

ACSC/STAT 3740

Predictive Analytics

WINTER 2023
Toby Kenney

In Class Examples

Story Time — Rumpelstilzkin

Stages in Analysis

Analysis Process

- 1 Identify statistical problem.
- 2 Determine and assess useable data.
- 3 Explore data for possible issues and approaches.
- 4 Research any relevant subject knowledge.
- 5 Fit initial models.
- 6 Validate models.
- 7 Fit better models
- 8 Report results.

Problem and Data Sources

Identifying Statistical Problems

Limitations

- Predictive modelling can only determine the relationship between variables. It cannot answer value judgements from the problem.

Criteria for success

- Explanation/Interpretation
- What prediction errors are acceptable?
- Are some errors worse than others?
- What is the relative importance of small vs large errors?
- How important is measuring the uncertainty?

Considerations

- Significance of problem.
- available data
- implementation challenges.

Problem and Data Sources

Identifying Statistical Problems

Example Problem

Should Dalhousie keep all courses online in FALL 2021?

Problem and Data Sources

Assessing Data Quality

Data Sources

- Could source be biased?
- Publication Bias.

Data Collection

- Survivorship Bias.
- Measurement error.
- Participation bias.

Processing

- Removed values.
- Binning.

Meaning of data

- Surrogate Variables

Data Visualisation

Considerations

Why Visualise Data?

- Patterns are often easier to identify from a figure than a table.
- Summary statistics can disguise important features, e.g. outliers.
- For complicated patterns, graphs can convey more information.
- Your eyes have fewer bugs than your R code.

What Data Should we Visualise?

- Sometimes omitting some data obscures the patterns.
- Conversely, putting too much in a single plot can make it difficult to see patterns.

Data Visualisation

Considerations

Who is Looking at the Graph?

- Yourself — e.g. when first exploring data.
- Experts in your field.
- Non-experts willing to spend time examining the graph.
- Non-experts reading quickly.

What do you Want to Show Them?

- General trends.
- Specific patterns

Possible Graph Problems

- Bad data.
- Bad perception.
- Distracting aspects.

Data Visualisation

Channels for Conveying Data

Channels for Conveying Continuous Data

- position
- length
- angle
- area
- depth
- brightness
- colour saturation
- shape

Channels for Conveying Categorical Data

- Hue (red, green, blue)
- Shape

Data Visualisation

Using `ggplot`

Creating a Plot with `ggplot`

- Specify data and mappings.

```
ggplot (data=courses,  
mapping=aes (x=students,y=average_grade,colour=  
subject,shape=term,linetype=as.factor(year)))+
```

- Specify plot type(s)

```
geom_point (colour="red",shape=2,aes (group=year))+
```

- Add labels captions, legends etc.

```
labs(x="This is the x axis (I used a log scale!)  
",y="This is the y axis",title="An example  
plot",subtitle="made with ggplot2")+  
guides (fill=TRUE,shape=FALSE)+
```

Data Visualisation

Using `ggplot`

Creating a Nice Plot with `ggplot`

- Specify axis transformations

```
scale_x_log10()+
```

- Split into subplots

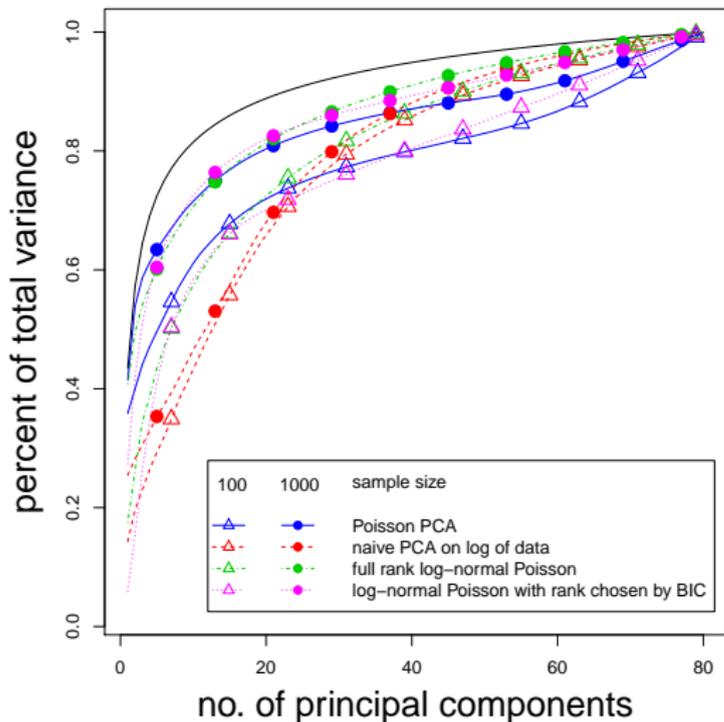
```
facet_wrap(~subject)+
```

- Make adjustments

```
theme(legend.position="left",  
axis.text=element_text(size=12),  
axis.title=element_text(size=14),  
plot.title=element_text(size=25,hjust=0.5),  
plot.subtitle=element_text(size=20,hjust=0.5))
```

Data Visualisation

Examples From Papers

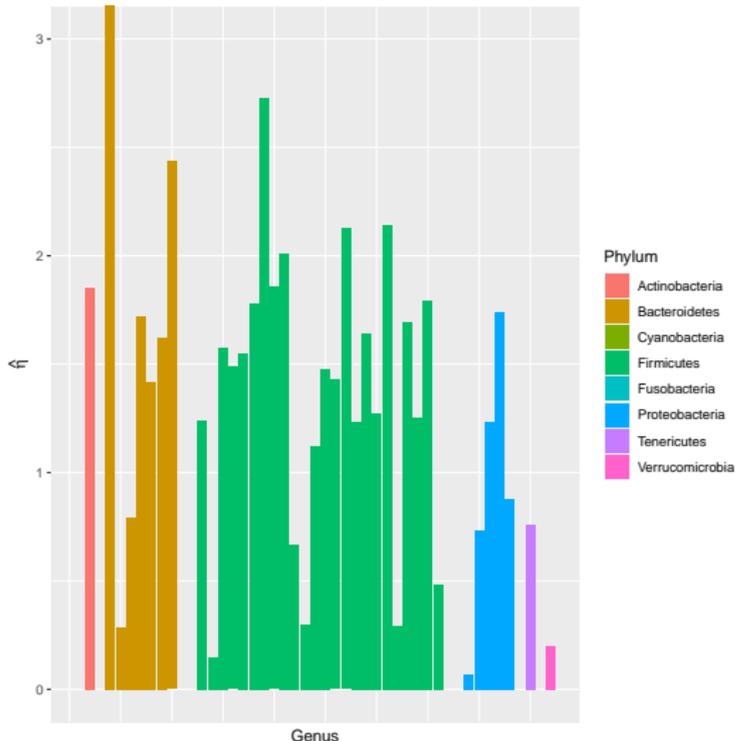


Comments

- Comparing 4 methods at two sample sizes.
- Black line shows theoretical maximum.
- Figure is in Black/White in printed article.

