Introduction to Machine Learning

Part 1: Supervised and Unsupervised Learning

John Pinney

November 2020

Intended learning outcomes

After attending the three sessions of this workshop, you will be better able to:

- Explain the difference between supervised and unsupervised learning.
- Select a suitable machine learning method for a given application.
- Prepare your own training and testing data sets.
- Evaluate the performance of a machine learning experiment.

Overview

What is machine learning?

Types of data

Unsupervised learning

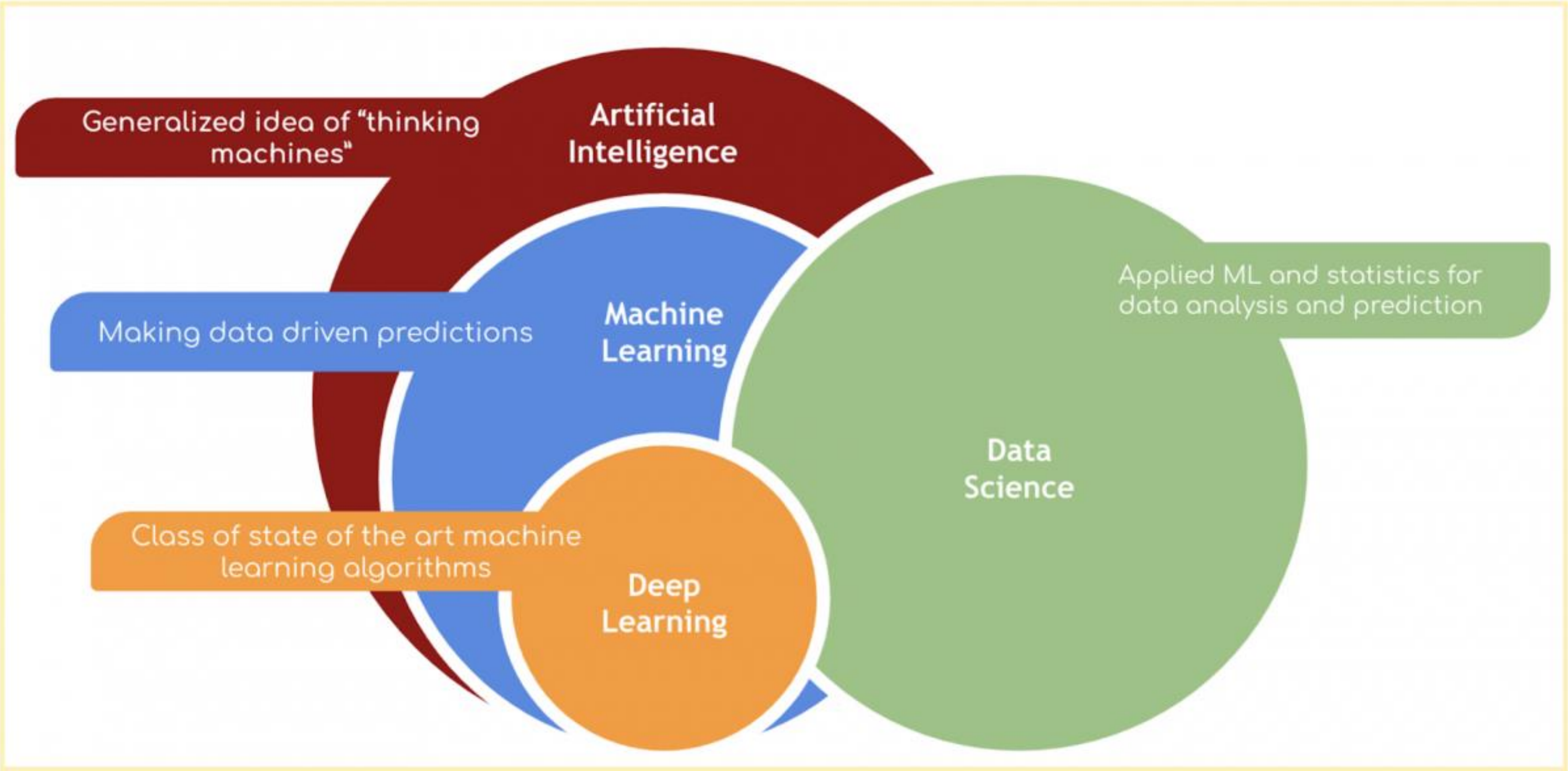
- Clustering
  - k-means

Supervised learning

- Regression
  - linear models
- Classification
  - logistic regression
  - decision trees

What is machine learning?

What is machine learning?



Statistical learning theory

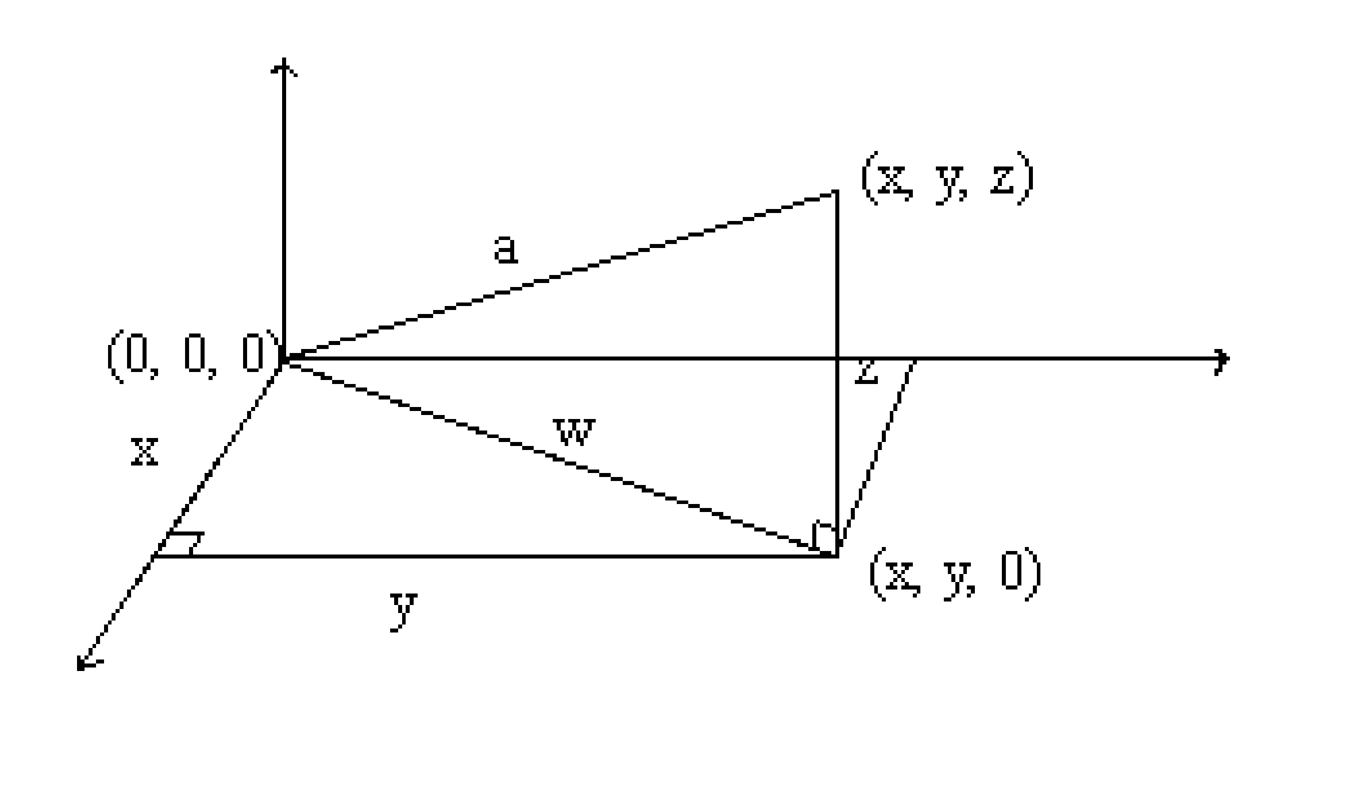
- Theory was introduced in the late 1960s.
- Became an applied science in 1990s.

- Allows us to
  - detect or learn structures and relationships in data.
  - assign observations to different classes.
  - make predictions based on current knowledge.

