STAT5003

Week 13 : Review and Final Exam

Jaslene Lin
The University of Sydney

THE UNIVERSITY OF SYDNEY

Exam Format

Final Exam Details

🗓️ Date: Monday 24 November 2025

🕓 Starting Time: 5:00pm

📍 Location: **Check Your Exam Timetable**

Final Exam Cover Sheet

THE UNIVERSITY OF SYDNEY

Room Number _____
Seat Number _____
Student Number _____
ANONYMOUSLY MARKED
(Please do not write your name on this exam paper)

CONFIDENTIAL EXAM PAPER

This paper is not to be removed from the exam venue

Mathematics and Statistics

EXAMINATION

Semester 2 - Final, 2025

STAT5003 Computational Statistical Methods

EXAM WRITING TIME: 2 hours
READING TIME: 10 minutes

EXAM CONDITIONS:
This is a RESTRICTED OPEN book exam - specified materials permitted

MATERIALS PERMITTED IN THE EXAM VENUE:
(No electronic aids are permitted e.g. laptops, phones)

Bilingual dictionary

Calculator - non-programmable

One A4 sheet of handwritten notes double-sided

MATERIALS TO BE SUPPLIED TO STUDENTS:
1 x 12-page answer book

Answer sheet: Gradescope MCQ (single-sided - 100 Qs)

INSTRUCTIONS TO STUDENTS:

This booklet contains 16 pages, including this cover sheet and 15 pages of questions. There are 16 questions.

Please tick the box to confirm that your examination paper is complete. ☐

For Examiner Use Only

Q	Mark
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	

Total _____

Exam Structure Overview

- Multiple Choice Section (9 questions 18 marks)
- Extended Answer Section (7 questions 42 marks)

In summary, the final exam accounts for 50% of the unit's total mark.

Multiple Choice Section

*There are **two correct answers** for each question. Choose all correct answers for each question. Each question is worth two marks. The answer needs to be completely correct to receive the two marks. The total mark for this section is 18.*

- You can select at most two options for each question. Otherwise, you automatically get 0 for the question.
- For every correct response that is selected, a mark is awarded. For every **incorrect** response that is selected, a mark is **deducted**.
- A mark for a question cannot be negative even if only incorrect responses are selected.
- Non-selected responses do not modify any awarded marks. There are no marks awarded for not answering a question.

Your answers must be entered on the Multiple Choice Answer Sheet.

Extended Answer Section

- Two groupset questions on scenarios with R outputs;
- Evaluation of a classification model;
- Compare and contrast of two model algorithms;
- Assessment of classification model performance using performance metrics;
- Scenario-based critique with proposed improvements to enhance the workflow; and
- Completion (via pseudo code) of a Monte Carlo to analyse a complex phenomenon

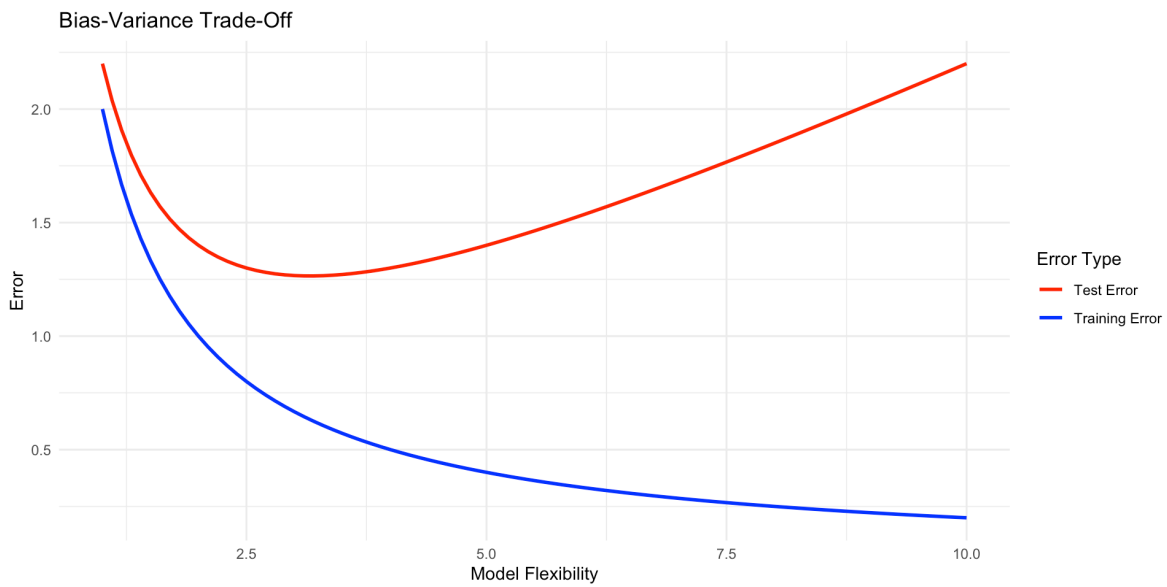
Review

Review: Statistical Learning Methods

- Regression
  - Linear regression
  - Univariate smoothing (nonlinear) regression
  - Regression Decision Trees
  - Random forests
  - Gradient descent boosting trees
- Classification
  - Logistic regression
  - LDA
  - kNN
  - SVM
  - Classification Decision Trees
  - Random forests
  - Adaboost
- Clustering and High Dimensional Viz
  - Hierarchical clustering
  - K-means clustering
  - PCA
  - t-SNE
  - MDS
- Statistical Methods
  - MLE and KDE
  - Resampling methods including Cross-Validation and Bootstrap
  - Monte Carlo methods and MCMC

Model Performance Metrics

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



Performance Metrics for Regression Models

These metrics focus on how closely predicted values align with actual values.

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

R-Squared: Measures the proportion of variance in the target variable explained by **a linear** model, providing an overall measure of goodness-of-fit.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Adjusted R-Squared: adjusted for the number of predictors, used for model selections.

