SMM636 Machine Learning

Rui Zhu

An introduction to machine learning

1. An introduction to machine learning
   - Machine learning applications
   - Supervised and unsupervised learning
   - Examples: classification
   - Some trade-offs
   - Machine learning versus statistical learning

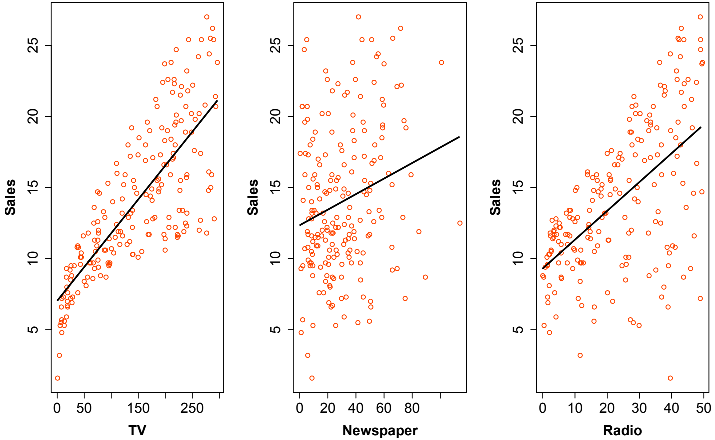
Machine learning applications

- Understand the association between wage and demographic variables in population survey data.[1]
- Understand the association between sales of a product and advertising in different media
- Predict whether an individual will default on his or her credit card payment.
- Predict whether a given day's stock market performance will increase or decrease using the past 5 days' percentage changes in the index.
- Identify handwritten digits.
- Classify a flower to one of three species of iris.
- Understand which colours are similar by human vision.
- Study whether there are groups, or clusters, among the cancer cell lines based on their gene expression measurements.

[1] Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

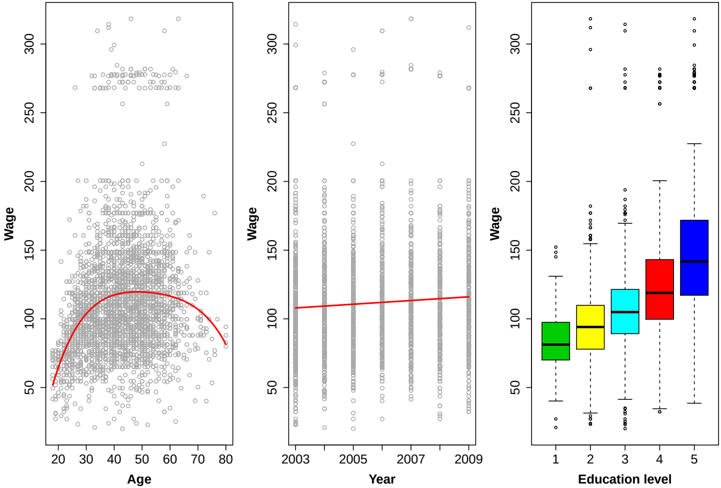
Advertising data

Consists of the sales of a product in 200 different markets, with advertising budgets for three media.



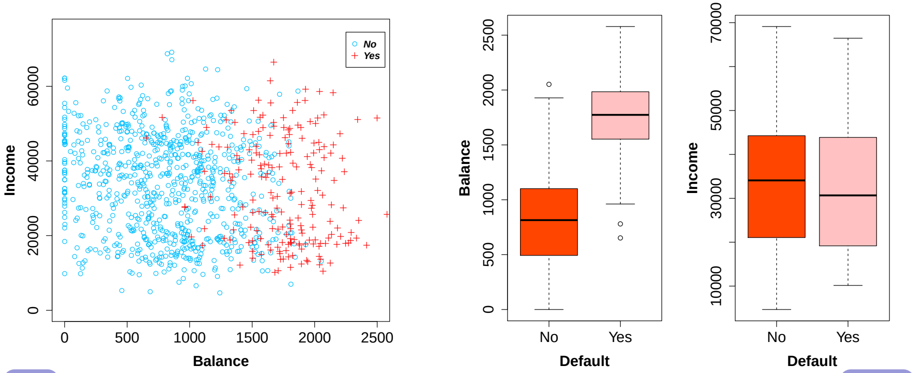
Wage data

Income survey information for 3000 males from the central Atlantic region of the United States.



Default data

A simulated dataset that contains information of ten thousand customers.