Some essential vocabulary…

- **vector**

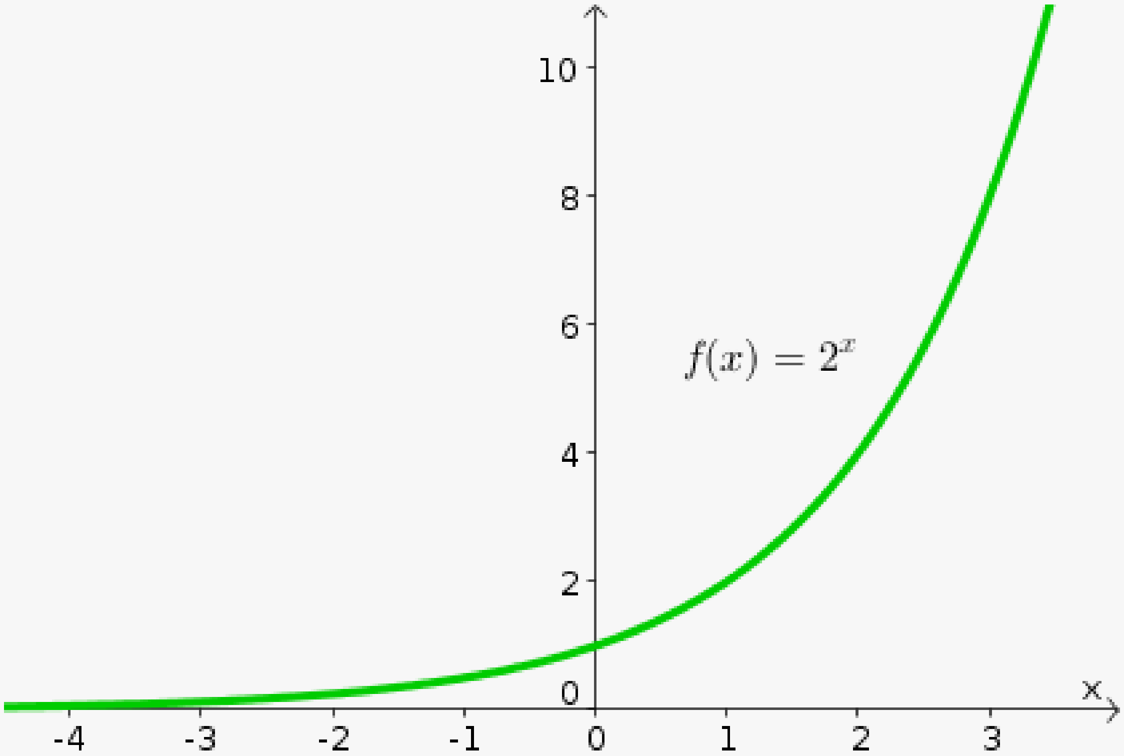
  A quantity within a multidimensional space.



Some essential vocabulary…

- **function**

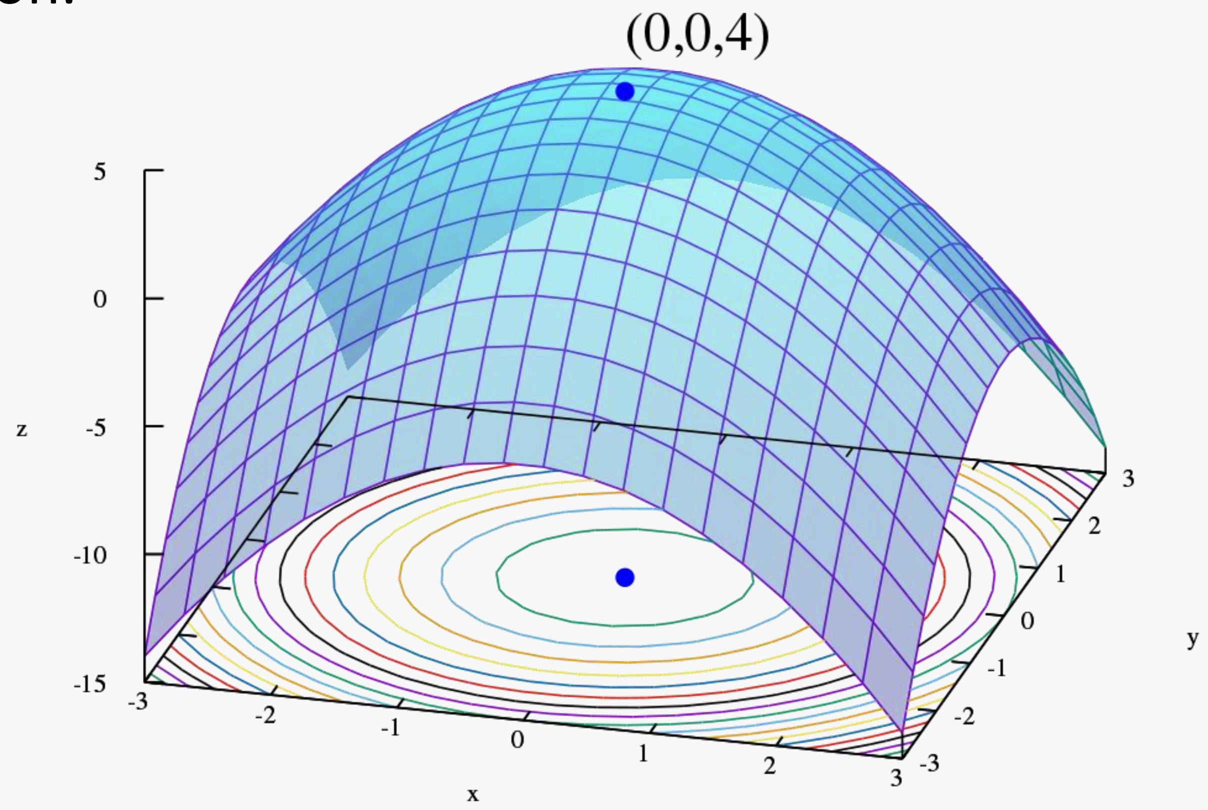
  A mapping from one vector space (input) to another (output).



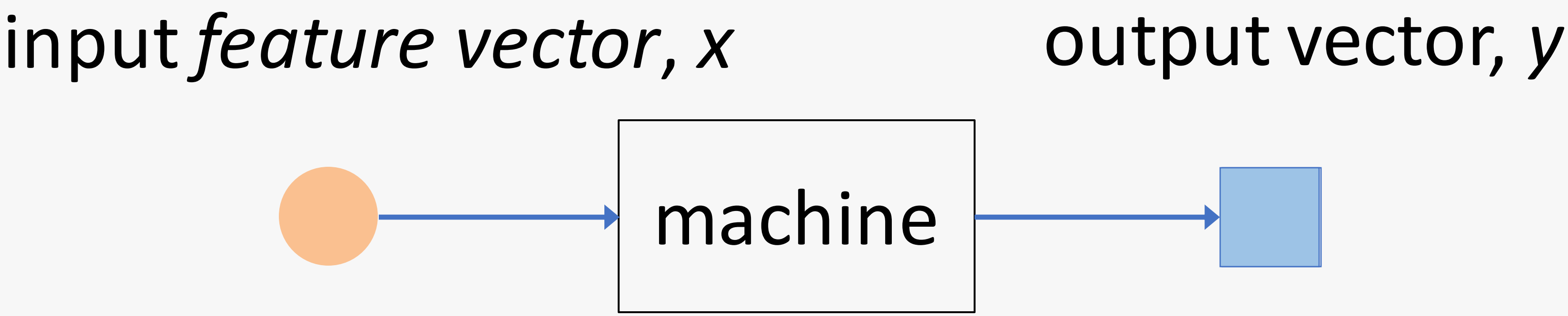
Some essential vocabulary…

- **optimisation**

  A procedure that attempts to find the minimum (or maximum) of a function.