Performance Metrics for Classification Models

For classification, metrics assess how well a model correctly classifies categorical outcomes.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

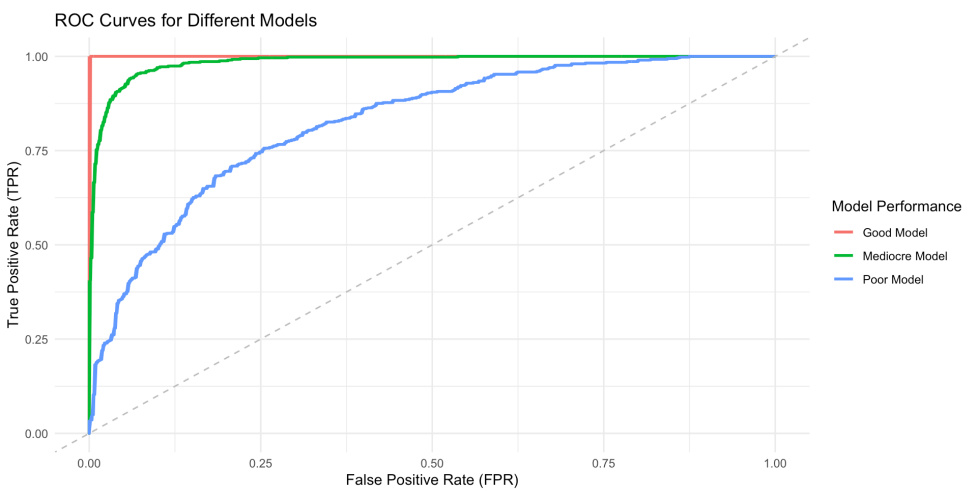
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Cohen's Kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Area Under the ROC Curve (AUC-ROC):



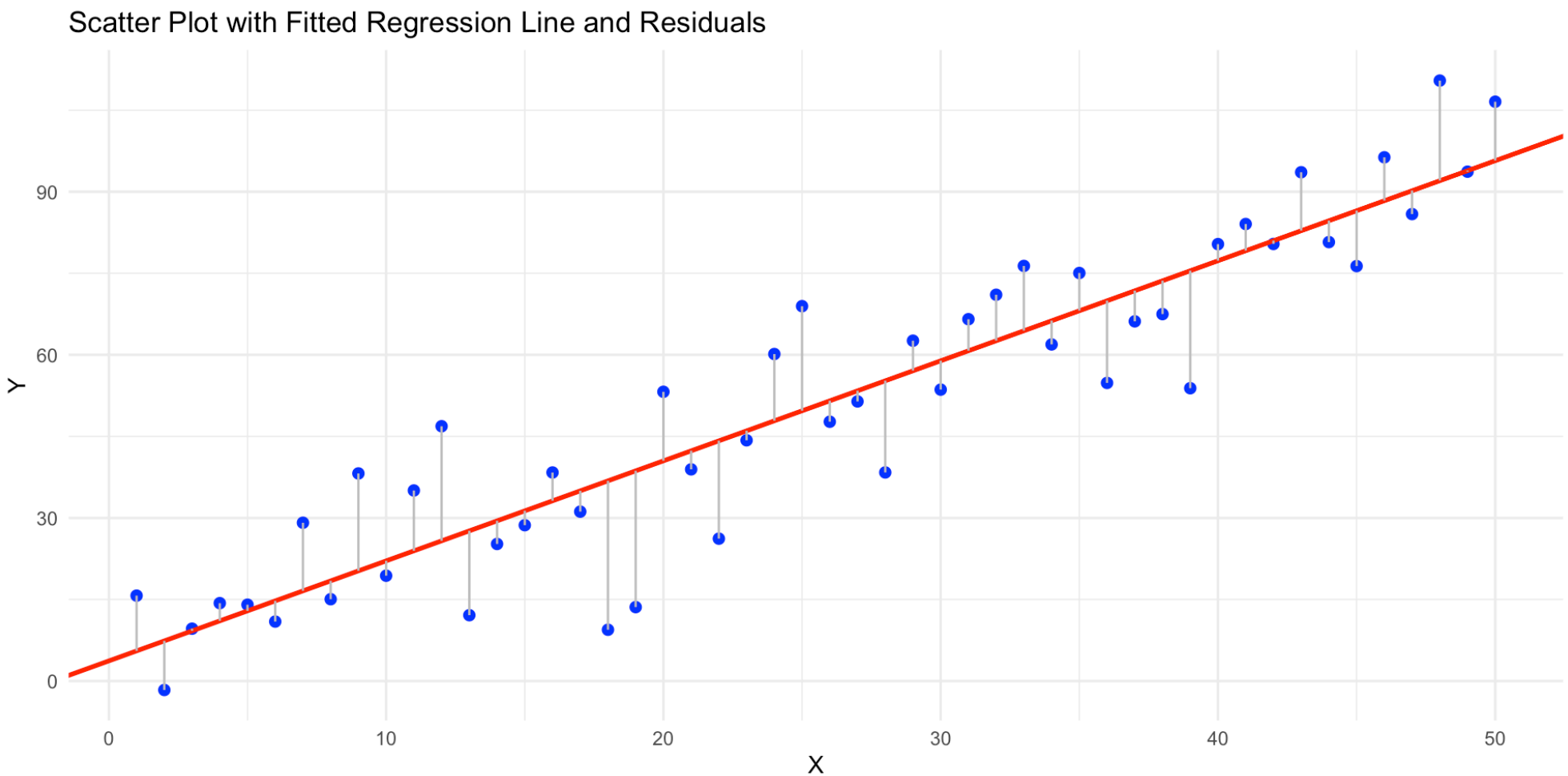
Performance Metrics for Classification Models

Question: When dealing with imbalanced classes in a classification problem, which performance metrics are most effective for evaluating model performance?

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Find coefficients to minimize the total sum of squares of the residuals



Linear Model Selection and Regularisation

Feature Selection

- Best subset selection
- Forward selection
- Backward selection
- Choose a model that minimises test error
    - ➟ Directly via test set (CV errors)
    - ➟ Indirectly via penalised criterion (Adjusted R-square; AIC and BIC)

Ridge Regression and Lasso

- Constrained optimisation techniques that minimise the squares with different constraints.
- Lasso has the extra benefit of feature selection as a free bonus.

