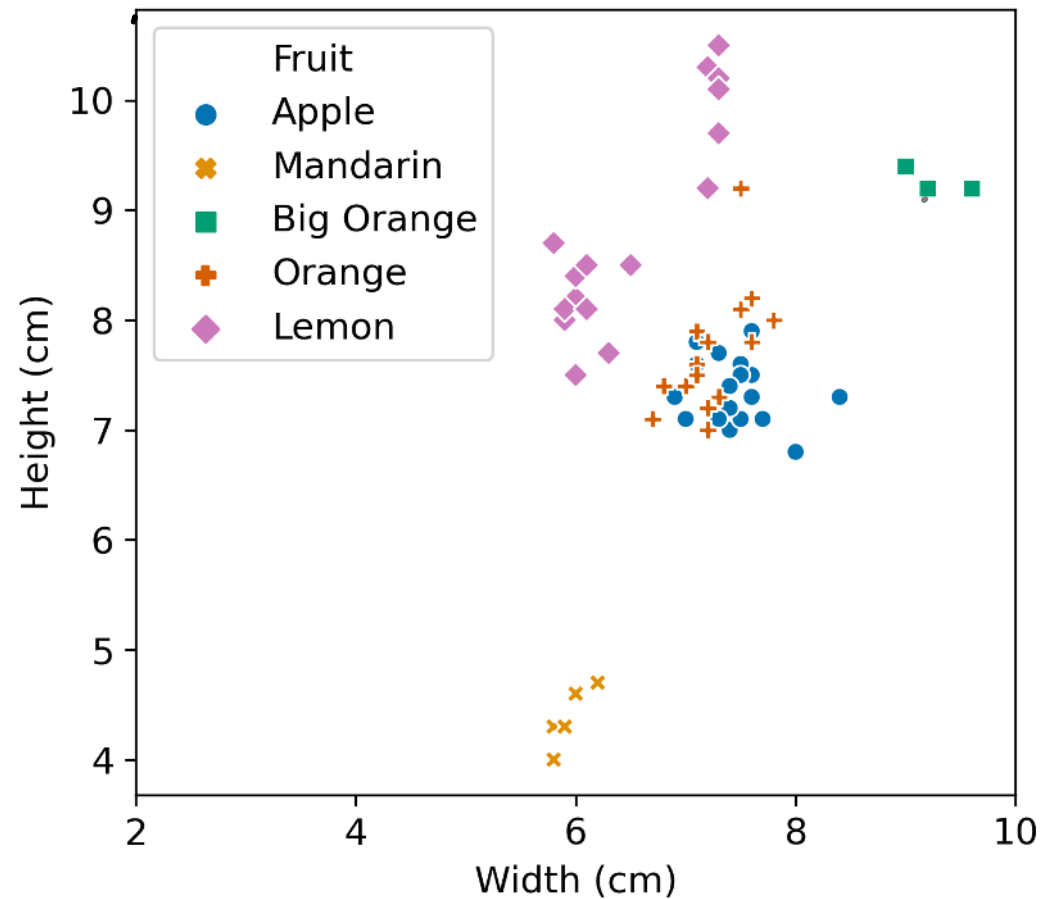
Overview

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

Supervised learning of classifications

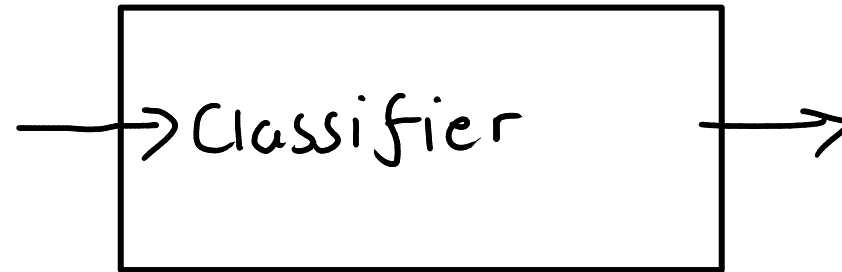
	Features	Feature vector		Label / class
	Income $\hat{=}$	Housing	Employment	Repaid
○ 人	20,000	Rent	Student	Yes
○ 人	50,000	Owns	Retired	No
⋮	⋮	⋮	⋮	⋮
○ 人	30,000	Rent	Retired	?