Data Visualisation

Examples From Papers



Comments

- This shows a single variable.
- However, the genus is arranged in a tree structure.
- In addition to the grouping by phylum, there are unshown subgroupings by class, order and family.

Data Exploration

Possible Approaches and Issues

Purpose of Data Exploration

- Identify (and hopefully correct) data issues.
- Decide on suitable modelling frameworks for the data.
- Identify unforeseen hypotheses. These might lead to future studies, or indicate confounding variables that need to be addressed.

Data Exploration Tools

- Data Visualisation
- Dimension reduction

Data Exploration

Possible Approaches and Issues

Missing Data

- Completely at random.
- At random.
- Not at random.

Outliers

- Large influence on results.
- May be data collection errors.
- Sometimes invalid values.

Duplicate Values

- Can influence the results.
- May be data collection errors.
- Can give misleading cross-validation/test results.

Data Exploration

Possible Approaches and Issues

Missing Data

- Completely at random.
- At random.
- Not at random.

Outliers

- Large influence on results.
- May be data collection errors.
- Sometimes invalid values.

Duplicate Values

- Can influence the results.
- May be data collection errors.
- Can give misleading cross-validation/test results.

Handling Data Issues

- Find correct value.
- Remove.
- Impute.

Data Exploration

Exploring Data

Questions to Answer

- Linear or non-linear model?
- Outliers?
- Important variables?
- Are residuals normal?
- Additional features?
- High correlation between predictors?

Simple Visualisations

- Histograms or density plots
- Draw scatterplots.

Dimension Reduction

- PCA

Use Summary Statistics to Identify:

- Outliers.
- Rare values.
- Failure of assumptions.

Data Exploration

Identifying Additional Features

Additional Features

- If most relations are linear, linear regression may be appropriate.
- Nonlinear functions can be fitted by adding transformations of original variables.
- Interaction terms can be added to model dependence between more than two variables.
- Very complicated models may be better modelled using random forest or other flexible methods.
- If predictors are strongly correlated and fairly high-dimensional, principal components may make good features.

Research any relevant subject knowledge

Subject Knowledge

- What sort of relationship is expected?
- Which modelling assumptions are expected to be true?

Validate models

Checking Assumptions

Normal Errors

- Q-Q plots

Independent Errors

- Difficult to detect unless good reason to suspect particular failures.
- Time series models make specific assumptions.

Homoskedasticity

- Conditional variance of response variable does not depend on predictor variables.
- Plot residuals against predicted values.

Validate models

Checking Assumptions

Question 1

The dataset `UKDriverDeaths` gives the monthly number of drivers killed or seriously injured in Great Britain.

A statistician uses the commands

```
library(forecast)
UK_driver_deaths<-
  data.frame(month=seq_len(192),
             deaths=as.vector(UKDriverDeaths))
UK_driver_arma<-auto.arima(UK_driver_deaths$deaths,d=0)
```

to fit an ARMA model to this data set.

Test the assumptions in this model.

Validate models

Measuring Performance

Information Criteria

- Training accuracy with correction for model complexity.
- Several versions — AIC, AICc, BIC, ...

Test Error

- Training error has overfit.
- Test data results more accurately assess model performance on new data.
- Results in smaller training data set.

Cross-Validation

- Multiple training-test splits, average test error over the splits.
- Provides more test data results.

Validate models

Measuring Performance

Question 2

The dataset `UKDriverDeaths` gives the monthly number of drivers killed or seriously injured in Great Britain.

A statistician uses the commands

```
library(forecast)
UK_driver_deaths<-
  data.frame(month=seq_len(192),
             deaths=as.vector(UKDriverDeaths))
UK_driver_arma<-auto.arima(UK_driver_deaths$deaths,d=0)
```

to fit an ARMA model to this data set.

Assess the performance of the model on this dataset.

Validate models

Measuring Performance

Question 3

The dataset `urine` from the package `boot` contains chemical analysis of urine samples. The objective is to predict the presence or absence of calcium oxalate crystals from the other predictors.

A statistician uses the commands

```
library(boot)
urine_logistic<-glm(r~.,data=urine,
                    family=binomial(link=logit))
```

to perform logistic regression to predict the outcome.

Assess the performance of the model on this dataset, and check the assumptions in the model.

Communication

Considerations

Audience

- Account for audience's level of technical and subject knowledge.
- The report may need to be targeted to multiple audiences.

Logical Structure

- Organise the report in a consistent way.
- Start with more general ideas, and develop into more details.

Communication

Parts of Report

Executive Summary/Abstract

- A short concise summary of the conclusions in the report.
- Should inform the reader of the main conclusions of your analysis.
- Often written last.

Introduction

- A short clear definition of the problem and its context.
- This should be precise enough to be answered from the data.
- Include a literature review where appropriate.
- Describe source and nature of data.
- May be appropriate to end with outline of remainder of report.

Data Characteristics

- A summary of the main observations in data exploration.

Communication

Parts of Report

Model Selection and Interpretation

- Start by clearly stating the recommended model.
- Interpret the model.
- Justify the model in comparison to alternative models.
- The statement and interpretation of the recommended model are the main conclusions for non-technical readers.
- Model justification should be streamlined.

Summary and Conclusions

- Repeat the main conclusions.
- Might be more technical than the abstract/executive summary. Usually more detailed.
- May also include suggestions for future studies.

Communication

Parts of Report

Tables and Graphs

- Tables and Graphs in the main document should all make some point.
- Tables and graphs should be easy to read.
- Should be self-contained.
- Only include necessary information.