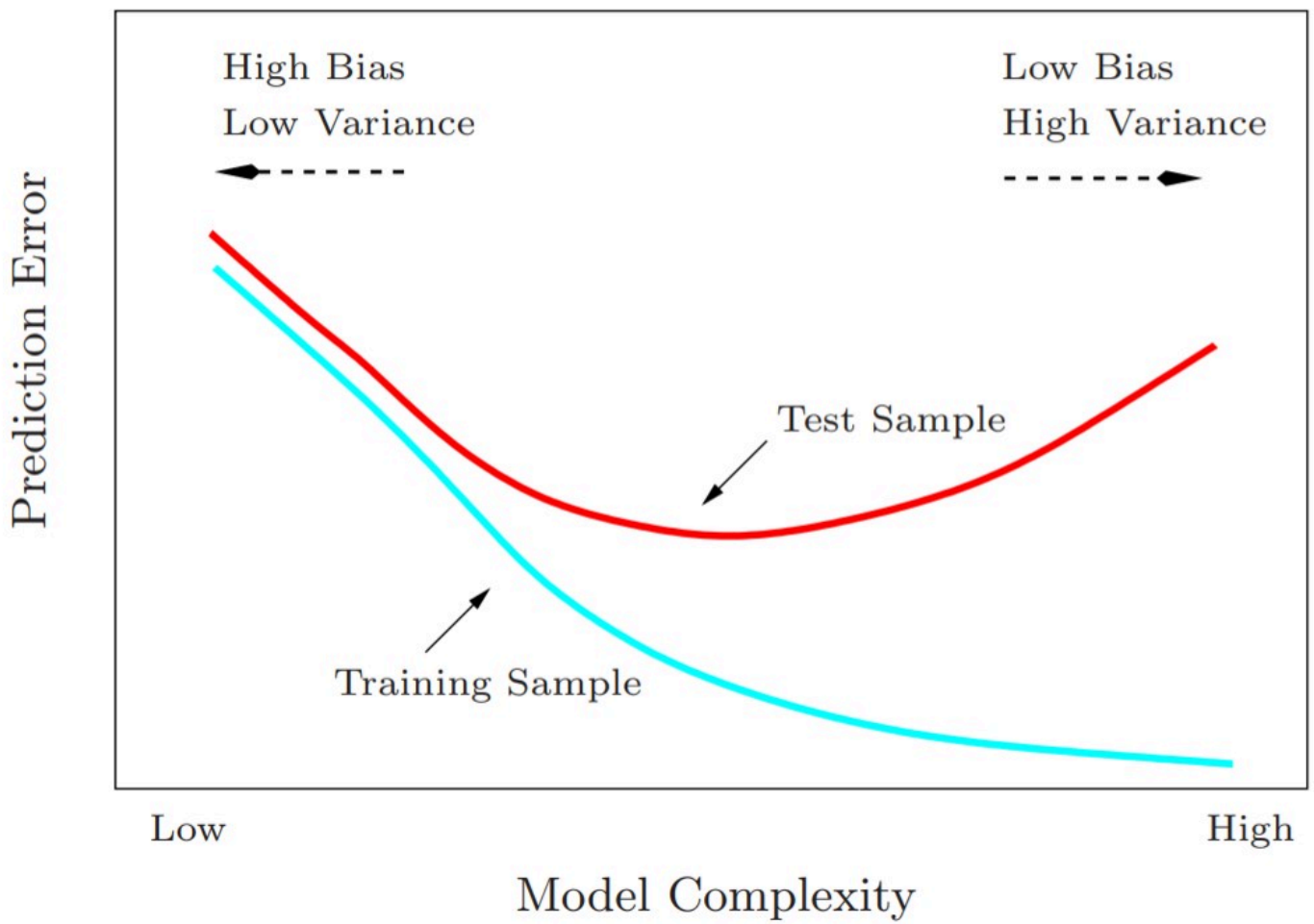
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 \qquad \text{subject to} \qquad \sum_{j=1}^{p} |\beta_j| \leq s.$$

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 \qquad \text{subject to} \qquad \sum_{j=1}^{p} \beta_j^2 \leq s.$$

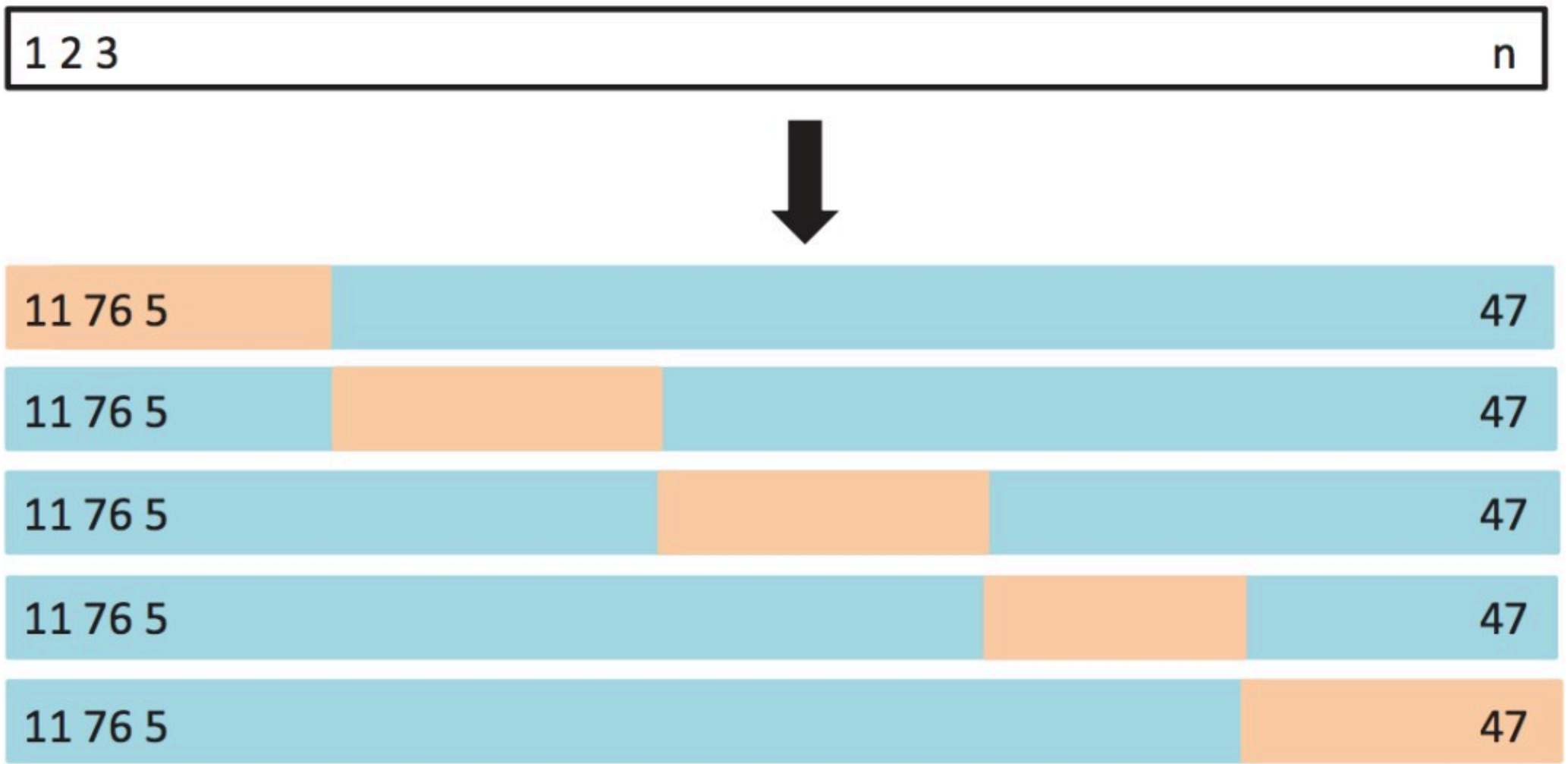
- Question: what is the function in `R` to fit lasso or ridge? If you want to fit a lasso model, what value you should set for the `$\alpha$` in that function?

Cross validation

- Fitting model to entire dataset can overfit the data and not perform well on new data
- Split data into training and tests sets to alleviate this and find the right bias/variance trade-off



5-fold Cross Validation

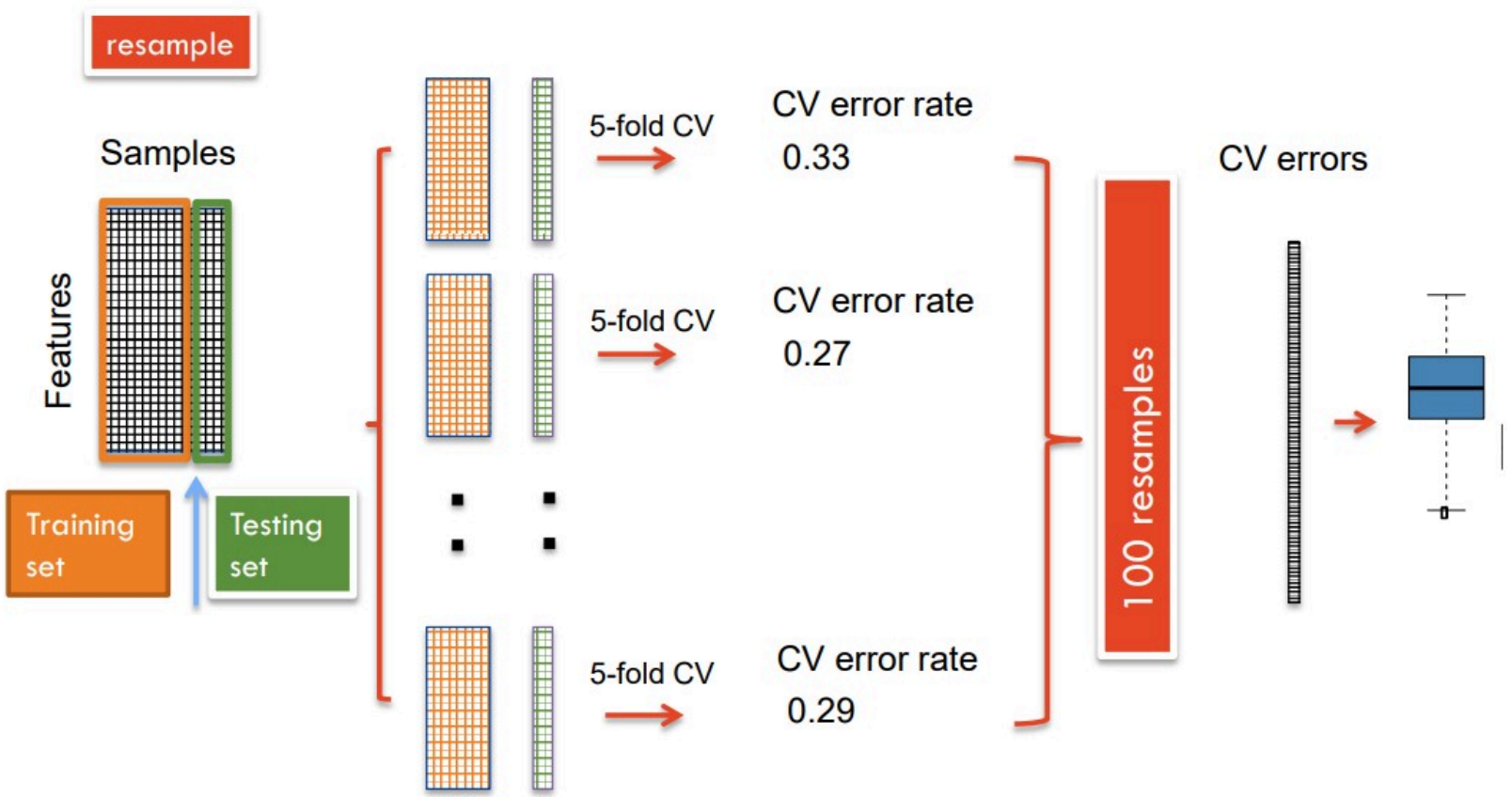


Repeated Cross Validation

It repeats the k-fold CV process multiple times, each with different random splits.

This helps to: provide a less biased CV test error estimate. provide the variance of the CV error.

It comes with **a computational cost**.

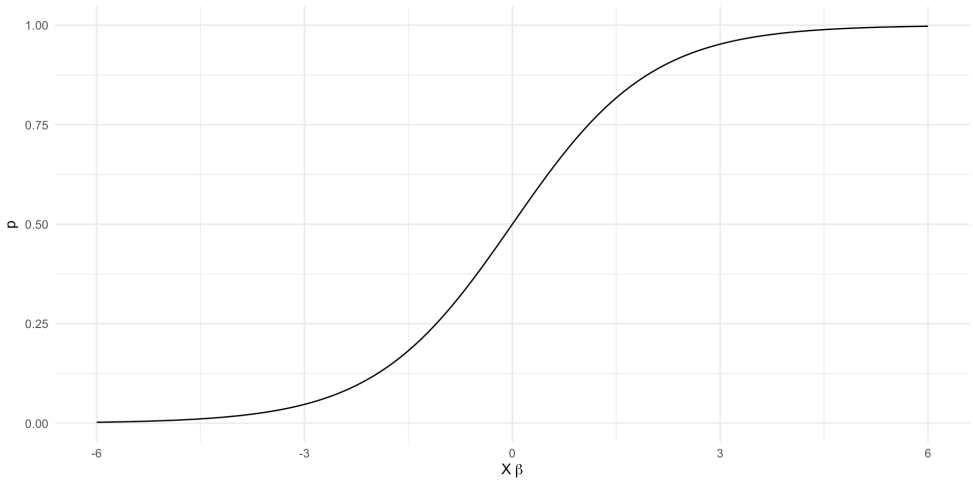


Logistic Regression model

- Model the mean non-linearly $g(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]) = \boldsymbol{X}\boldsymbol{\beta} = \mu$
- $\log\left(\frac{p}{1-p}\right) = \boldsymbol{X}\boldsymbol{\beta}$
- Solve for $\boldsymbol{p}$ gives

  ➠ $p = P(Y = 1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{X}\boldsymbol{\beta})}$

```r
xs <- seq(-6, 6, length = 512)
y <- 1/(1 + exp(-xs))
ggplot(data.frame(x = xs, y = y)) + geom_line(aes(x
  labs(x = bquote(X~beta), y = bquote(p))
```



Linear Discriminant Analysis (LDA)

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

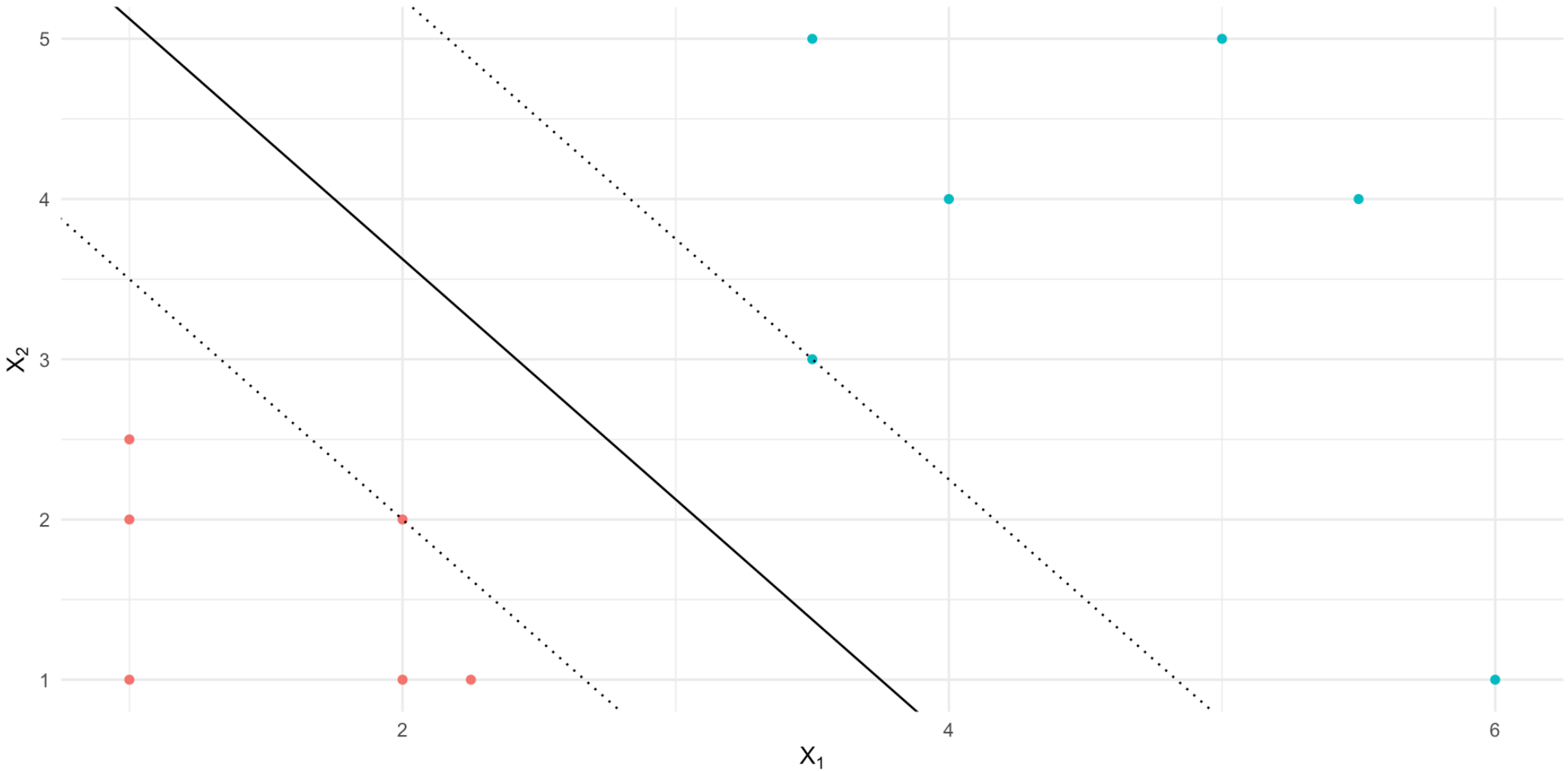
Posterior: The probability of classifying observation to group k given it has features x

Prior: The prior probability of an observation in general belonging to group k

- $f_k(x)$ is the density function for feature x given it's in group k

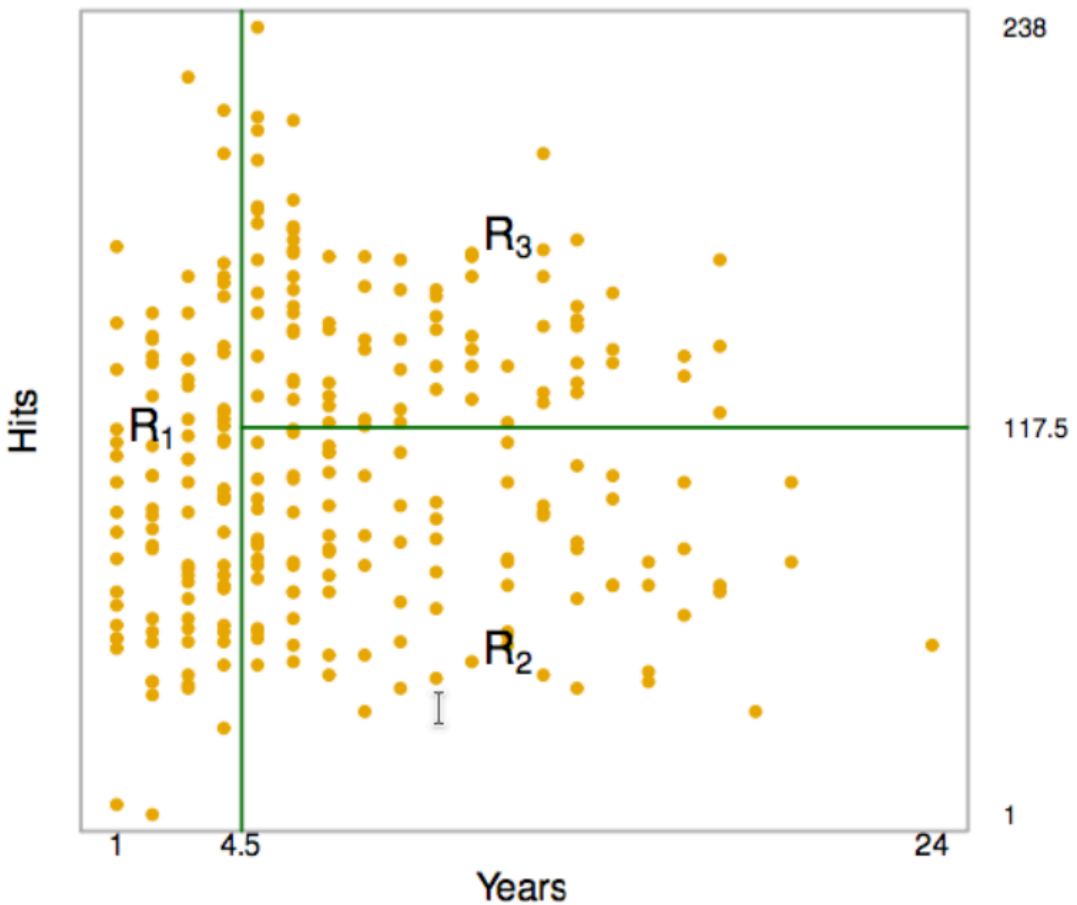
Support Vector Machines (SVM)