Visualising the classification problem



Definition of a classifier

Feature vector



Properties

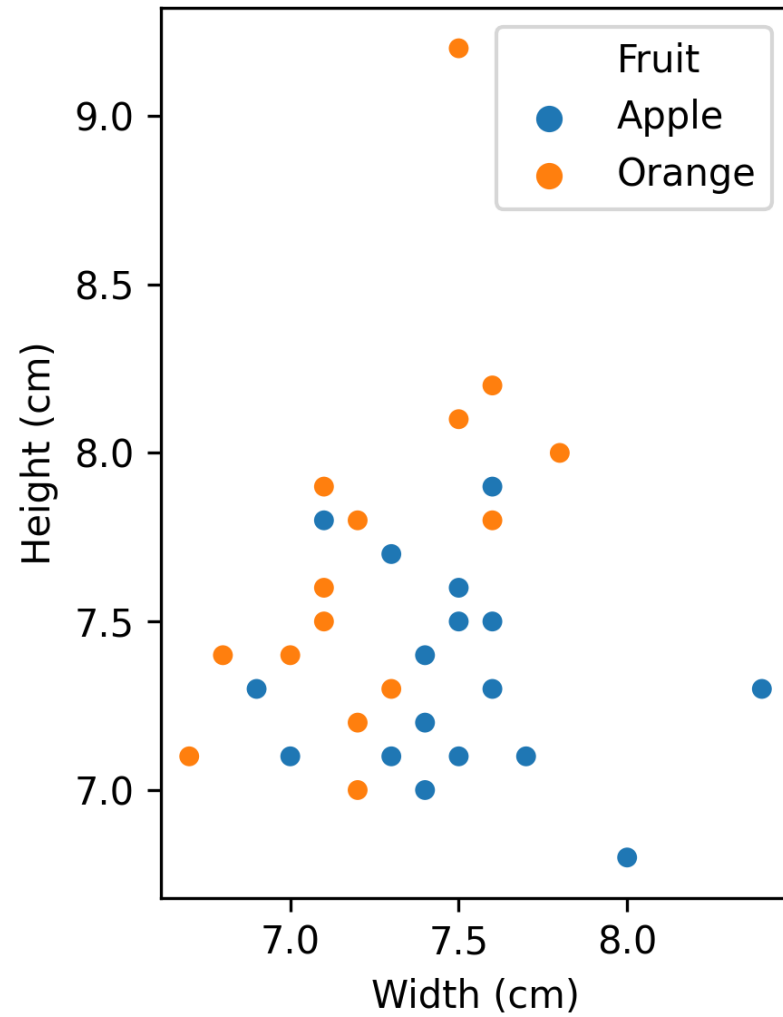
Inf2 - Foundations of Data Science: Intro to supervised learning - Principle of the logistic regression classifier



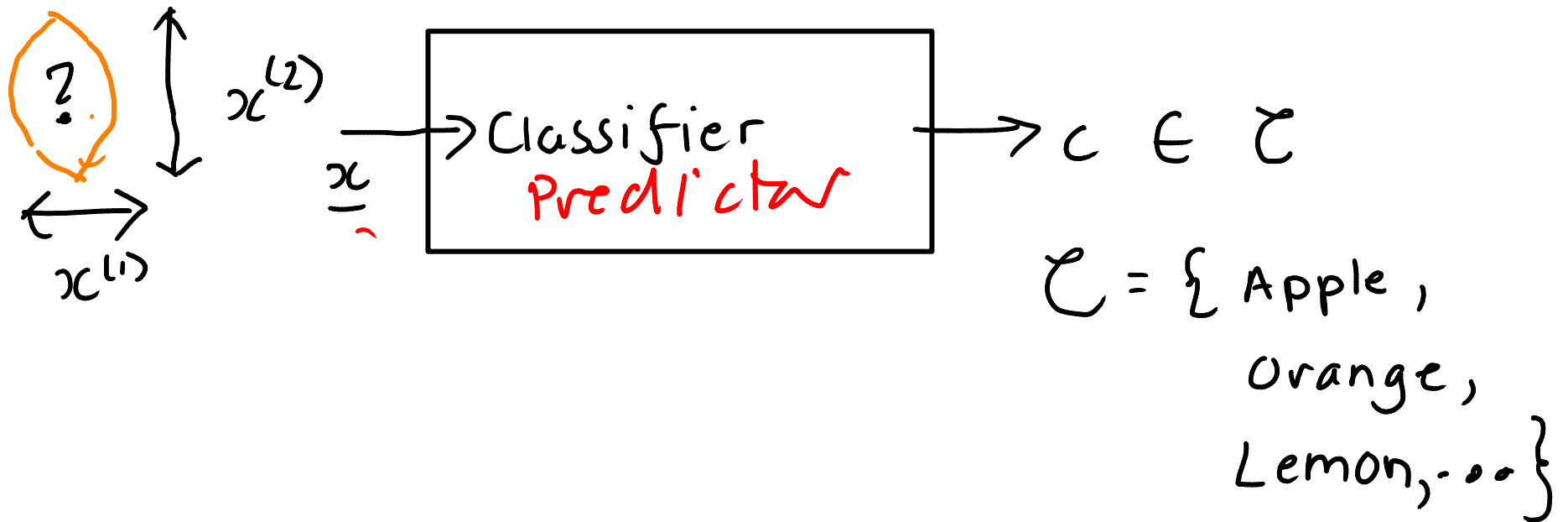
THE UNIVERSITY *of* EDINBURGH
informatics