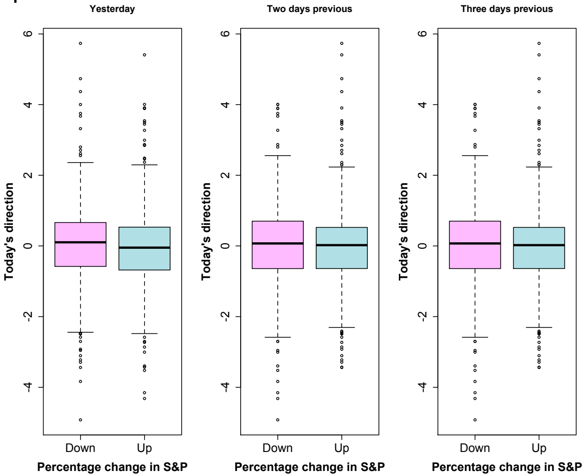


back

forward

Stock market data

Contain the daily movements in the Standard & Poor's 500 stock index over a 5-year period.

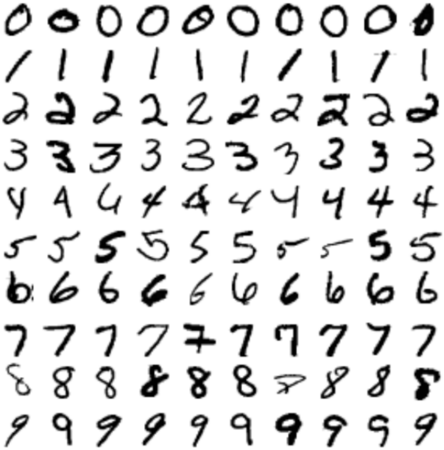


back

forward

MNIST Handwritten digits data

Contain 70,000 images of handwritten digits.

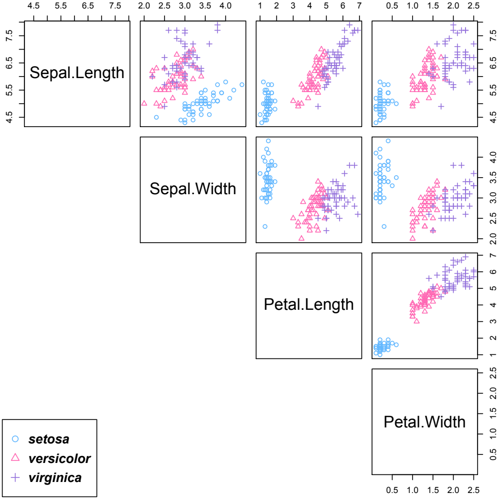


back

forward

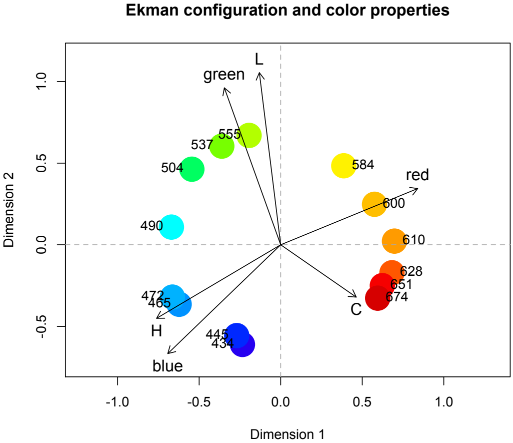
Iris data

Fisher's iris dataset contains four measurements for 50 flowers from each of 3 species of iris.



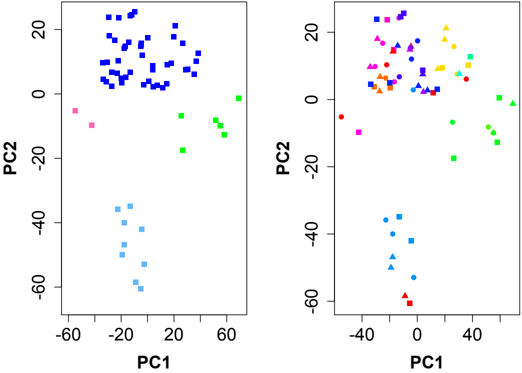
Ekman data

Contain similarities for 14 colours which are based on a rating by 31 subjects.



NCI60 gene expression data

Consist of 64 samples of human tumour with 6830 gene expression measurements.



back

forward

A wide range of applications in computer vision

- Facial recognition
- Person re-identification
- Image restoration
- Object recognition
- . . .