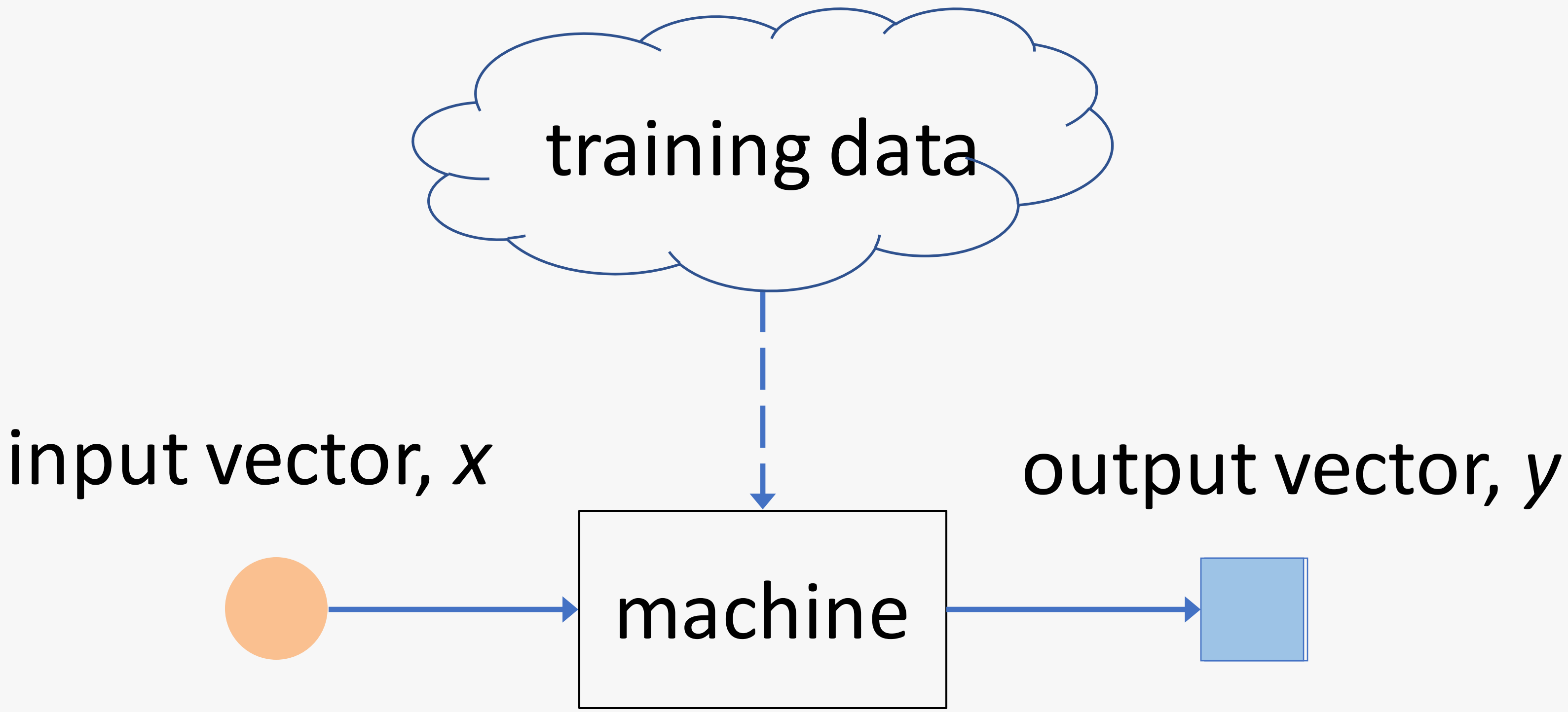


A ‘machine’ has inputs and outputs.



(acts like a function)

The machine has parameters that we might need to *fit* (optimise) using **training data.**



Types of data

Categorical data
(no numerical relationship between values)

- **Nominal data**: no obvious ordering of categories.

  e.g. favourite colour:

  green / blue / orange / yellow

  When there are only 2 possible categories, data is called *dichotomous* or *binary.*

- **Ordinal data**: there is a natural order for the categories.

  e.g. Likert scale:

  strongly disagree / disagree / neutral / agree / strongly agree

Quantitative data
(numerical data from counts or measurements)

- **Discrete data**: can only take specified values.
  e.g. number of children in a family (integer)

- **Continuous data**: can take any value in an interval.
  e.g. blood pressure

Example dataset

Take a look at the **iris** dataset.

What are the **features** and what are their data types?

Unsupervised learning

Unsupervised learning

- In unsupervised learning, we are looking for structure in the inputs without any knowledge of associated outputs: the data are considered to be *unlabelled.*

- We are seeking to “discover new knowledge”

- Examples include:
  - Dimensionality reduction, e.g. principal component analysis
  - Self-organising map
  - Clustering

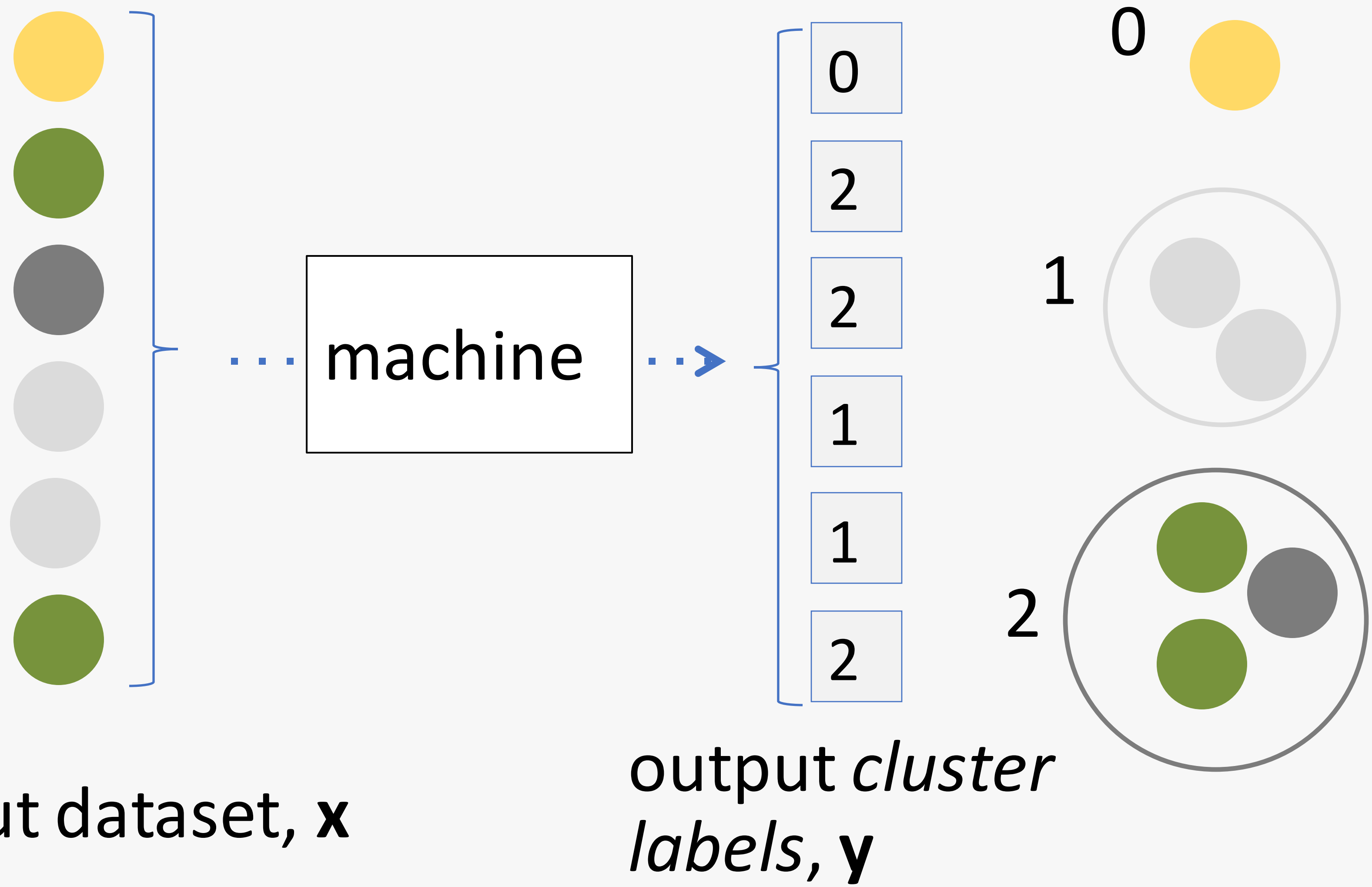
Clustering

To look for structure within a dataset, we often make use of **clustering** techniques.

A set of objects is grouped in such a way that objects in the same cluster are more similar (in some sense) to each other than to those in other clusters.

It is a central task in exploratory data mining.

Clustering



Clustering

- **Feature-based** clustering

  takes as input the set of input feature vectors.

- **Distance-based** clustering

  takes as input a matrix of **distances** that are calculated between each pair of input feature vectors.

  e.g. Euclidean distance.

Clustering methods may be **flat** (just reporting cluster labels) or **hierarchical** (reporting a *dendrogram* of nested clusters).

k-means clustering

A feature-based technique for *flat* clustering.