F O U N D A T I O N S
O F
D A T A
S C I E N C E

Constructing a classifier by hand



Constructing a classifier by supervised learning



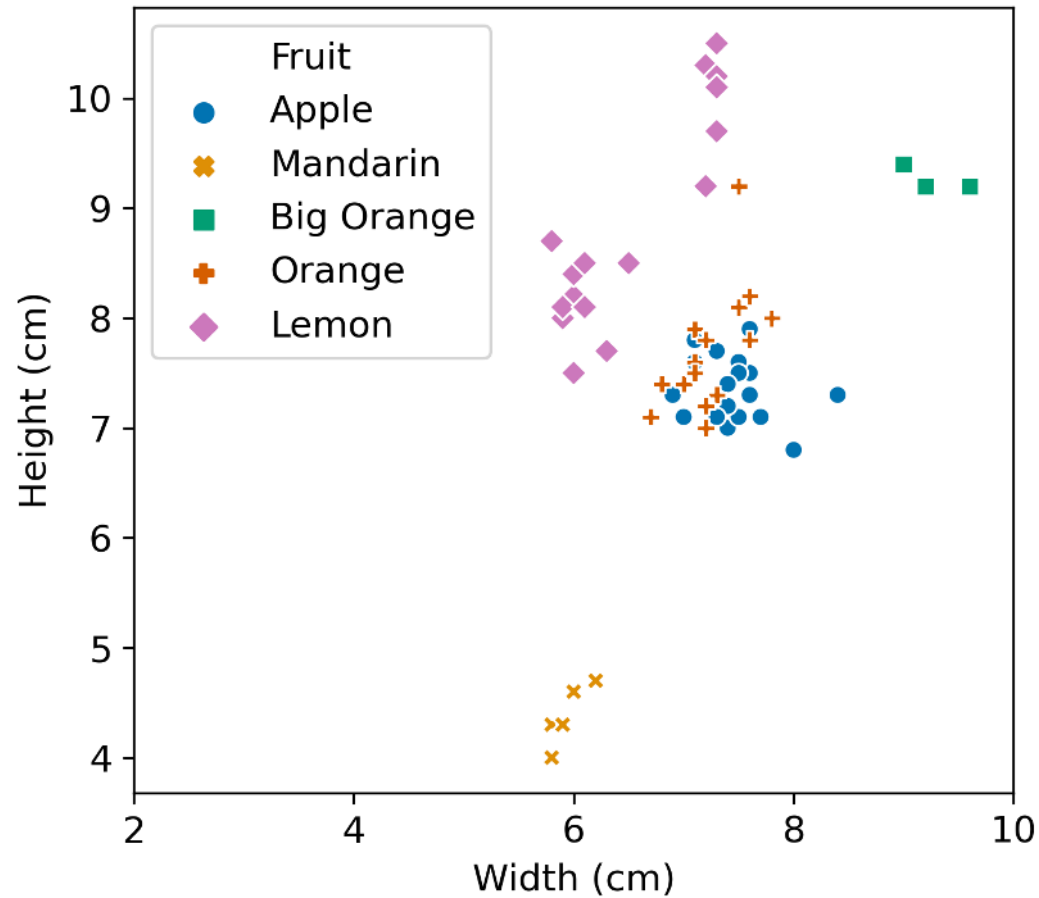
Inf2 - Foundations of Data Science: Intro to supervised learning - Nearest neighbour classification