Any similarities or dissimilarities between these tasks?

Supervised learning

1. **Regression**: the outcome measurement is *quantitative*.
2. **Classification**: the outcome measurement is *qualitative*, i.e. categorical, discrete variables or factors.

We will focus on classification problems in supervised learning.

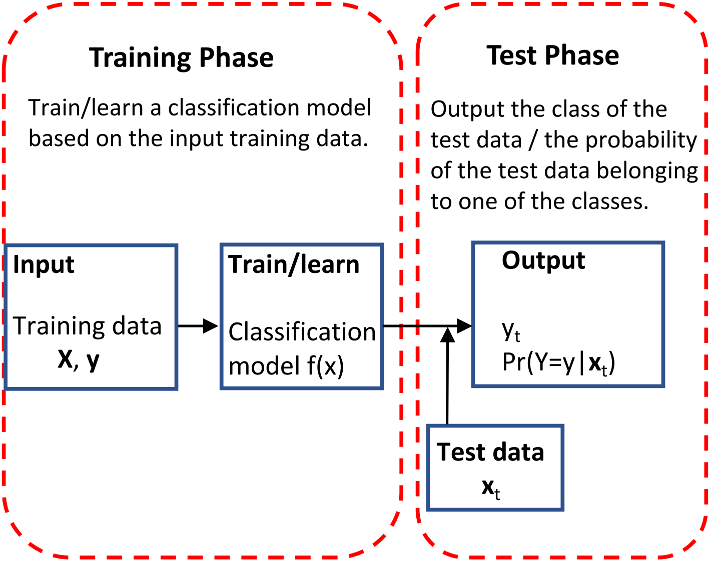
Supervised learning: some notations

1. Measurement variables:
   - Outcome measurement Y (also called dependent variable, response in regression or target, class, label in classification).
   - Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
2. Observed values:
   - $\mathbf{x}_i$: a p-dimensional column vector, the ith observation (also called instance, example or sample) of X.
   - $\mathbf{X}$: an $N \times p$ matrix, a set of N observations of X.
   - y_i: a scalar (or factor), the ith observation of Y.
   - $\mathbf{y}$: an N-dimensional column vector, a set of N observations of Y.
3. Training data: a set of N observations of both X and Y
   - $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$
   - $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$
   - $\mathbf{X}$, $\mathbf{y}$

Supervised learning: objectives

1. Training and test data
   - Training data: $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$
   - Test data: $\mathbf{x}_t$ for one test instance, or $\mathbf{X}_t \in \mathbb{R}^{N_t \times p}$ if we have N_t test instances.
2. Given the training data, we aim to
   - Understand the association between outcomes and inputs.
   - Predict the response/class, $\hat{y}_t$, of the test data $\mathbf{x}_t$.
   - Assess the quality of the predictions and inferences.

Classification



Training Phase

Train/learn a classification model based on the input training data.

Input

Training data

X, y

Train/learn

Classification model f(x)

Test Phase

Output the class of the test data / the probability of the test data belonging to one of the classes.

Output

y_t

$\Pr(Y=y|\mathbf{x}_t)$

Test data

$\mathbf{x}_t$

Classification

Unsupervised learning

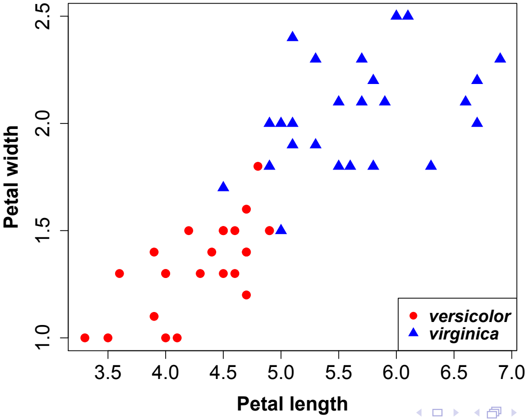
1. No outcome measurement is available: we only have $\mathbf{X}$.
2. Objective: discover interesting patterns in predictor measurements.
    - Visualise the data in a low-dimensional space, i.e. unsupervised dimension reduction.
    - Detect subgroups in the data, i.e. clustering or cluster analysis.
    - ...
3. More challenging than supervised learning.
4. Hard to assess the results from unsupervised learning.