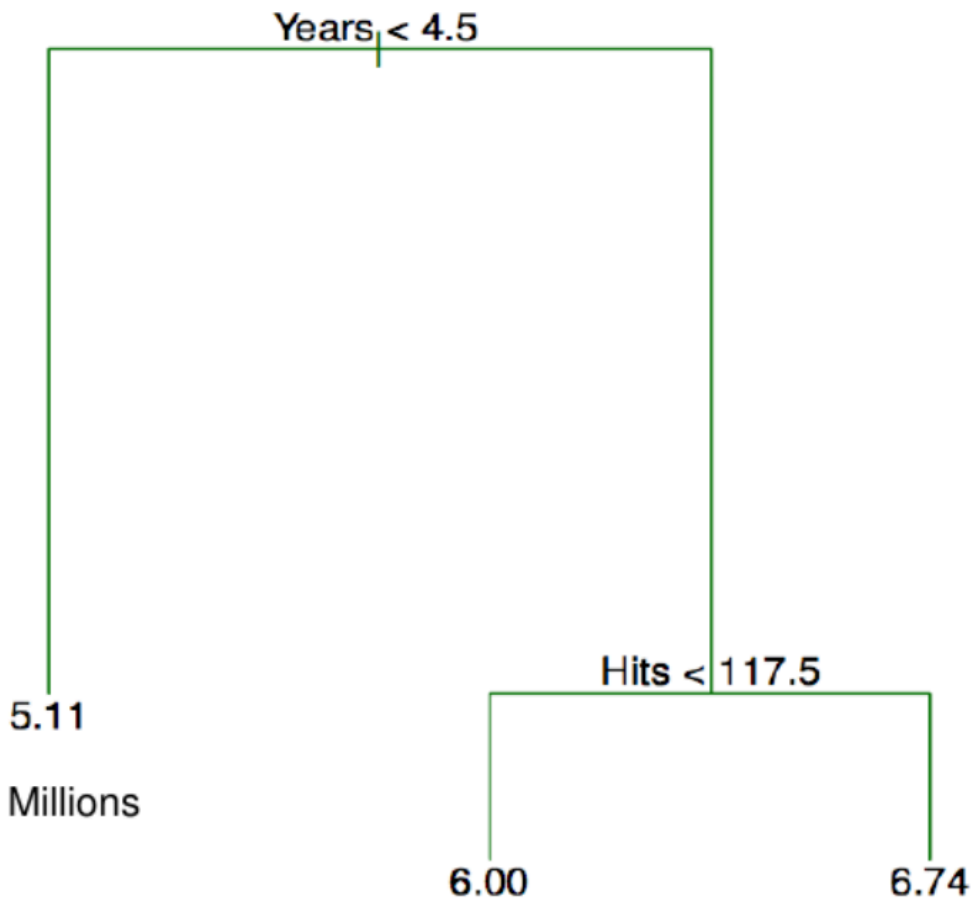
- Find the best hyperplane or boundary to separate data into classes.



Basic decision trees

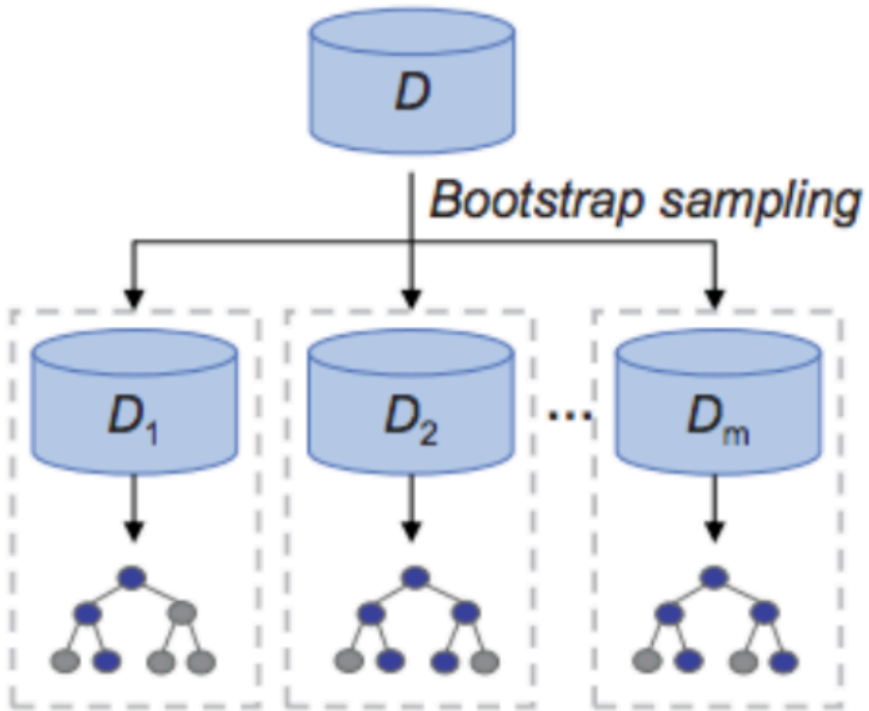
- Partition space into rectangular regions that minimise loss in predictions.





Bagging trees and random forests

- Use bootstrap technique to create resampled trees and average the result
- $\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b^*(x)$
- Random forests do further subsampling of predictors at split to improve model



How can the out-of-sample (test) error be estimated when using a bagging model?