THE UNIVERSITY *of* EDINBURGH
informatics

F O U N D A T I O N S
O F
D A T A
S C I E N C E

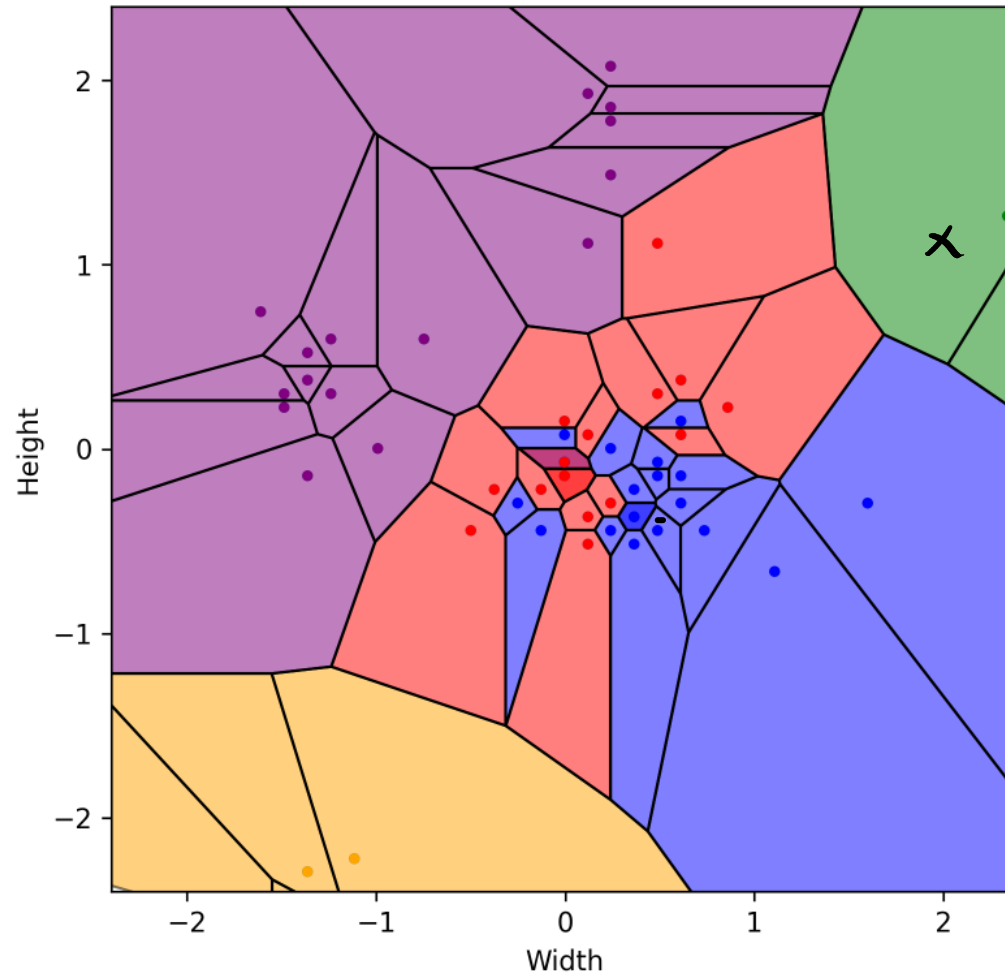
Straight decision boundaries don't fit the data