Summarising Tables and Graphs

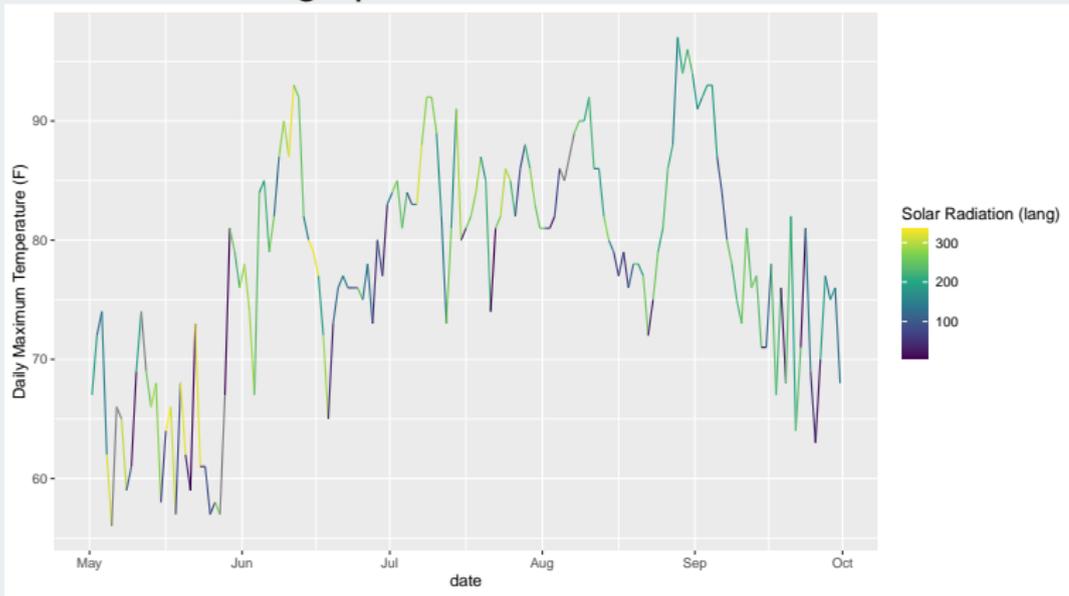
- Focus on interpretation.
- Identify your conclusions. These should require as little context as possible. They should be related to the problem statement.
- Identify the aspects of the data that support the conclusion.

Communication

Parts of Report

Question 1

(a) How should this graph be edited to better show the conclusions?



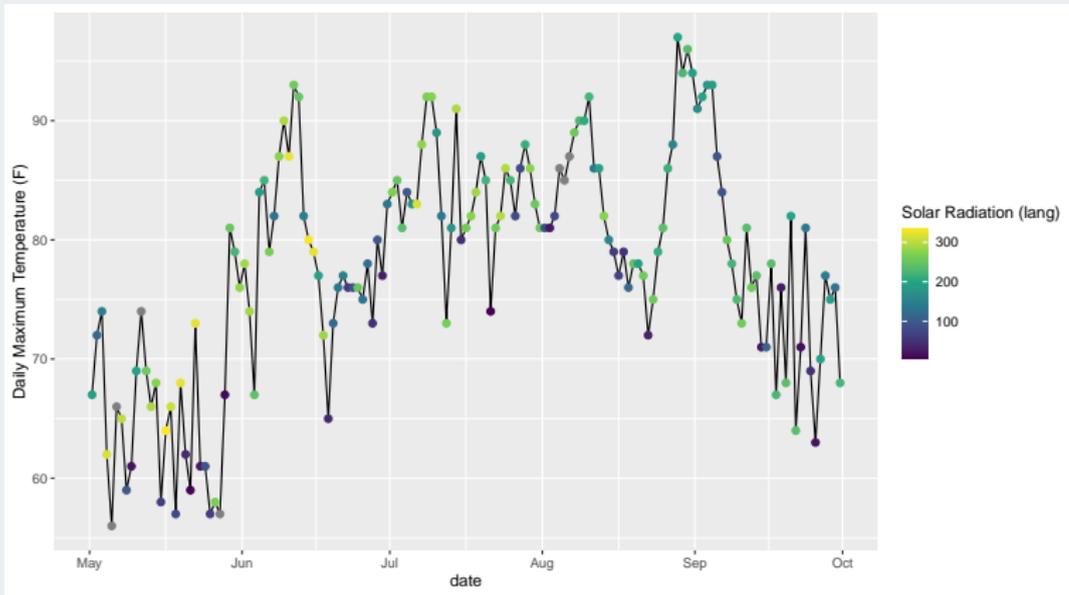
(b) Summarise the main features of the graph.

Communication

Parts of Report

Answer to Question 1

(a)

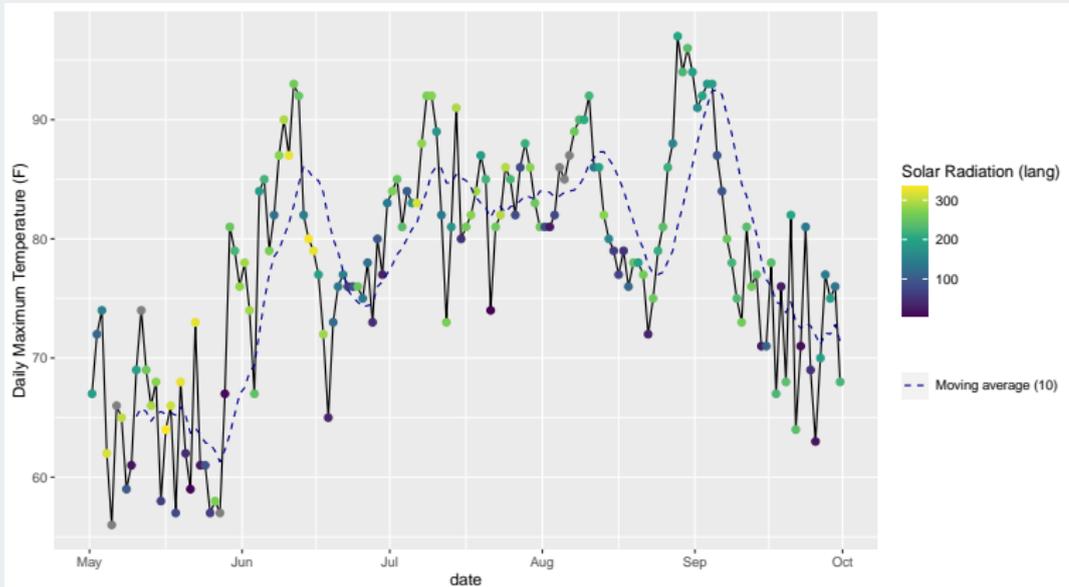


Communication

Parts of Report

Answer to Question 1

(a)



Answer to Question 1

(b)

- Very low solar radiation corresponds to low temperatures for the time of year.
- High solar radiation is more common in spring and early summer.
- Low solar radiation is common throughout the year.
- Temperature is highest in summer.
- Daily temperature fluctuation was highest in late September
- Fluctuations in the 10-day moving average were higher in summer.