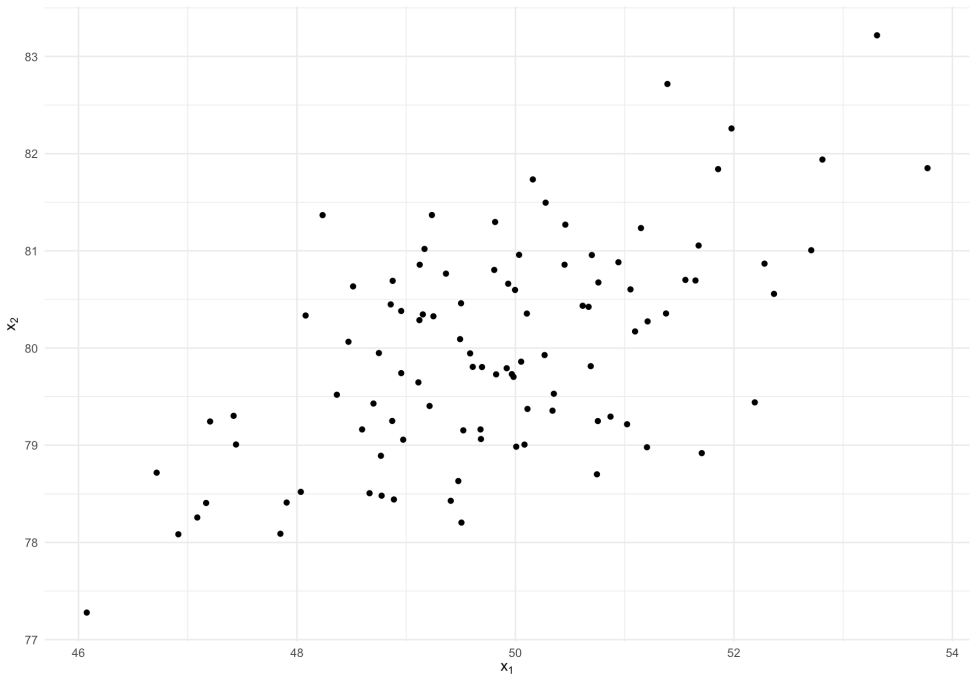
Boosting

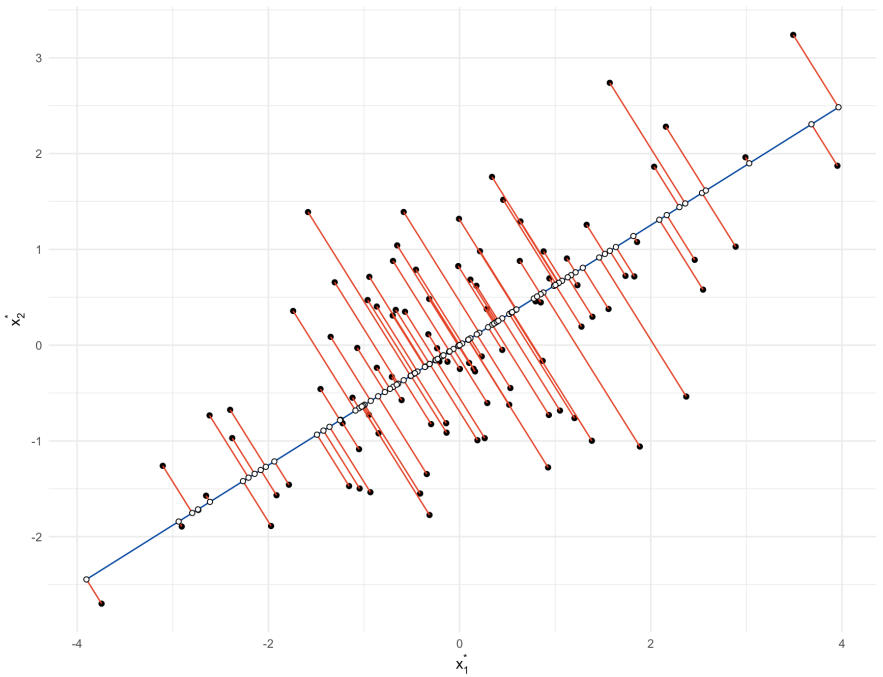
- Fit tree to residuals and learn slowly
- Slowly improve the fit in areas where the model doesn't perform well
- Some boosting algorithms discussed
    - AdaBoost
    - Stochastic gradient boosting
    - XGBoost

What are the key hyper-parameters need to be tuned when fitting gradient boosting? How does these hyper-parameters impact on the bias-variance trade-off in the model performance?

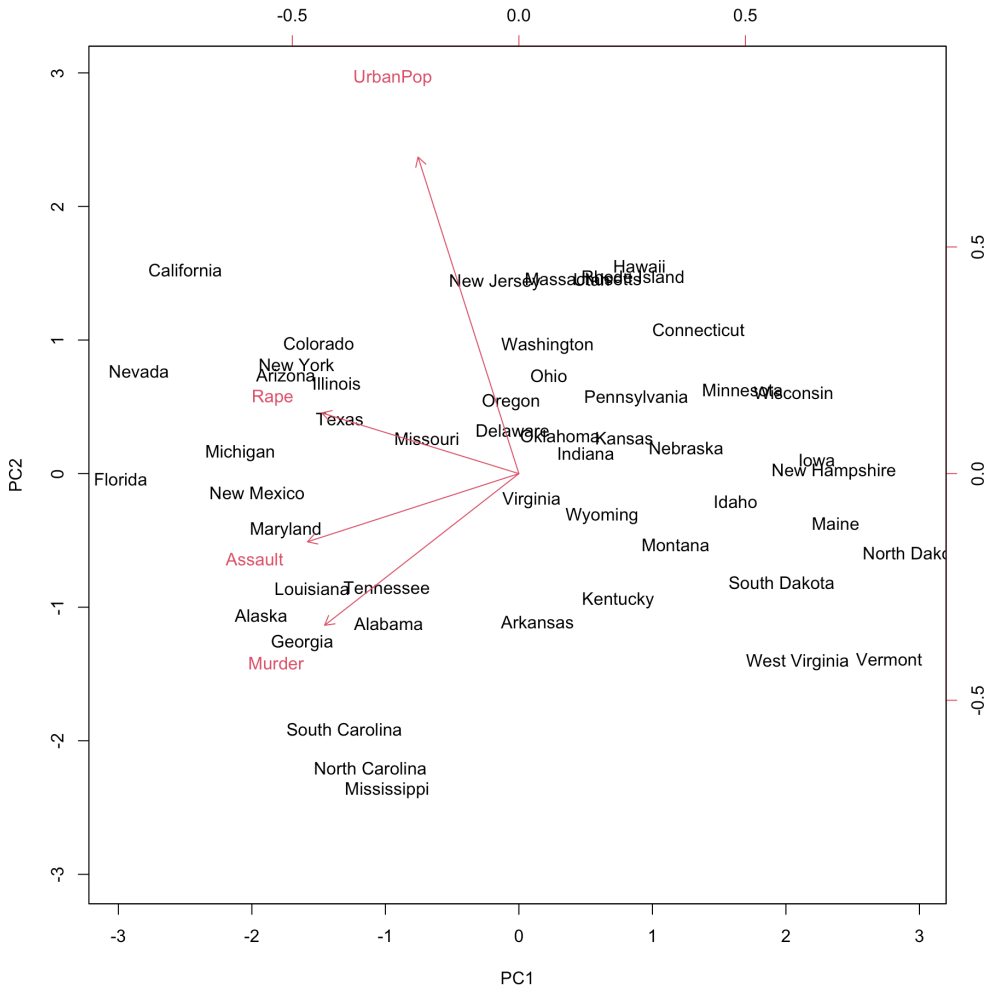
Principal Components Analysis (PCA)