Requires a prior decision of the number of clusters (**k**) – in practice a good value for **k** for a given data set may be found by *post-hoc* analysis (e.g. silhouette score).

k-means clustering aims to partition **n** observations into **k** clusters, in which each observation belongs to the cluster with the nearest mean.

k-means clustering algorithm

1. Initialise positions for ***k*** cluster centroids (at random).
2. **Assignment step:** Assign each observation to the cluster whose centroid is "nearest" according to the chosen distance metric.
3. **Update step:** Calculate the new centroid positions according to the observations assigned to each cluster.
4. Check for convergence (cluster assignments did not change). If not converged, go to **2**.

k-means clustering

k-means is often fast in practice, but is a *heuristic method* so is not guaranteed to find the global optimum. Re-running several times with different starting points is therefore advisable.

Note that this is an example of an *expectation maximisation* approach.

k-means example

Using only the numerical features,

cluster the **iris** dataset.

k-means exercise

Look at the **abalone** dataset.

Considering only the numerical features, perform k-means clustering. Use the *silhouette score* to determine how many clusters the data appear to fall into.

What do the two clusters appear to correspond to?

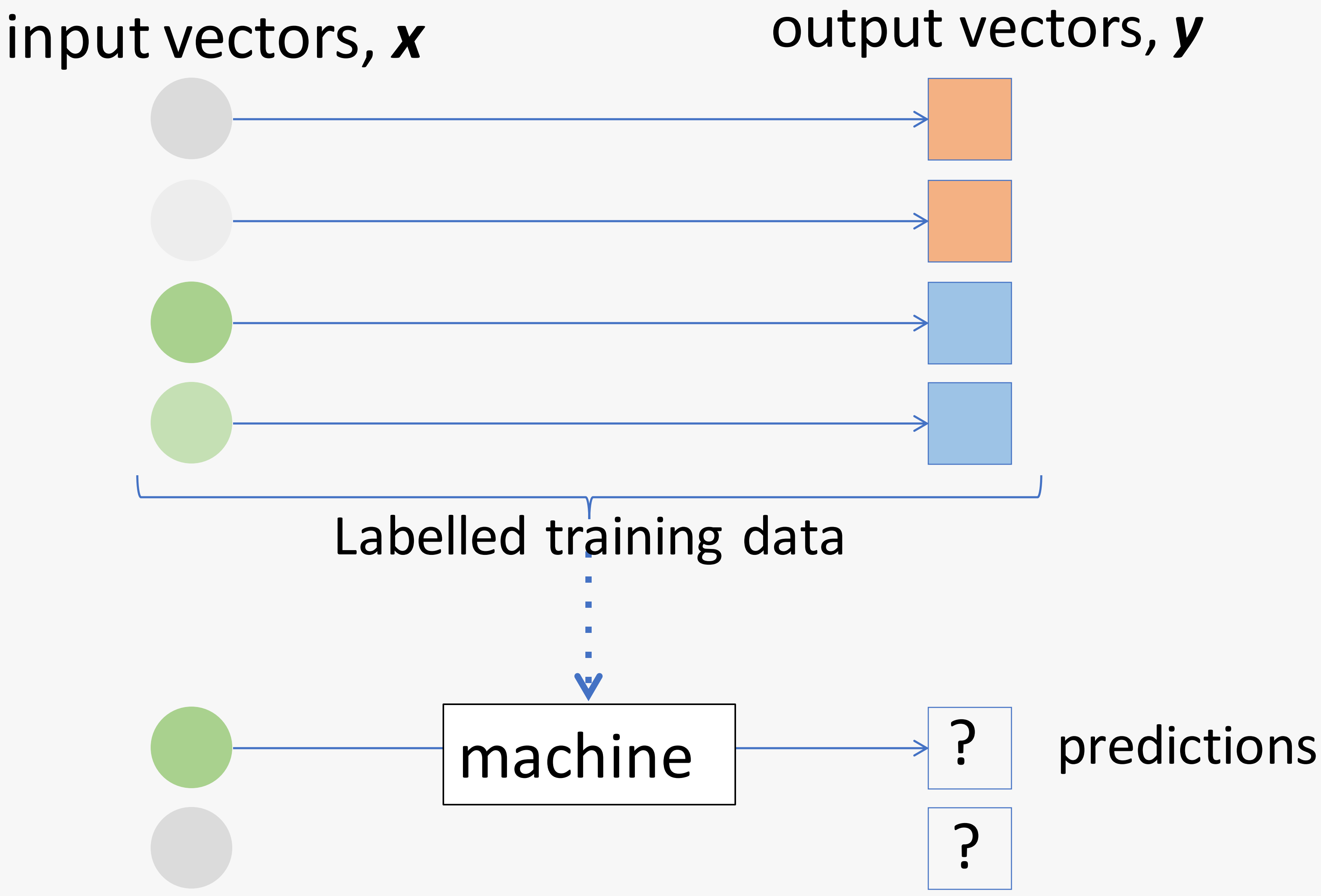
Supervised learning

Supervised learning

- Here, labelled data are used to “train” a machine learning algorithm, which is then used to classify or predict the response of new input data.

- We want to learn the function $f : x \rightarrow y$

Supervised learning



Two types of supervised learning

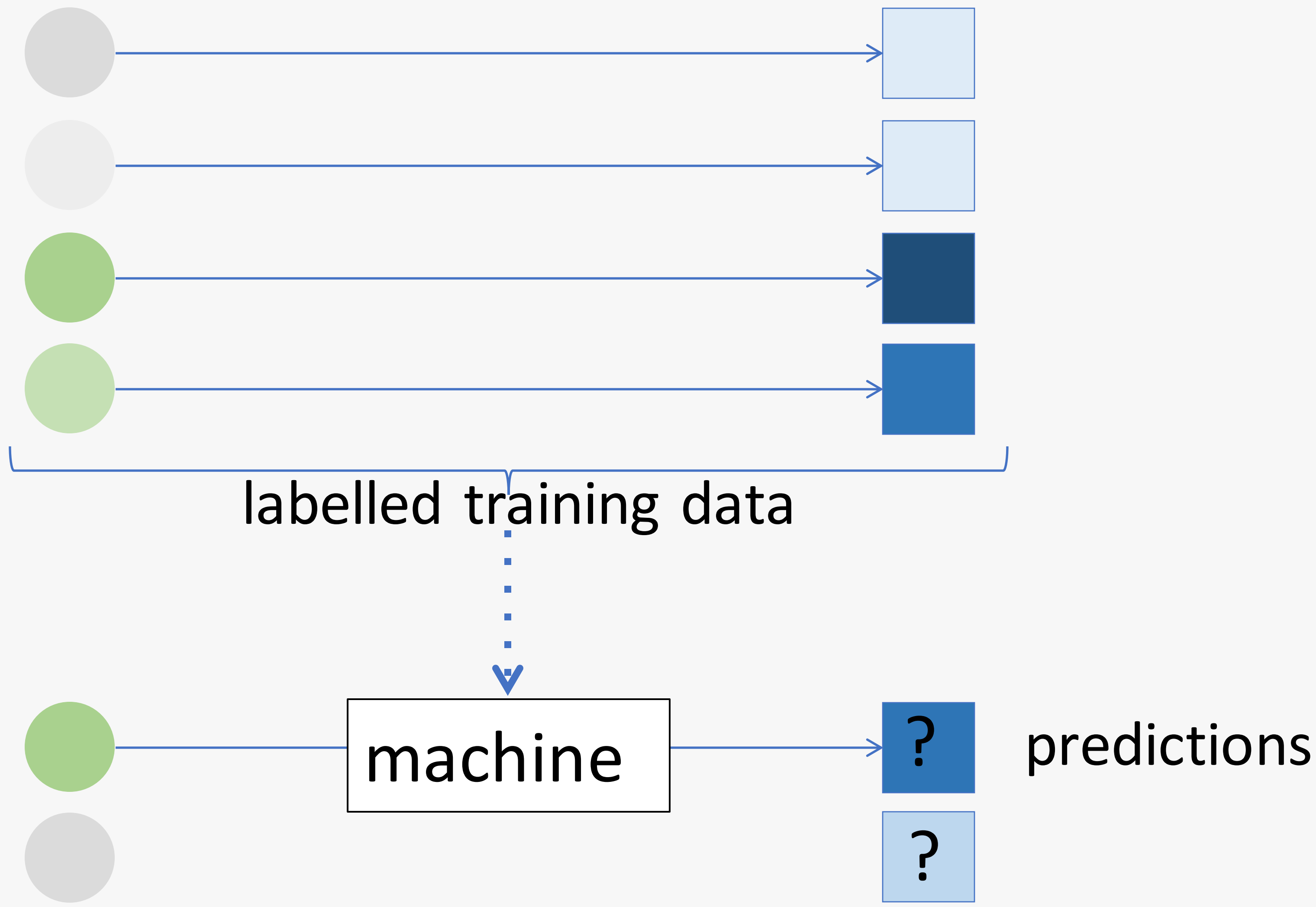
y is a continuous value
=> Regression
(estimate the response to a given input)

y is a discrete-valued class label
=> Classification
(identify the class of a given input)