Linear Regression

Revision

Linear Regression

- Fits models of the form $y = X\beta + \beta_0 + E$ for some vector β .
- Can add functions of existing predictors as new predictors.
- Fitted by least squares. This is MLE for normal residuals.

Assumptions

- Error is normal.
- Errors are independent.
- Homoskedasticity

Diagnostics

- Residuals vs. fitted values
- Q-Q plot of residuals.

Limitations

- High dimensions.
- Correlation between predictors.

Linear Regression

Revision

Question 1

The data set `Boston` contained in the `MASS` package in `R` describes house prices.

- (a) Perform a linear regression of median value on the other variables.
- (b) Perform diagnostics to assess whether the linear regression model is suitable.
- (c) Use a transformation of median value to improve the regression.
- (d) Add additional predictors to improve the regression.

Time Series

Introduction

Idea

- Assumption of independent errors is not valid.
- Add previous values as predictors.
- Also need to add time (and possibly functions of it) as a predictor.

Assumptions

- Error is normal.
- Homoskedasticity
- Stationarity

Diagnostics

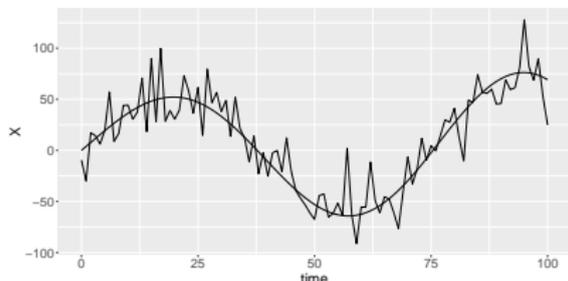
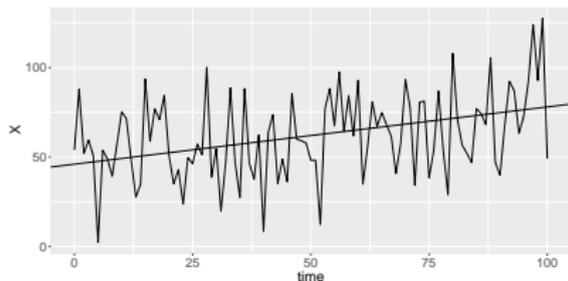
- Residuals vs. fitted values
- Q-Q plot of residuals.
- Dickey-Fuller Test

Time Series

Modelling Trends

Idea

- Model $y_t = f(t) + \epsilon$ a time trend plus random error.
- Time trend is usually long-term trend plus (or multiplied by) seasonal trend.



Note

- Error is i.i.d..
- Top figure shows linear trend
- Bottom figure shows linear trend multiplied by seasonal trend.

Question 1

The data set `EuStockMarkets` in R contains daily stock market data for four european markets between 1991 and 1998.

[The data are every working day, so are not evenly spaced. However, for the purpose of this analysis, we will assume they are evenly spaced.]

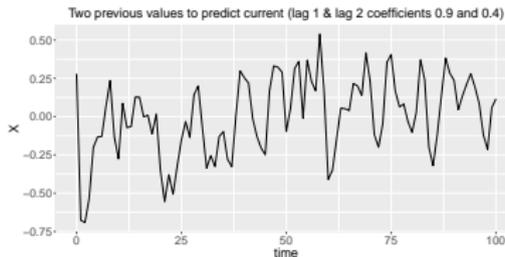
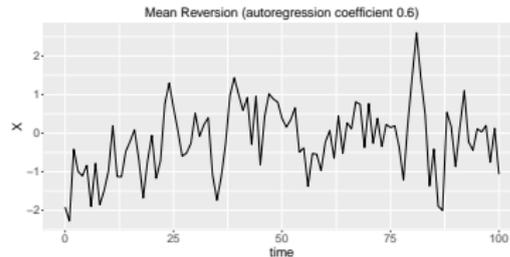
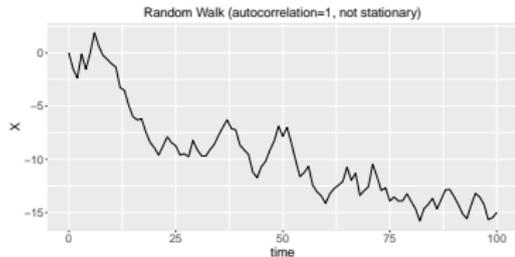
- (a) Fit a quadratic trend to the log-DAX value.
- (b) Plot the residuals over time and other diagnostic plots.

Time Series

Autocorrelation

Idea

- Even after removing trend, the values at different time points are not independent, with adjacent time points much nearer.
- Deal with this by including previous time points as predictors.



Time Series

Autocorrelation

Question 2

Using the detrended data from the previous question:

- (a) Fit an autoregressive model on the DAX.
- (b) Plot the residuals over time and other diagnostic plots.

Time Series

Moving Averages

Idea

- Linear moving average $s_t = \frac{x_t + \dots + x_{t-k+1}}{k}$
- Exponential moving average $s_t = (1 - w) \sum_{i=1}^t w^{t-i} y_i$

Notes

- Exponential moving average of i.i.d. variables is AR 1
- Still assuming Homoskedasticity
- Weighted linear moving average of AR process is i.i.d.