Principle of nearest neighbour classification

Euclidean distance

Nearest neighbour classification of fruit



- Standardised variables
- Often better for distance-based algorithms

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

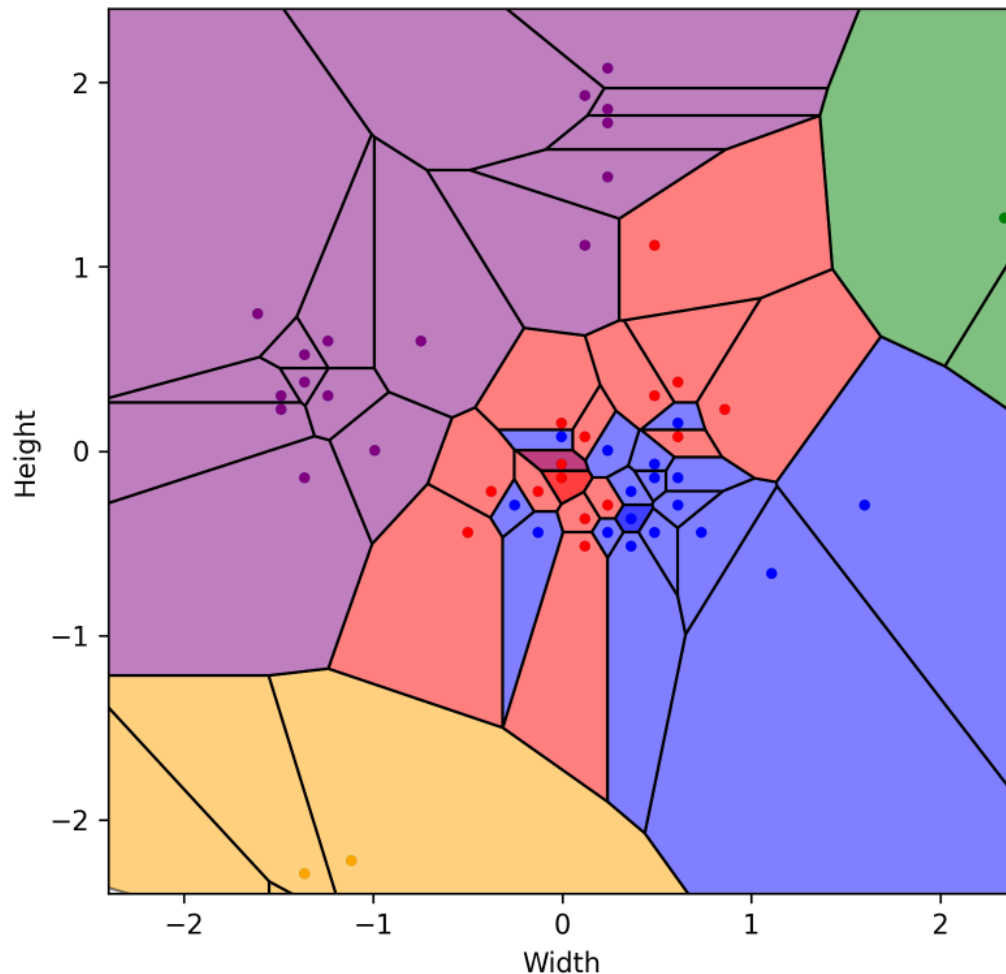


THE UNIVERSITY *of* EDINBURGH
informatics

F O U N D A T I O N S
O F
D A T A
S C I E N C E

One metric: classification error rate

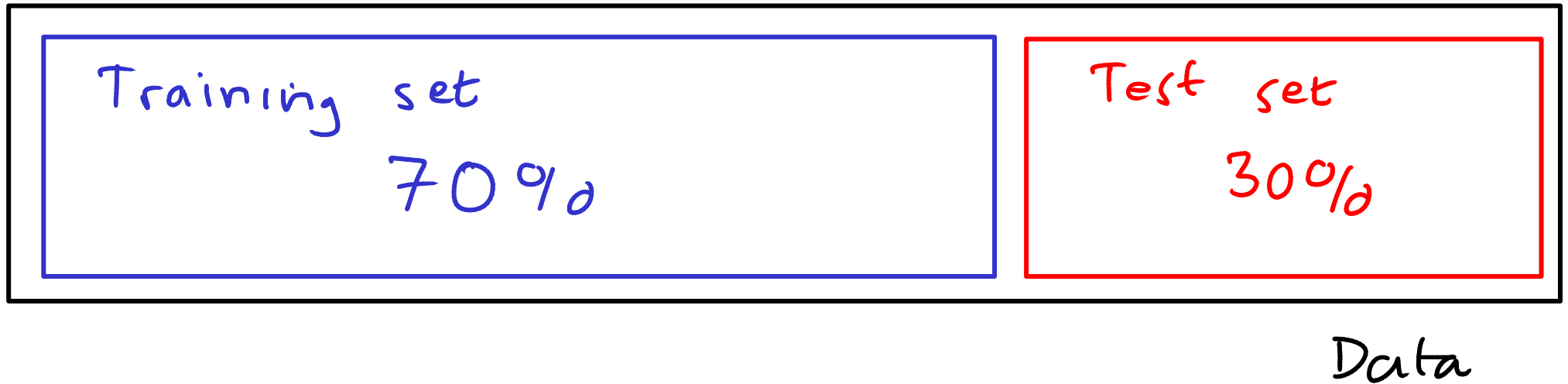
$$\text{Classification error rate} = \frac{\# \text{ misclassifications}}{\# \text{ examples}}$$



What is the classification error rate on the training set?