Supervised learning:

Regression

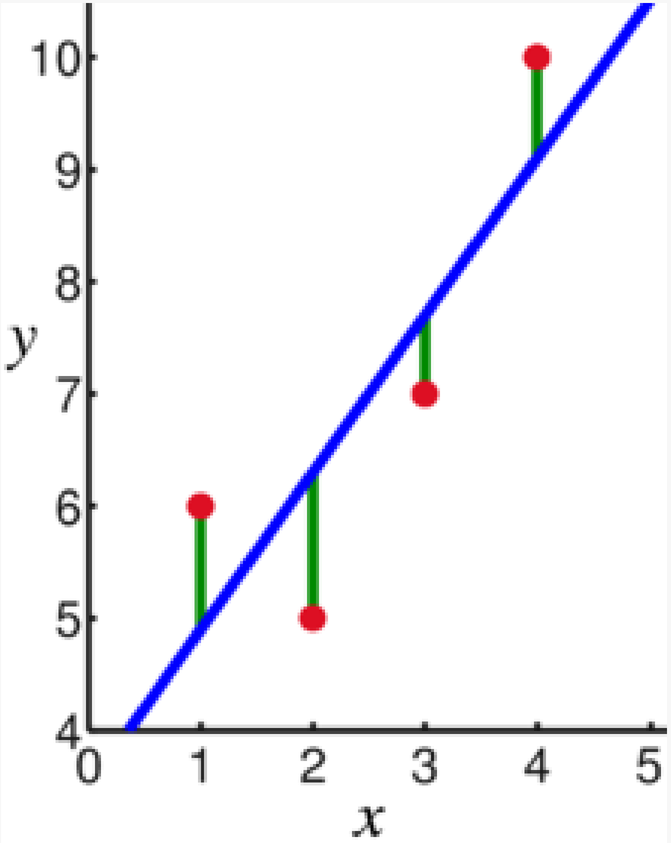
Regression



Linear regression

Predict ***y*** from the features of **x** by fitting a linear function.

Fitting is an optimisation procedure: e.g. minimise the sum of squared errors.



Linear regression example

With the **iris** dataset:

Considering only *iris virginica:*

1. Split the data into **training** and **testing** sets.
2. Use linear regression to predict **sepal length** from **petal length.**

Linear regression exercise

With the **abalone** dataset:

Considering only adults:

1. Split the data into **training** and **testing** sets.
2. Use linear regression to predict **rings** from the numerical features.

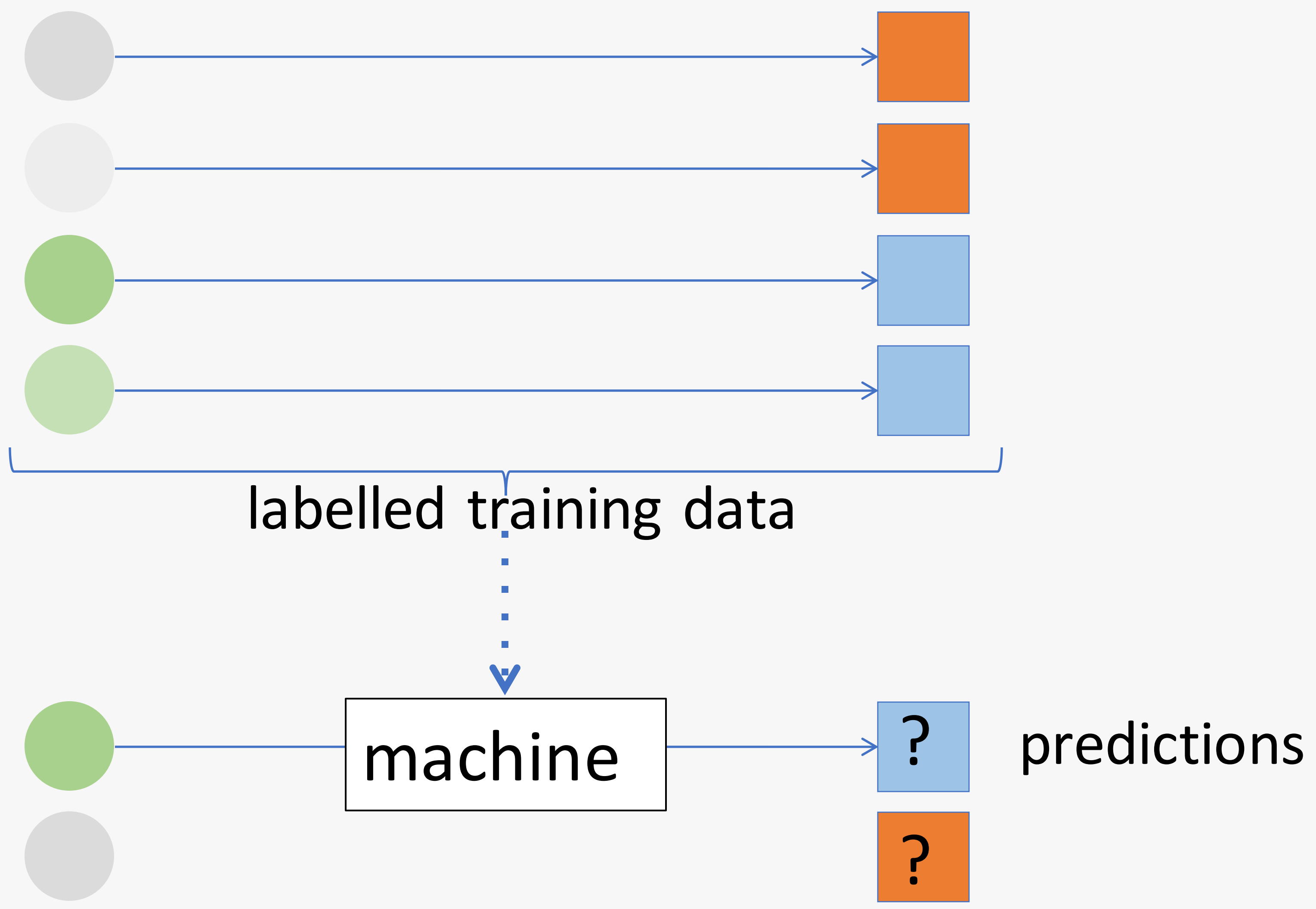
Linear regression with many features

We often want to apply some kind of **regularisation** to our model, so that small coefficients are pushed to zero. E.g. *ridge regression*, *lasso* or *elastic net*.

This makes models simpler and easier to interpret, and potentially shows which features are informative for predicting **y**.

Supervised learning:
Classification

Classification



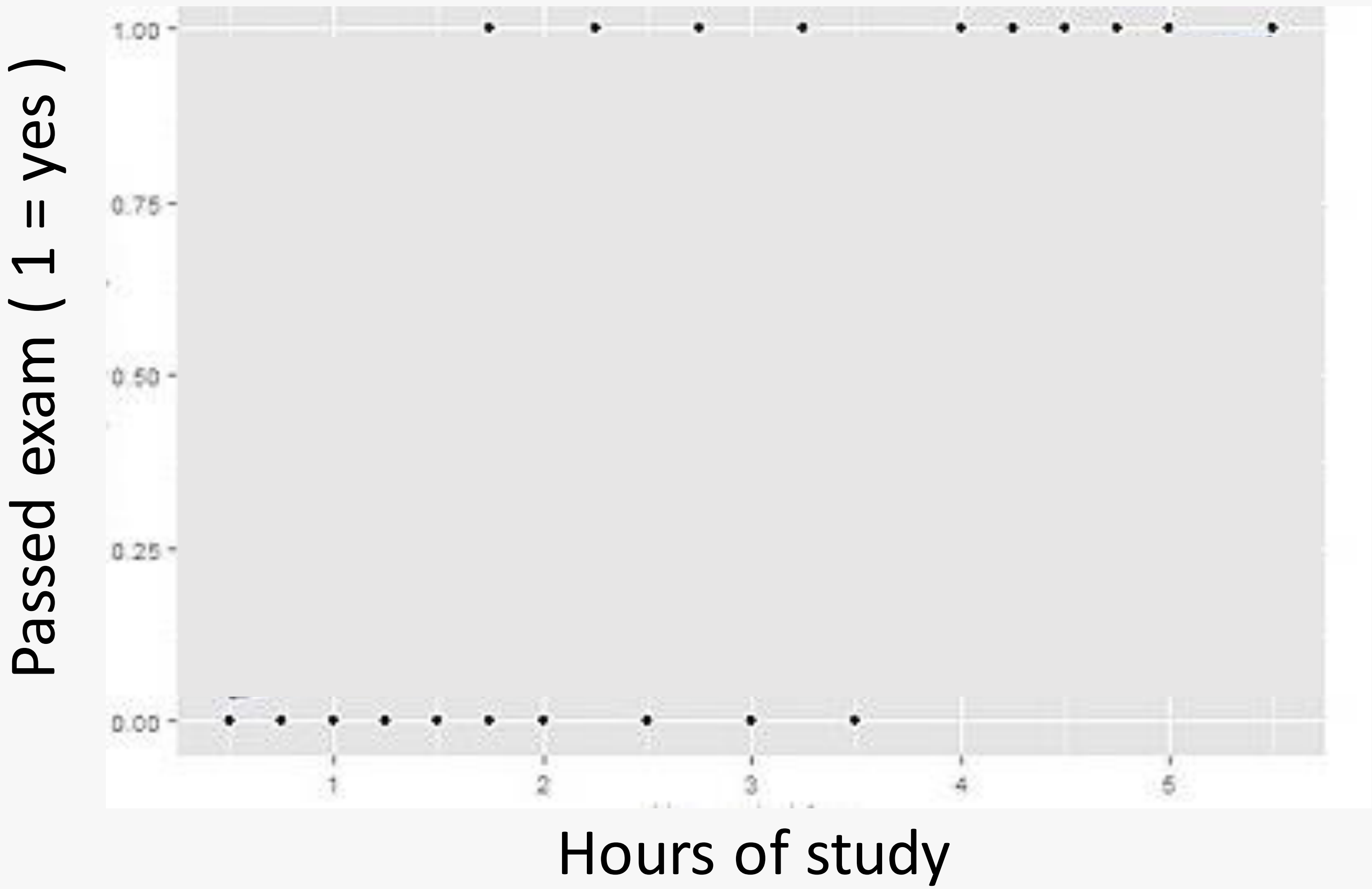
Logistic regression

Confusingly, logistic regression is an algorithm for **classification.**

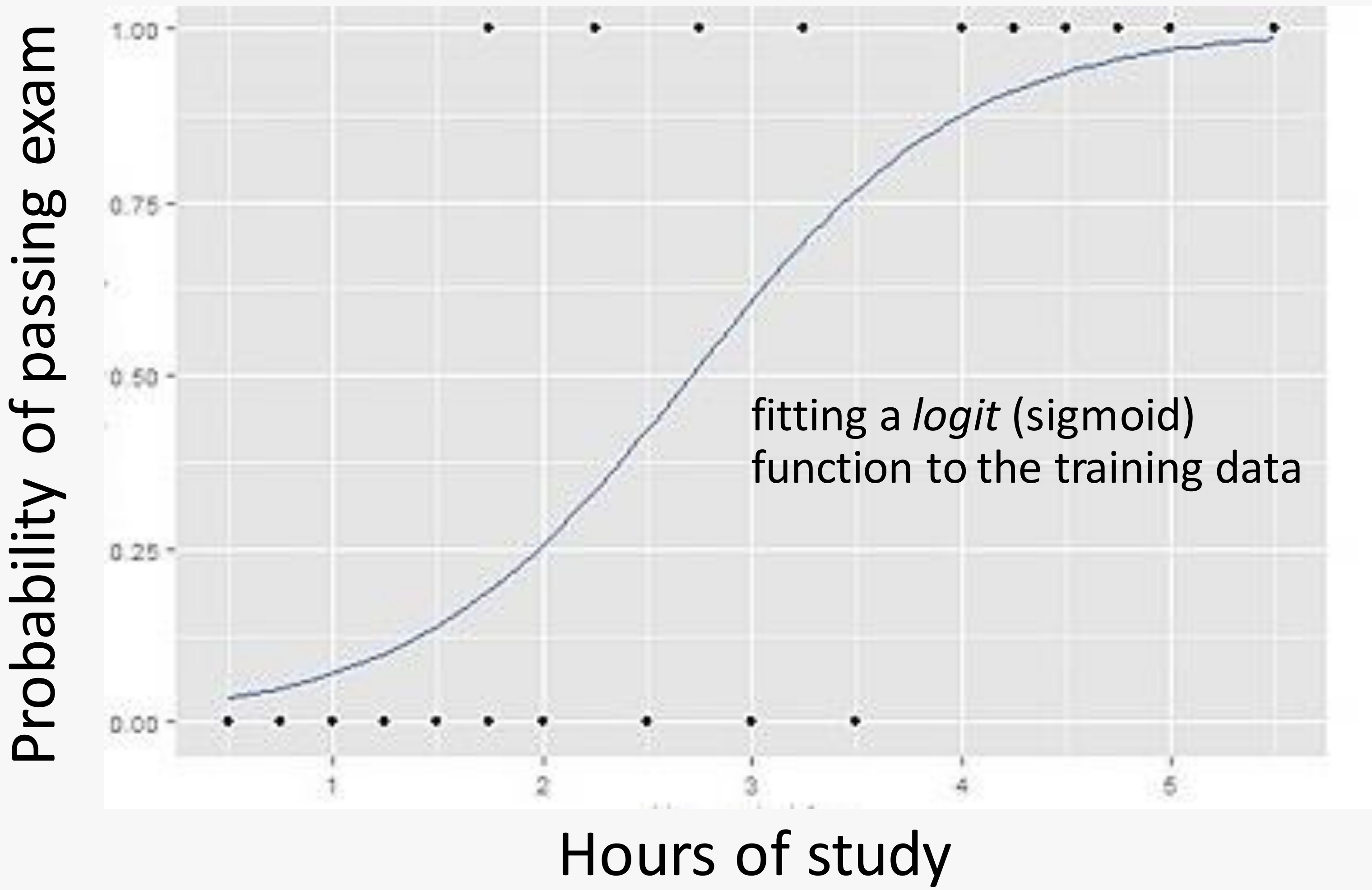
Consider a binary classification, with classes labelled 0 and 1.

For our training data, we can plot the **probability** that a particular value of **x** is labelled as class 1.