ARMA model

- moving averages follow AR process
- arises naturally as sum of AR processes

Time Series

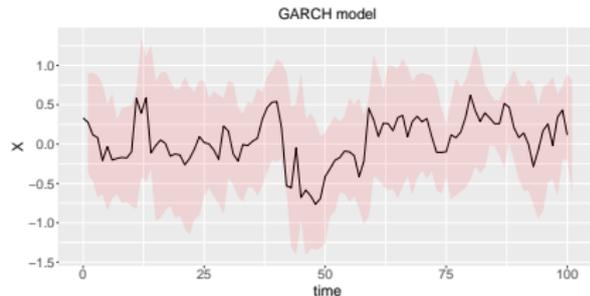
GARCH — Generalised Autoregressive Changing Heteroscedasticity

Idea

- Sometimes the variance of a time series follows a time series of its own.
- ARCH(p) — Conditional variance $\sigma_t^2 = \omega + \gamma_1 \epsilon_{t-1}^2 + \dots + \gamma_p \epsilon_{t-p}^2$
- GARCH(p,q) —
 $\sigma_t^2 = \delta_1 \sigma_{t-1}^2 + \dots + \delta_q \sigma_{t-q}^2 + \omega + \gamma_1 \epsilon_{t-1}^2 + \dots + \gamma_p \epsilon_{t-p}^2$

Notes

- Assume error is normal.
- Usually fitted by MLE
- Variance often important for time series, particularly financial.



Time Series

GARCH — Generalised AutoRegressive Changing Heteroscedasticity

Question 3

Fit a GARCH model to the DAX data studied in the previous questions.

Generalised Linear Models

Revision

Idea

- Specified conditional distribution for response (e.g. Bernoulli)
- Transformed conditional expectation $f(\mu_i)$ is fitted via regression.
- Coefficients fitted via MLE.

Assumptions

- Conditional distribution of response follows specified distribution.
- Homoskedasticity or fixed value of other parameters

Diagnostics

- Raw residuals $y_i - \mu_i$ don't have good properties.
- Several alternative residuals.
- Deviance residuals — root of log-likelihood difference.

Limitations

- High dimensions.
- Correlated predictors.

Question 1

The dataset `iris` in R contains measurements of three different species of iris plants.

- (a) Use logistic regression to classify the samples from the `versicolor` and `virginica` species.
- (b) Plot the deviance residuals.
- (c) The predictor `Sepal.Width` is not strongly correlated with the other predictors. Fit a model with this predictor removed. Why do the coefficients change?

Variable Selection and Regularisation

Revision

Idea

- Too many predictors result in bad models or even no model.
- Select only the most important predictors, get better results.

Variable Selection Methods

- Search based on goodness of fit.
- Penalty based — LASSO, ridge regression.

Search methods

- Forward Selection
- Backward Selection

Goodness of Fit

- Information Criteria.
- (Generalised) Cross validation
- Hypothesis testing

Penalties

- $L^0 - |\{i|\beta_i \neq 0\}|$
- LASSO - $\sum |\beta_i|$
- Ridge Regression - $\sum \beta_i^2$

Variable Selection and Regularisation

Revision

Question 1

The data set `longley` in R contains a number of economic data points.

- (a) Fit a linear model to predict the variable `Employed` from the other variables.
- (b) Use forward selection to select only the important variables in this model.
- (c) Use backward selection to select the important variables.
- (d) Use LASSO to select the important variables.
- (e) Use ridge regression to fit a model.

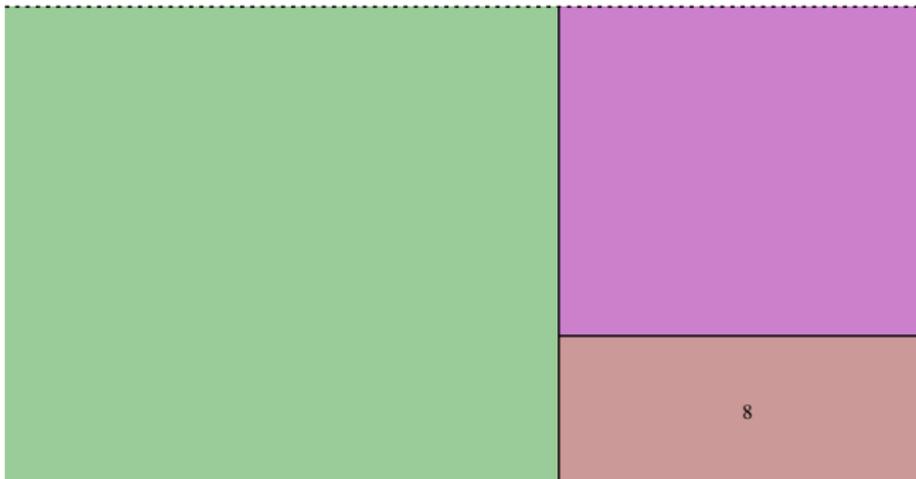
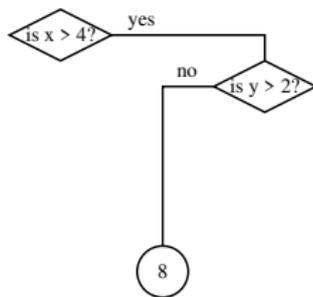
Tree-based Methods

Revision



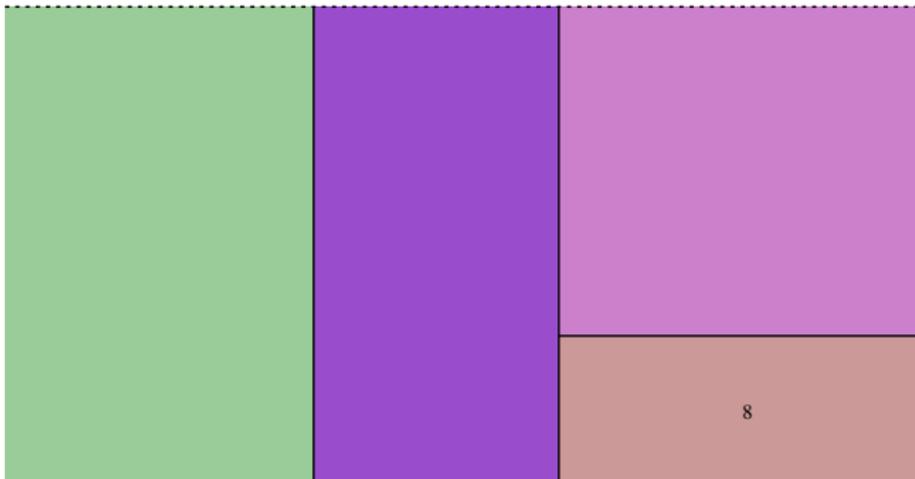
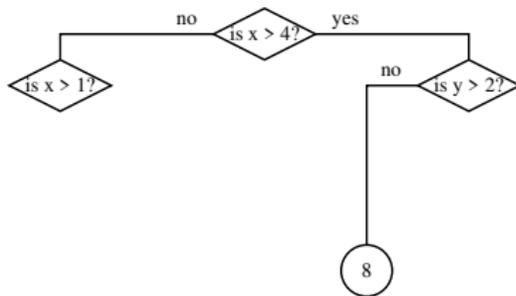
Tree-based Methods