What we will learn in SM636

1. We will focus on popular supervised and unsupervised learning methods.
2. Other learning tasks:
   - Semi-supervised learning: sometimes labelling is expensive and we have a small number of labelled training data while a large number of unlabelled data.
   - Multi-label classification: each instance can belong to several classes at the same time.
   - Reinforcement learning: take an action in a give situation to maximise a reward.
   - ...

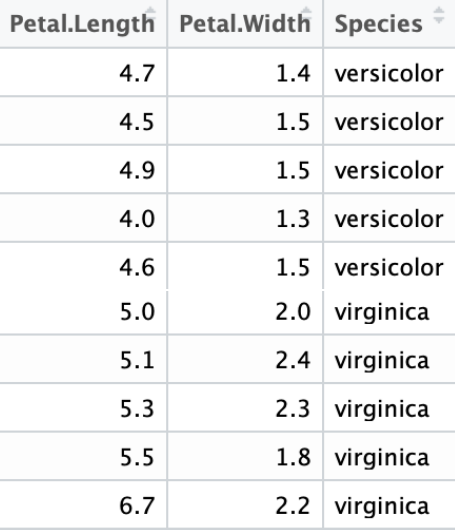
Examples: classification (iris data)

- Task: we aim to classify a new flower to one of the two classes, versicolor and virginica.
- Training data: 1) 25 versicolor flowers and 25 virginica flowers, 2) each of them is described by two features, petal length and petal width.



Examples: classification (iris data)

- Some examples of the training data.



Petal.Length	Petal.Width	Species
4.7	1.4	versicolor
4.5	1.5	versicolor
4.9	1.5	versicolor
4.0	1.3	versicolor
4.6	1.5	versicolor
5.0	2.0	virginica
5.1	2.4	virginica
5.3	2.3	virginica
5.5	1.8	virginica
6.7	2.2	virginica

Examples: classification (iris data)

1. Measurement variables:
   - Outcome measurement (class labels) $Y \in \{\text{versicolor, virginica}\}$.
   - Predictor measurements (features) $X=$ (petal length, petal width).
2. Observed values:
   - $\mathbf{x}_i$: a 2-dimensional column vector, $i = 1, 2, \ldots, 50$.

$$\text{The first flower}(i = 1) : \mathbf{x}_1 = \begin{bmatrix} 4.7 \\ 1.4 \end{bmatrix}_{2\times 1}$$

   - $\mathbf{X}$: a 50×2 matrix

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \cdots \\ -\mathbf{x}_{50}^T- \end{bmatrix}_{50\times 2}$$

Examples: classification (iris data)

- y_i: versicolor or virginica $(i = 1, 2, \dots, 50)$

$$y_1 = \text{versicolor}$$

- $\mathbf{y}$: a 50×1 column vector

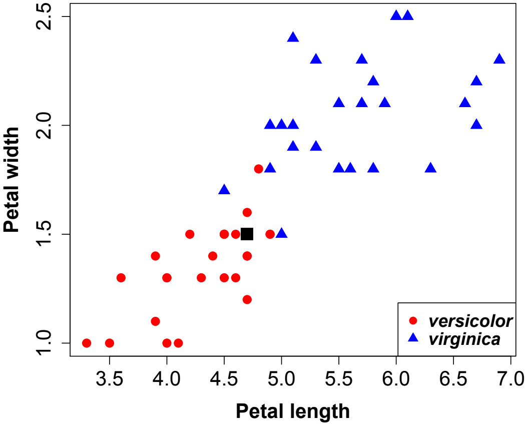
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_{50} \end{bmatrix}_{50\times1} = \begin{bmatrix} \text{versicolor} \\ \text{versicolor} \\ \dots \\ \text{virginica} \end{bmatrix}$$