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

Evaluating generalisation to unseen data



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

F O U N D A T I O N S
O F
D A T A
S C I E N C E

Principle of k-NN

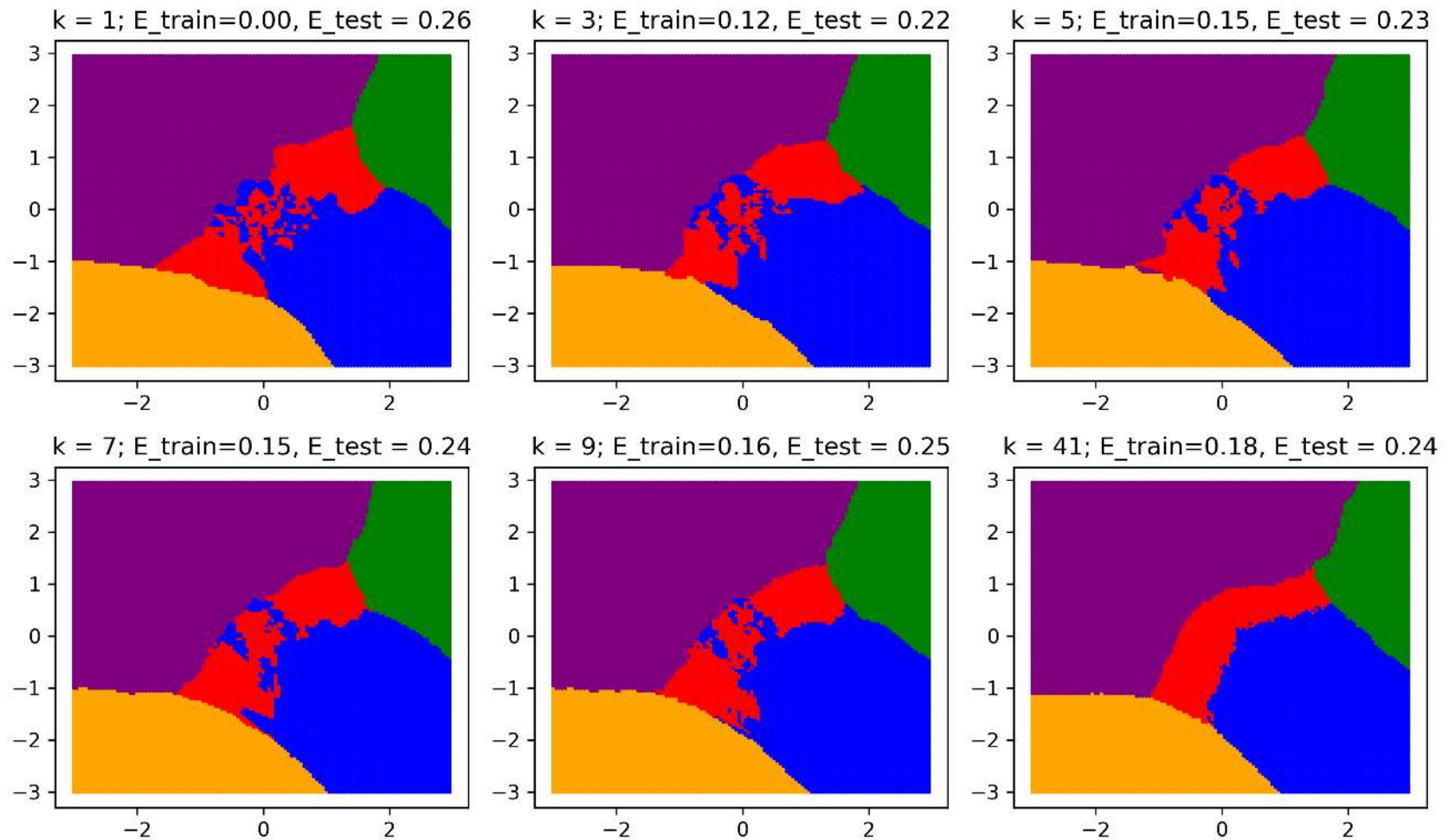
To improve generalisation look at...

k	R	B	Winner
1			
2			
3			
4			
5			

Ties: random resolution or
weighting

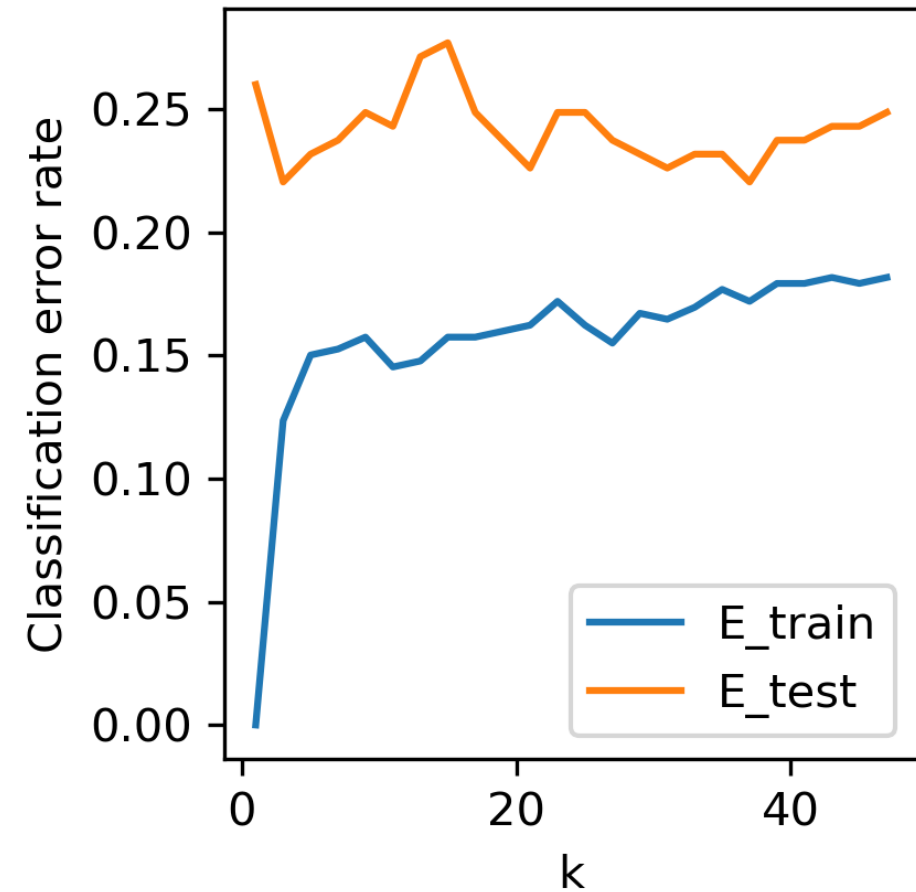
Odd numbers of k

Effect of k-NN

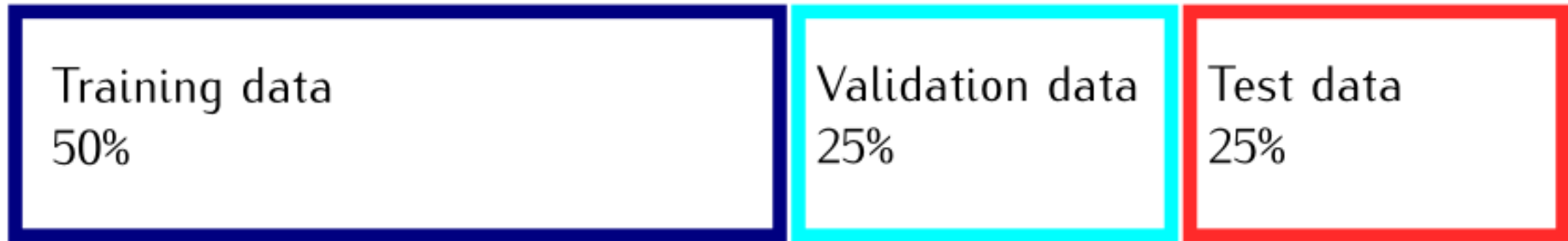


How to pick k?

k is a hyperparameter



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