Revision



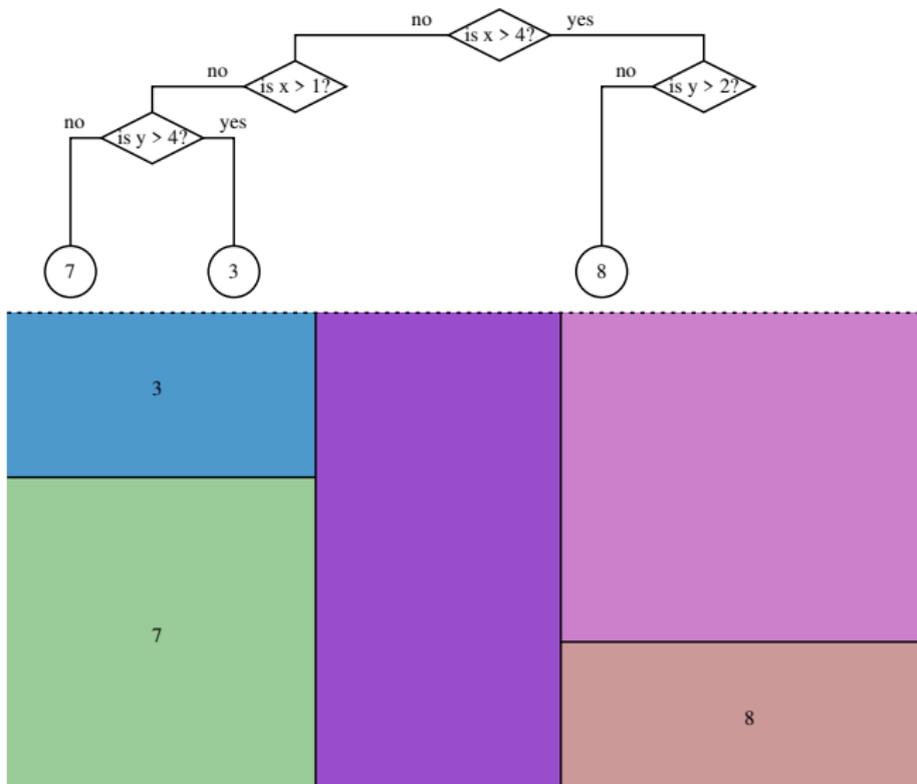
Tree-based Methods

Revision



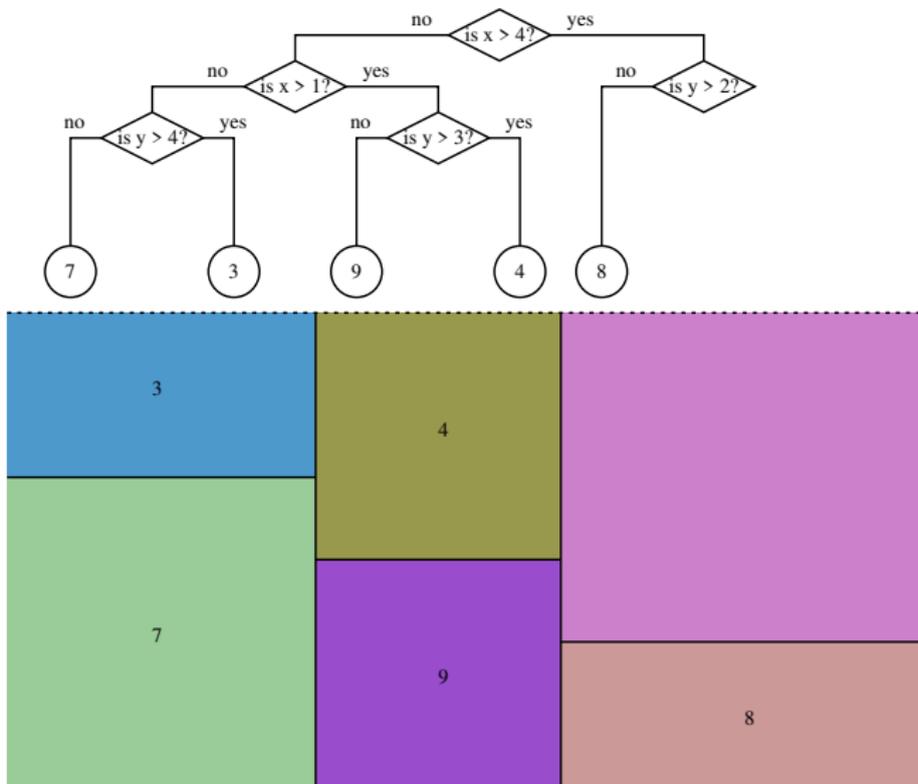
Tree-based Methods

Revision



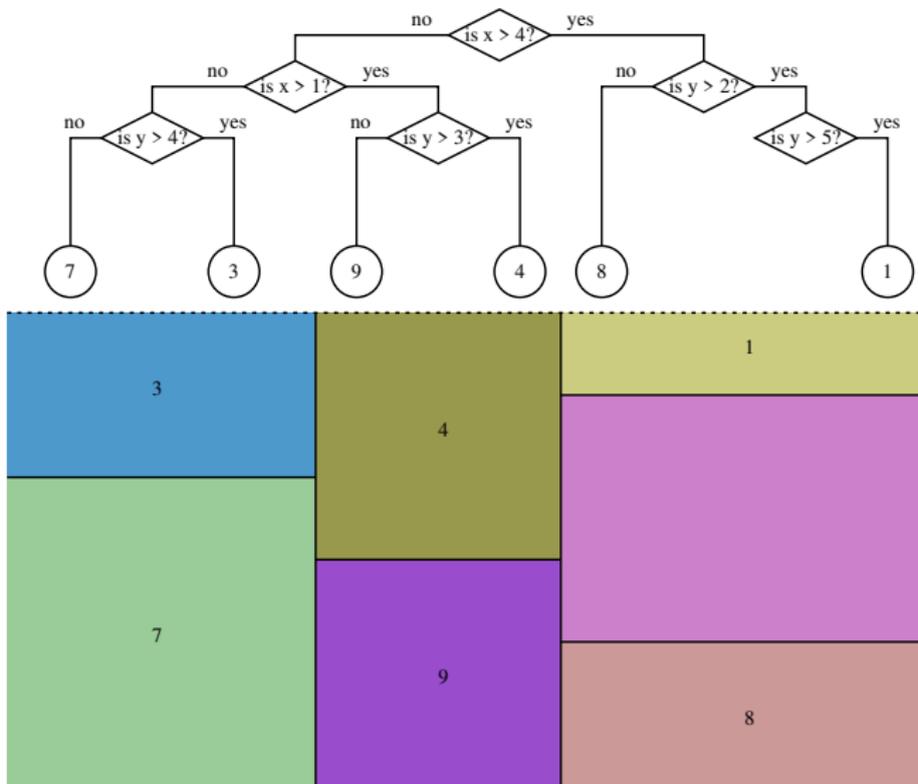
Tree-based Methods

Revision



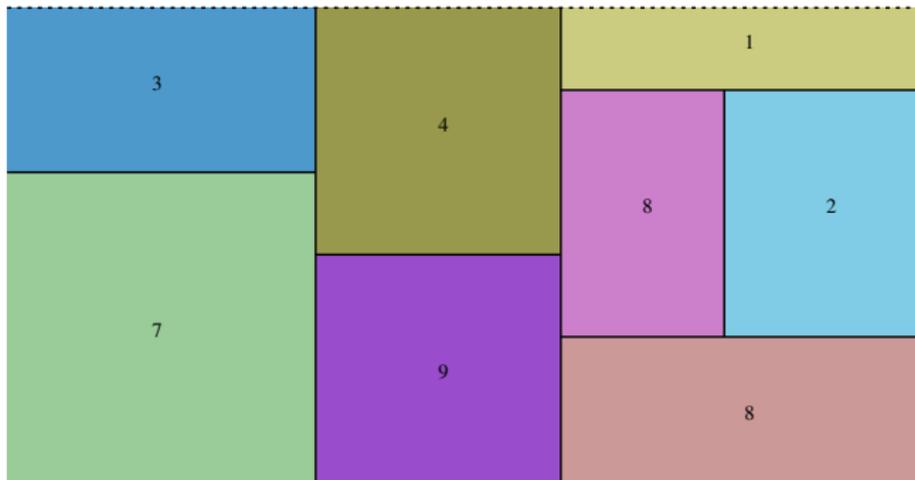
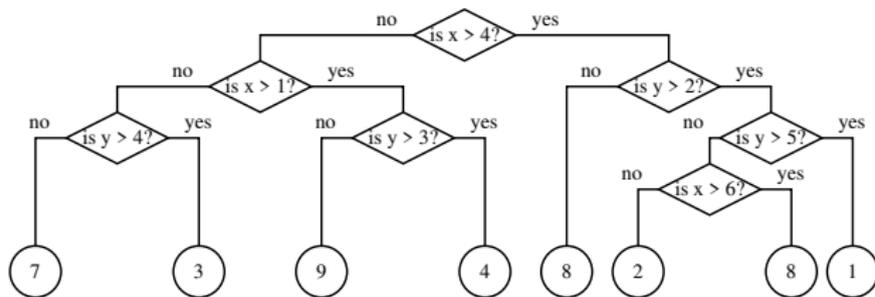
Tree-based Methods

Revision



Tree-based Methods

Revision



Tree-based Methods

Revision

Idea

- Divide region into rectangular blocks, assign value to each block.
- Equivalent to a decision tree.
- Various methods to avoid overfitting.
- One tree not flexible enough, so average many trees.

Decision Trees

- Cut each leaf node to best improve results.
- Limit complexity either with maximum depth or minimum node size.

Random Forest

- Fits many decision trees.
- Subsets data and variables to make trees different.

Boosted trees

- Fits trees using residuals from current model.

Tree-based Methods

Revision

Question 1

The data file `pollution.txt` contains pollution data from McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', *Technometrics*, vol.15, 463– 482.

(Downloaded from

<http://lib.stat.cmu.edu/datasets/pollution>)

- (a) Fit a decision tree to estimate Nitrous Oxide pollution (NOX). Use a number of tuning parameter values on a training sample, and compare the accuracy on test data.
- (b) Fit a random forest model to estimate NOX , using several different tuning parameters, and compare the accuracy.
- (c) Use boosted trees to estimate NOX .

PCA and Clustering

Revision