3. Training data: $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_{50}, y_{50})\}$

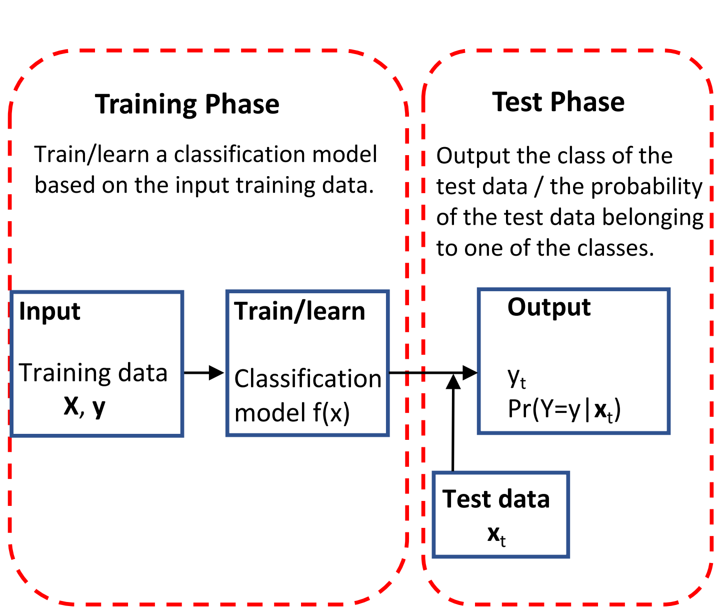
$$\mathbf{X} = \begin{bmatrix} 4.7 & 1.4 \\ 4.5 & 1.5 \\ \dots & \dots \\ 5.7 & 2.1 \end{bmatrix}_{50\times2}, \quad \mathbf{y} = \begin{bmatrix} \text{versicolor} \\ \text{versicolor} \\ \dots \\ \text{virginica} \end{bmatrix}_{50\times1}$$

Examples: classification (iris data)

- How to classify a test flower (black square), $\mathbf{x}_t = (4.7, 1.5)^T$, to one of the two classes?

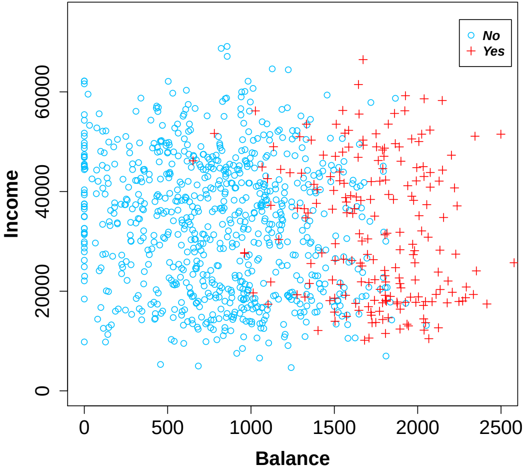


Examples: classification (iris data)


Training Phase
Train/learn a classification model based on the input training data.
Input
Training data
X, **y**
Train/learn
Classification model f(x)
Test Phase
Output the class of the test data / the probability of the test data belonging to one of the classes.
Output
y_t
$\Pr(Y=y|\mathbf{x}_t)$
Test data
$\mathbf{x}_t$

Examples: classification (default data)

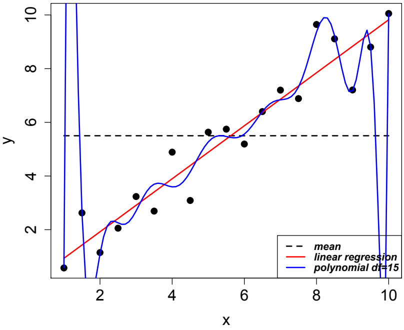
- How to estimate the probabilities that an insurance claim is fraudulent, based on the training data?

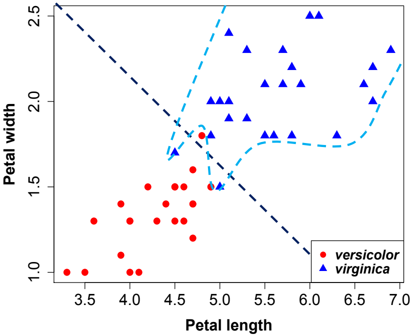


Goodfitting versus overfitting and underfitting

- Generalisation ability: we prefer a model that can well describe the pattern in the training data, as well as well generalise to unseen/new/test data.
- Fitting the data using a very simple model, we probably can't well capture the underlying relationship between features and responses (underfitting).
- As the flexibility of a model increases, we can well capture the pattern in the training data, however we can't well predict an unseen data, because the very flexible model also models the random noise in the training data (overfitting).

Goodfitting versus overfitting and underfitting





Prediction accuracy versus interpretability

- Linear models are easy to interpret, but may not make accurate predictions.
- Sophisticated models may provide accurate predictions, but are difficult to interpret.
- Depends on your aim: more accurate or more interpretable.

Machine learning versus statistical learning

From Prof Hastie and Prof Tibshirani [2]

- Machine learning is a subfield of artificial intelligence while statistical learning is a subfield of statistics.
- There is much overlap. The distinction has become more and more blurred.
- Machine learning has the upper hand in marketing.

[2] https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf