MATH5743M: Statistical Learning - Lecture 1

Dr Seppo Virtanen, s.virtanen@leeds.ac.uk

Semester 2: 24 Jan 2022

Course arrangements and assessment

Each week will be focused on a particular topic on statistical learning, with two weekly lectures followed by a practical session. The lectures will be kept weekly on Mondays at 9am; the first lecture is on 24 Jan, week 1. The practicals will be arranged on Mondays at 11am on even semester weeks; the first practicals is on Monday 31 Jan, week 2.

Assessment will be by a mixture of an exam (50%) and three pieces of assessed practical work (50%). The deadlines for the three assessed practicals are:

1. 11:59pm Friday March 18
2. 11:59pm Friday April 8
3. 11:59pm Friday May 13

You need to return each **assessed** practical report by the deadline in Minerva using Turnitin. I will return to you marks and feedback. Details about the exam will be provided later.

Course outline

Statistical learning is the practice of using statistical methods to recognise patterns in data, to identify the processes that produced the data and to make predictions using that knowledge.

Course outline

Statistical learning is the practice of using statistical methods to recognise patterns in data, to identify the processes that produced the data and to make predictions using that knowledge.

Examples related to *prediction*:

- Stockmarket traders use statistical models to *predict* the value of stocks in the future based on previous stock values, company earnings and national economic trends.
- Companies like Facebook, Google and Amazon use statistical models to *predict* which adverts or products you are most likely to find interesting.
- Police try to *predict* where, when and what type of crime will occur.

Course outline

Statistical learning is the practice of using statistical methods to recognise patterns in data, to identify the processes that produced the data and to make predictions using that knowledge.

Examples related to *inference*:

- Security services such as GCHQ use statistical models to *infer* possible threats from voice recordings, email records and video tracking.
- Criminologists aim to *infer* causes of crime based on socio-demographic and descriptive spatial features.

Course outline

Statistical learning is the practice of using statistical methods to recognise patterns in data, to identify the processes that produced the data and to make predictions using that knowledge.

Examples related to both *prediction* and *inference*:

- Biologists and social scientists use statistical models to *infer* what drives human and animal behaviour, and to *predict* what individuals will do when their situation changes.
- Doctors use statistical models to *infer* which patients are likely to respond to a specific drug, and to *predict* which individuals are likely to get sick based on previous medical records.

Course outline

Statistical learning is the practice of using statistical methods to recognise patterns in data, to identify the processes that produced the data and to make predictions using that knowledge.

Examples related to *decision making*:

- Self-driving cars use statistical models to *decide* on the best action to take, using cameras, GPS and data from human drivers.
- Computer programs playing GO or chess against human opponents.

Statistical Learning

- high demand for *data scientists* in academia and industry
- intersection of computer science, maths and stats and domain/application knowledge
- key developments based on statistical modelling

This course will introduce you to the foundations of statistical learning. You will learn how to

- select the right statistical model to analyse a data set,
- how to use statistical models to understand and predict data and
- see how real world examples make use of statistical models to solve both academic and industrial problems.

Lecture topics

A variety of statistical modelling techniques: how they work, when they are used, and how complex statistical models can be constructed from a few fundamental ingredients.

1. Introduction
2. Optimisation and learning
3. Linear regression 1
4. Linear regression 2
5. Model selection
6. Logistic regression
7. Decision trees
8. Ensemble learning and collective intelligence
9. Random Forests
10. Further topics/Revision

Both theoretical and applied aspects

Apply these techniques and understanding practically and efficiently using the R statistical software package. Learn all the details of the analysis are performed.

- How a method is developed,
- how it works
- and important factors to consider when using it

Using R

In R,

```
help(runif)
```

Also, try Googling **R function for X**, or **how to X in R**.

When you get errors

- Pause and think.
- Check the help files and try the examples.
- Google the error message
- Ask me! In practicals I can help you with any problems quickly.

Supervised learning

Predict the values of some *outputs* based on matching sets of *inputs*. (In classical stats refer to dependent/independent variables.)

$$P(\text{OUTPUT} = y \mid \text{INPUT} = x, \theta) = f(y, x, \theta) \tag{1}$$

- *Choose* an appropriate functional form for $f()$, alternatively, a probability density function for the outputs conditioning on the inputs and parameters
- Using a data set of matched inputs and outputs, *learn* the best values for the free parameters, θ
- *Interpret* the values of θ - what do they tell us about the real system we are studying?
- Use $f(\cdot, \cdot, \theta)$ to *predict* the values of new outputs, given their matching inputs.
- Model evaluation and selection. Which model performs best?

Supervised learning

Based on the output domain/type, often refer to *regression* or *classification*. Useful probability distributions include

- normal for continuous and numeric outputs,
- Bernoulli for binary yes/no outputs and
- multinomial/categorical for one-out-of-K classes.

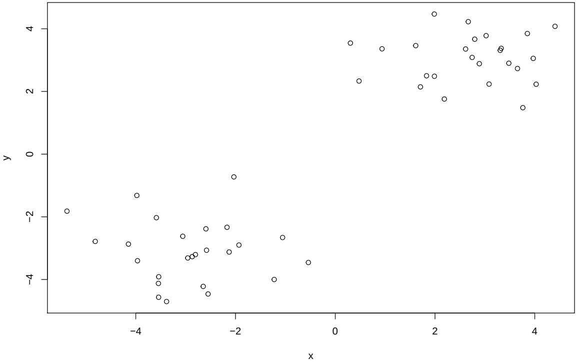
Revisit the prediction examples

- Stockmarket traders use statistical models to *predict* the value of stocks in the future based on previous stock values, company earnings and national economic trends.
- Companies like Facebook, Google and Amazon use statistical models to *predict* which adverts or products you are most likely to find interesting.
- Police try to *predict* where, when and what type of crime will occur.
- Face recognition.
- Weather prediction.

Unsupervised learning

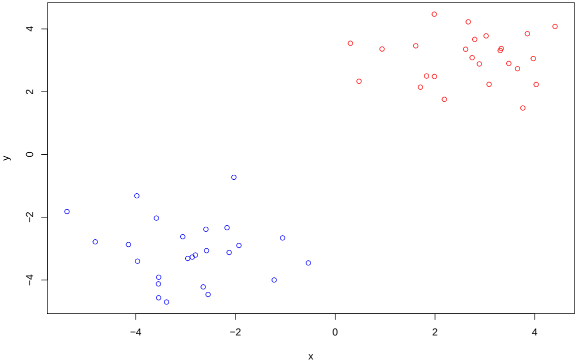
Example: *Clustering*, the partitioning on a number of data points into two or more distinct groups.

```
x1 = cbind(rnorm(25, -3, 1), rnorm(25, -3, 1))
x2 = cbind(rnorm(25, 3, 1), rnorm(25, 3, 1))
D = rbind(x1, x2)
plot(D[, 1], D[, 2], xlab="x", ylab="y")
```



Unsupervised learning

```
kmeans_object = kmeans(D, 2)
cluster_id = kmeans_object$cluster
cols = c("Red", "Blue")
plot(D[, 1], D[, 2], col=cols[cluster_id],
     xlab="x", ylab="y")
```



Unsupervised learning

- often no labels/outputs
- try to find structure/patterns in the data

Revisit the examples for *inference*:

- Security services such as GCHQ use statistical models to *infer* possible threats from voice recordings, email records and video tracking.
- Criminologists aim to *infer* causes of crime based on socio-demographic and descriptive spatial features.
- Biologists and social scientists use statistical models to *infer* what drives human and animal behaviour, and to *predict* what individuals will do when their situation changes.
- Doctors use statistical models to *infer* which patients are likely to respond to a specific drug, and to *predict* which individuals are likely to get sick based on previous medical records.

Optional resources

- The Elements of Statistical Learning. J.H. Friedman, R. Tibshirani and T. Hastie.
- Pattern Recognition and Machine Learning. C. Bishop.
- Machine Learning: A Probabilistic Perspective. K.P. Murphy.