Idea

- Transform data into smaller number of principal components.
- Principal components are linear combinations of variables.
- PCs are uncorrelated, and minimise squared error.
- Sometimes standardise (correlation matrix instead of covariance).

Assumptions

- If we assume the data are multivariate normal, then:
 - Squared error is supported by likelihood theory.
 - Principal components are independent.

Diagnostics

- Scree plot — used to choose number of principal components.

Limitations

- High dimensions
- Interpretability

Question 1

The data set `iris` contained in the `datasets` package in R contains measurements of several iris plants.

- (a) Perform a principal component analysis to find the main directions of variation..
- (b) Make a scree plot to assess how many principal components to analyse.
- (c) Plot the loadings and show how this relates to species.
- (d) Repeat this using correlation instead of covariance.

Clustering

Idea

- Identify unknown groups within the data.
- Individuals from different groups follow different distributions.

Different clustering method

- *K*-means — normal groups with identity variance
- Mixture model — i.i.d. sample from a mixture of components.
- Hierarchical clustering — treat clusters as points and cluster them.

Choosing No. of Clusters

- Plot sum of squared distances vs. no. of clusters. Find elbow point.
- Gap statistic — compare with an expected graph (using simulations).

Limitations

- Sometimes no true answer
- Can be sensitive to outliers.