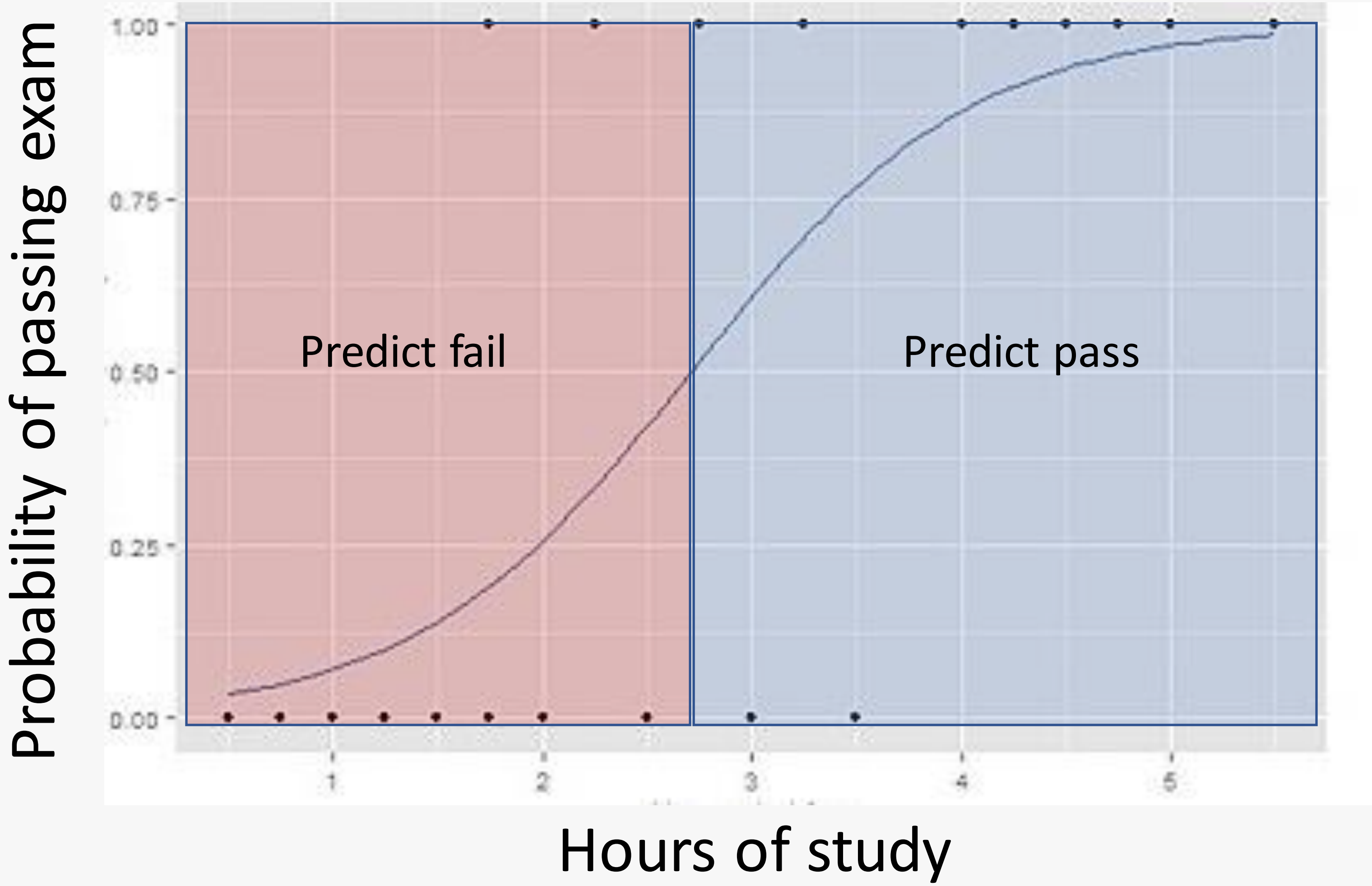
Logistic regression



Logistic regression



Logistic regression



Logistic regression example

With the **iris** dataset:

Considering *iris versicolor **and** virginica:*

1. Split the data into **training** and **testing** sets.
2. Predict **iris** (the species) from **petal length.**
3. Use a *confusion matrix* to examine the results.
4. Do the results improve if the other numerical features are included?

Do the same for a three-class logistic regression.

Logistic regression exercise

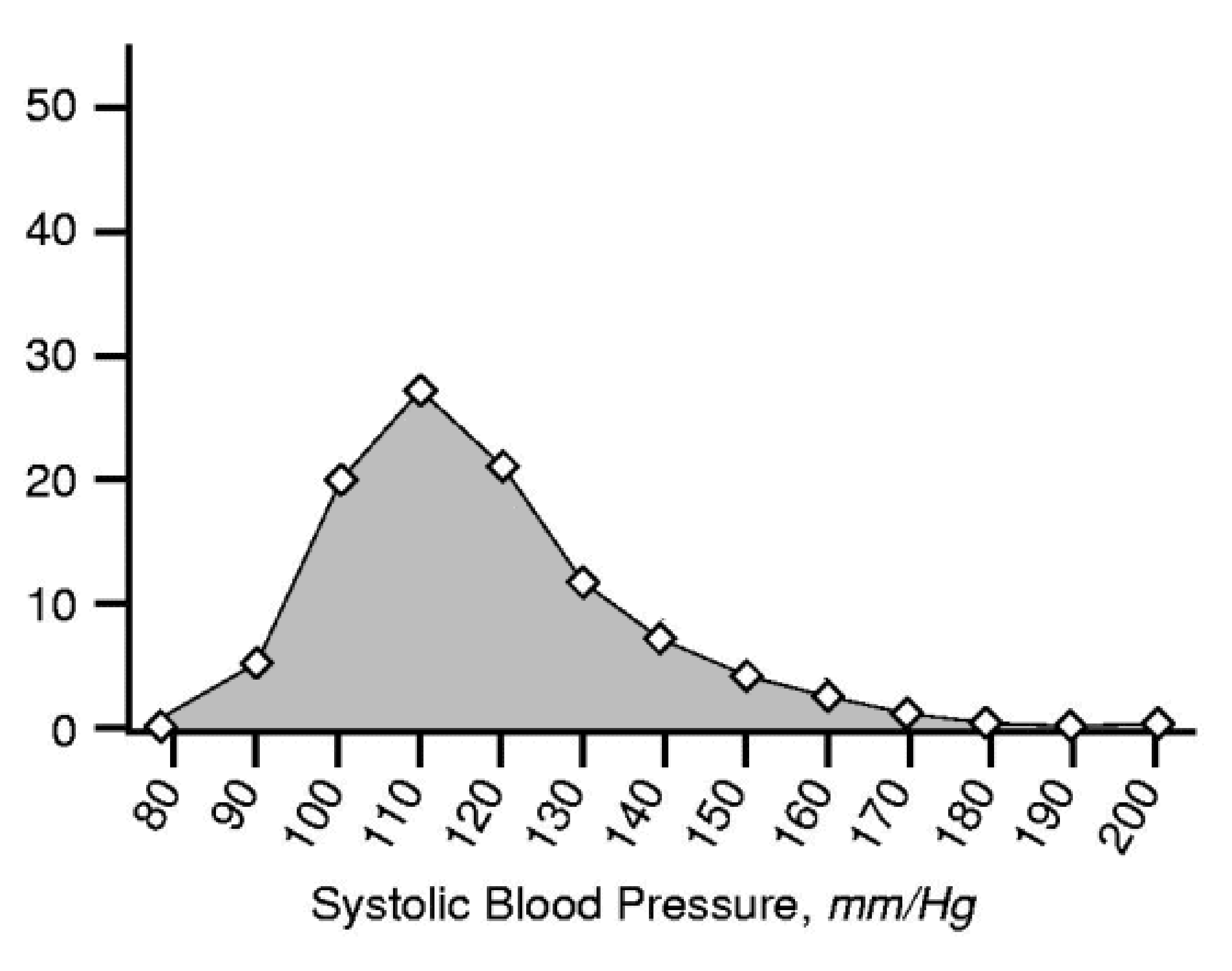
With the **kickstarter** dataset:

1. Split the data into **training** and **testing** sets.
2. Predict **funded** from the numerical features.
3. Use a contingency table to examine the results.
4. How could we make use of the **type** feature, which is a *nominal* data type?

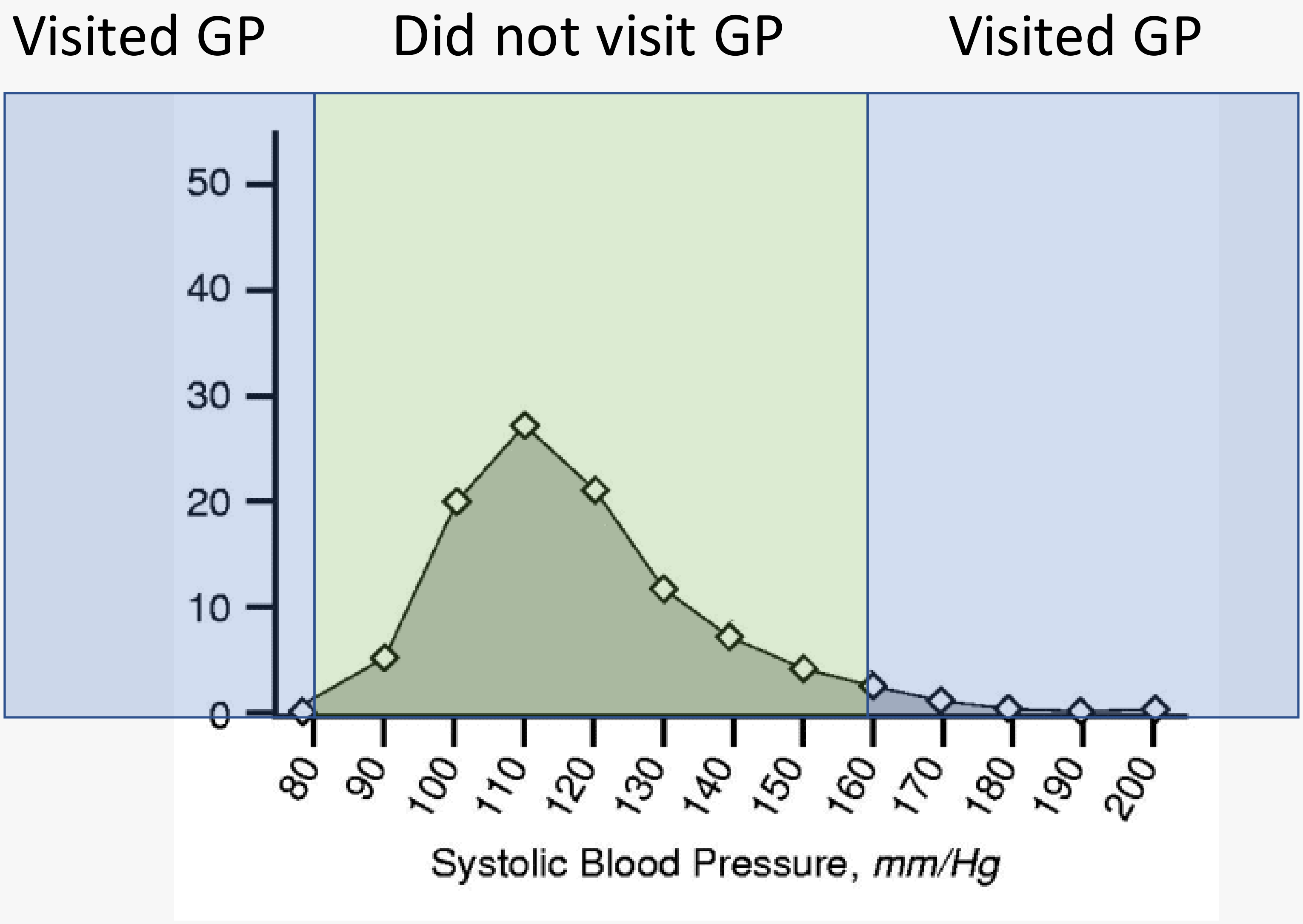
‘One-hot’ encoding

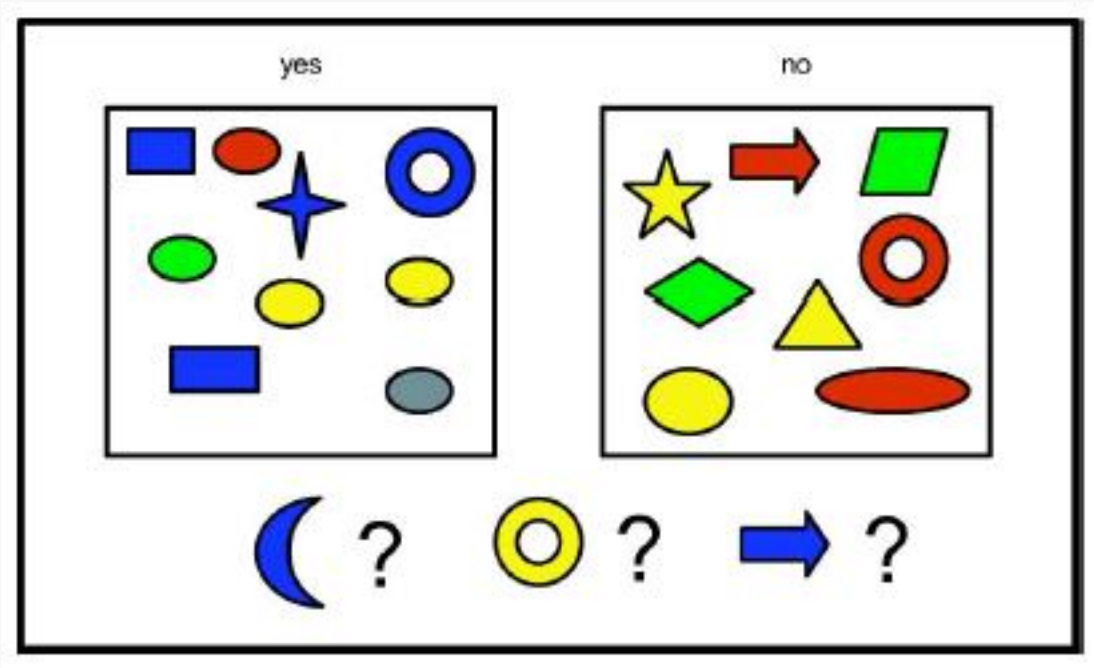
Useful for converting a categorical variable into multiple binary features, which can be used in algorithms that require numerical inputs.

What about non-linear classification?



What about non-linear classification?



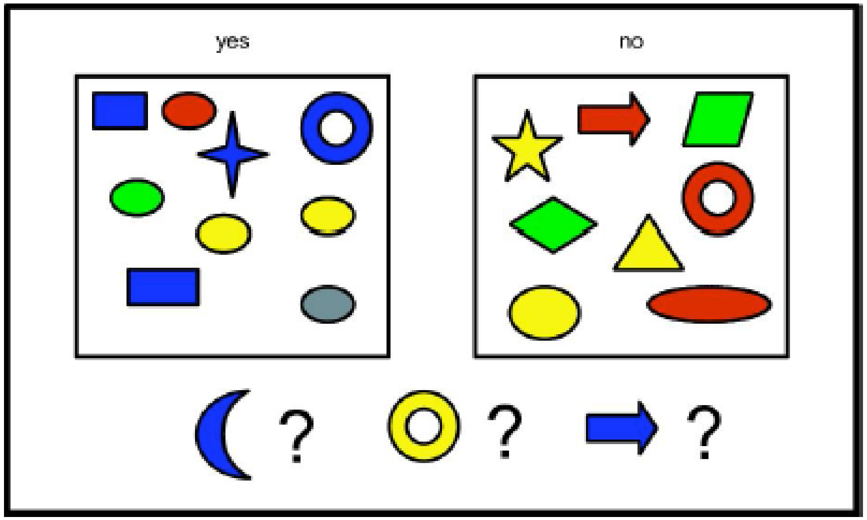


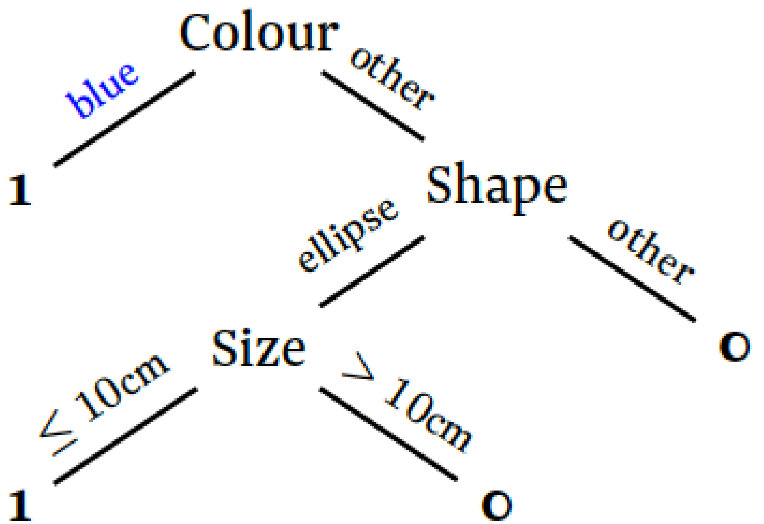
D features (attributes)

N cases

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

Decision tree





Finding an optimal tree is very difficult. In practice, we use a *greedy algorithm*, which builds the tree step by step, optimising the result at each stage.

Decision tree example

With the full **iris** dataset:

1. Split the data into **training** and **testing** sets.
2. Predict **iris** (the species) from the other features.
3. Use a *tree viewer* to examine the resulting decision tree.
4. Use a *confusion matrix* to examine the results.

Decision tree exercise

With the **titanic** dataset:

1. Split the data into **training** and **testing** sets.
2. Predict **survived** from the other features.
3. Use a *tree viewer* to examine the resulting decision tree.
4. Use a *confusion matrix* to examine the results.

Summary of Part 1

Machine learning is a subfield of artificial intelligence, concerning *data-driven predictions*.

Clustering (e.g. *k-means*) is an *unsupervised* approach. It can be used to discover structure in unlabelled data.

Regression (e.g. *linear regression*) is a *supervised* approach. It predicts a numerical output from the input features.

Classification (e.g. *logistic regression, decision tree*) is also a *supervised* approach. It predicts a categorical output from the input features.

Next time…

How can we **evaluate and compare performance** in supervised learning?

How can we **improve performance** beyond the basic algorithms?