Clustering

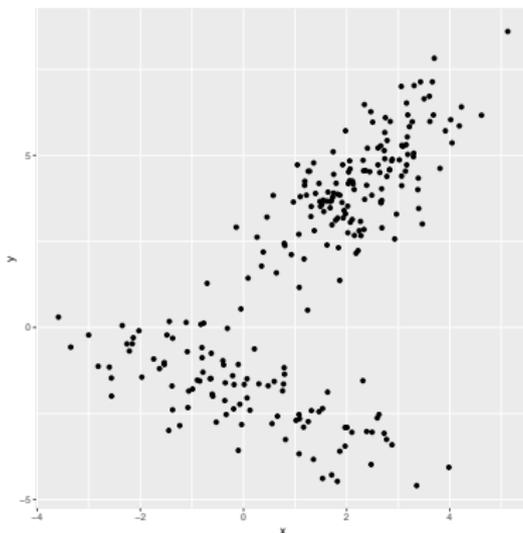
K-means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Clustering

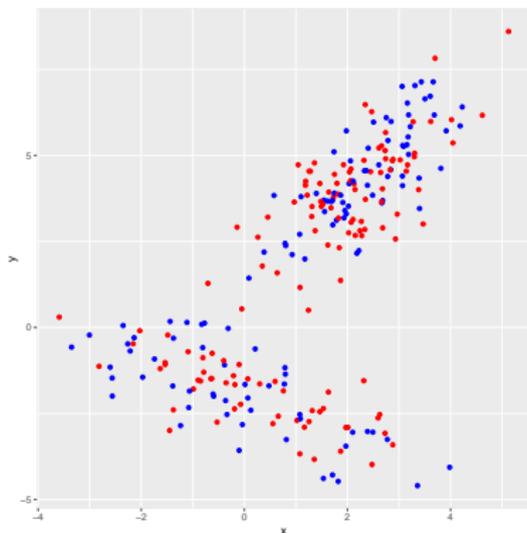
K -means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Clustering

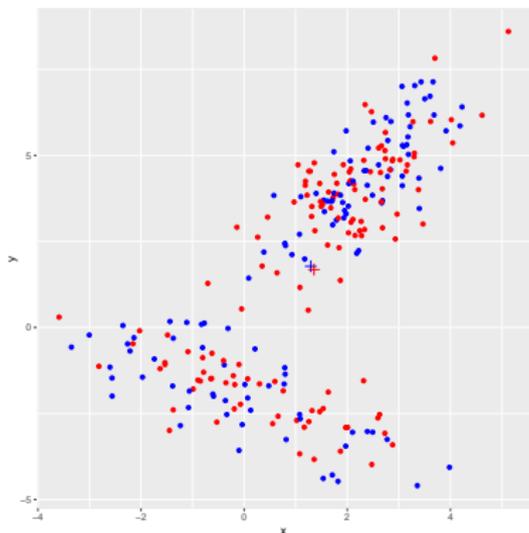
K -means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Clustering

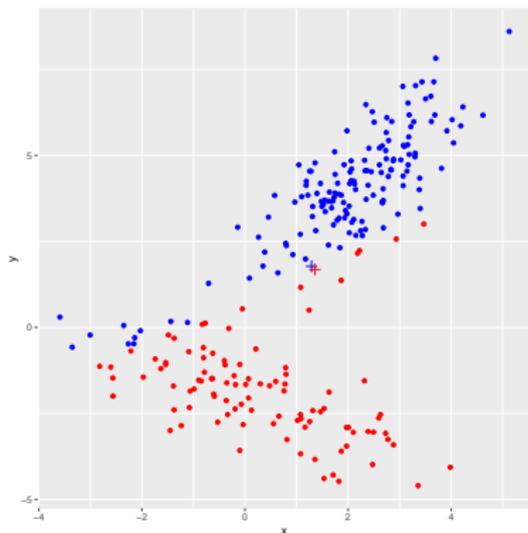
K -means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Clustering

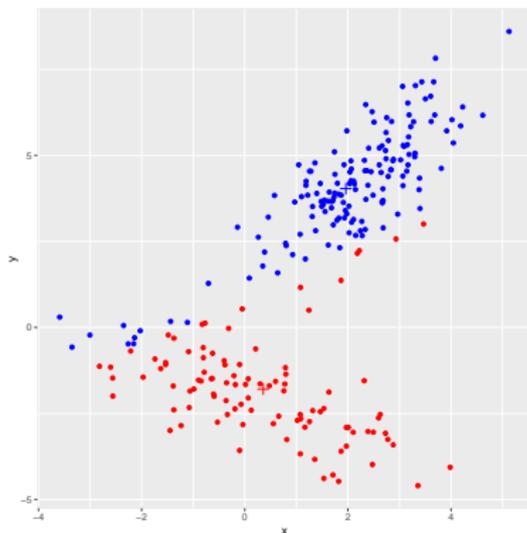
K -means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- Reassign points to clusters.
- Repeat until convergence.



Clustering

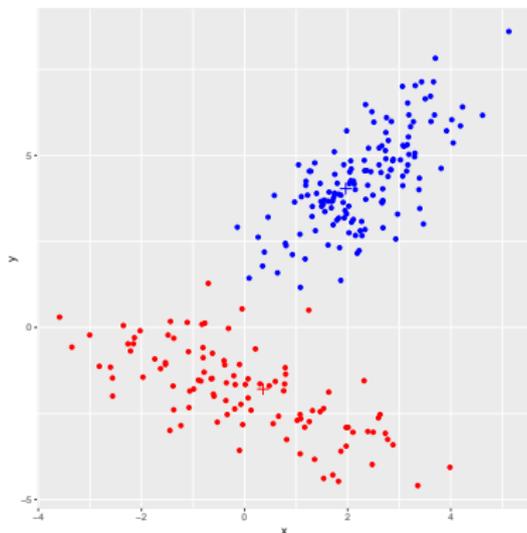
K-means clustering

Problem

- Fix no. K of clusters.
- Minimise sum of squared distance to cluster centres.

Algorithm

- Start with a random assignment to clusters.
- Calculate the cluster centres
- **Reassign points to clusters.**
- Repeat until convergence.



Clustering

K -means clustering

Question 1

For the iris data set:

- (a) use K -means clustering on the measurements (not using the species) to cluster the plants.
- (b) Choose the appropriate number of clusters.
- (c) How do the clusters compare with the species of the plants?