- Find linear combinations of variables that maximise the variability

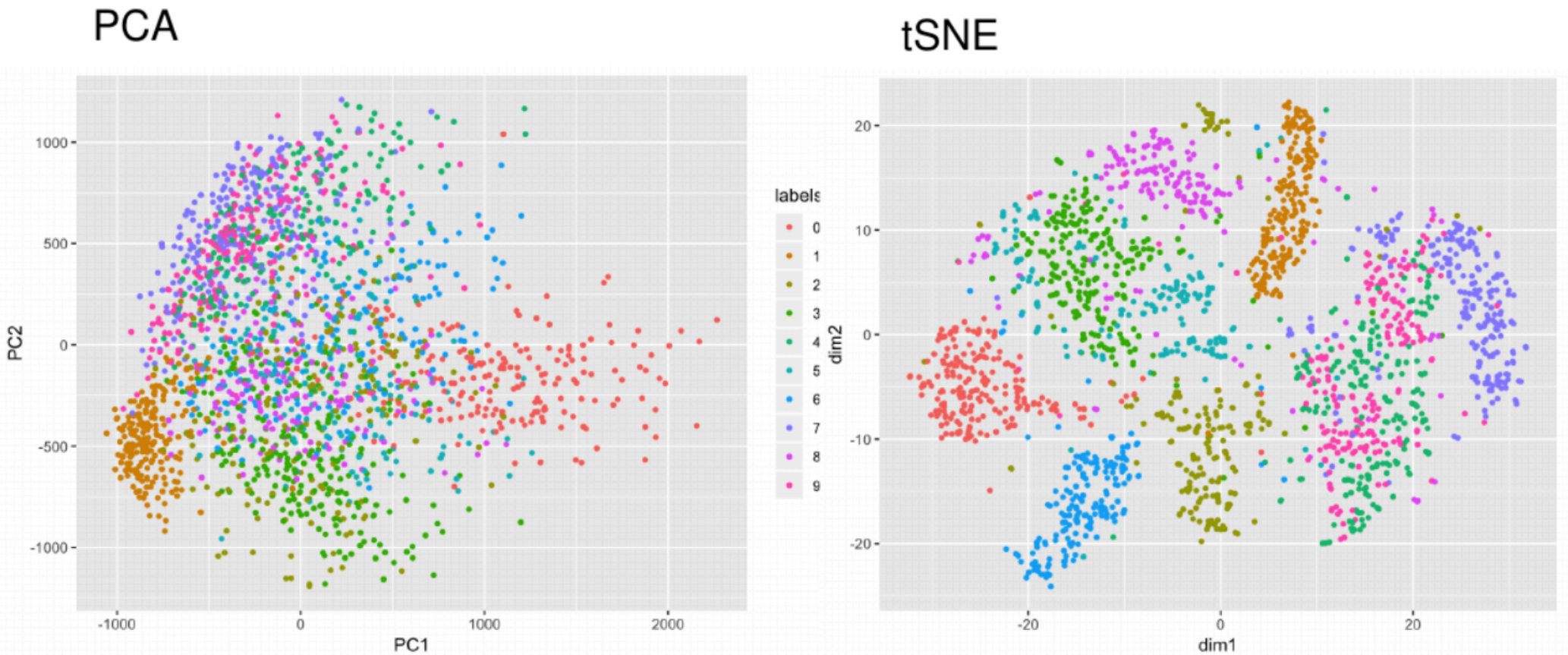




PCA


-0.5
0.0
0.5
UrbanPop
California
New Jersey
Hawaii
Connecticut
Colorado
Washington
Nevada
New York
Arizona
Illinois
Ohio
Rape
Oregon
Pennsylvania
Minnesota
Wisconsin
Texas
Missouri
Delaware
Oklahoma
Kansas
Nebraska
Michigan
Indiana
Iowa
New Hampshire
Florida
New Mexico
Virginia
Idaho
Wyoming
Maine
Maryland
Montana
Assault
South Dakota
Louisiana
Tennessee
Kentucky
Alaska
Alabama
Arkansas
Georgia
Murder
West Virginia
Vermont
South Carolina
North Carolina
Mississippi
PC2
PC1
3
2
1
0
-1
-2
-3
0.5
0.0
-0.5

PCA and t-SNE


PCA
PC2
PC1
1000
500
0
-500
-1000
-1000
0
1000
2000
tSNE
dim2
dim1
20
10
0
-10
-20
-20
0
20
labels
0
1
2
3
4
5
6
7
8
9

Clustering Methods

K-Means Clustering

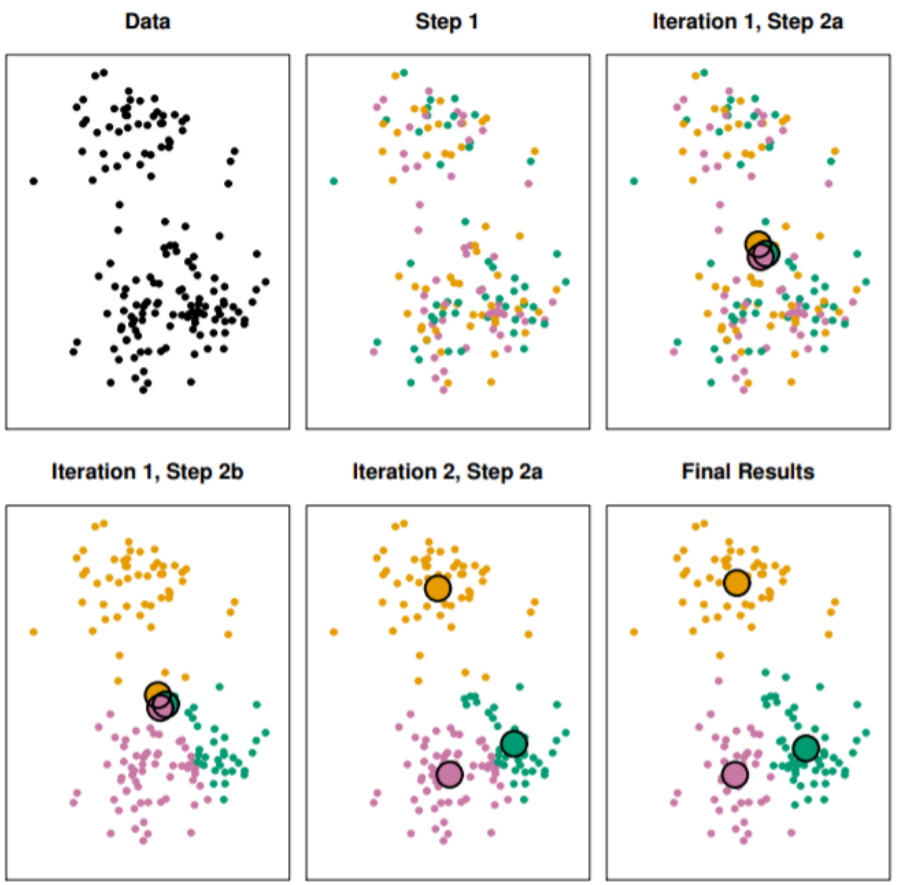
- Initialise each observation at random to a cluster.
- Iterate the following until convergence.
    1. Find cluster means with cluster memberships fixed

$$\widehat{\overline{x}}_j = \text{argmin}_m \sum_{i:\text{cluster}(i)=j} ||x_i - m||^2$$

    2. Find cluster memberships with cluster means fixed

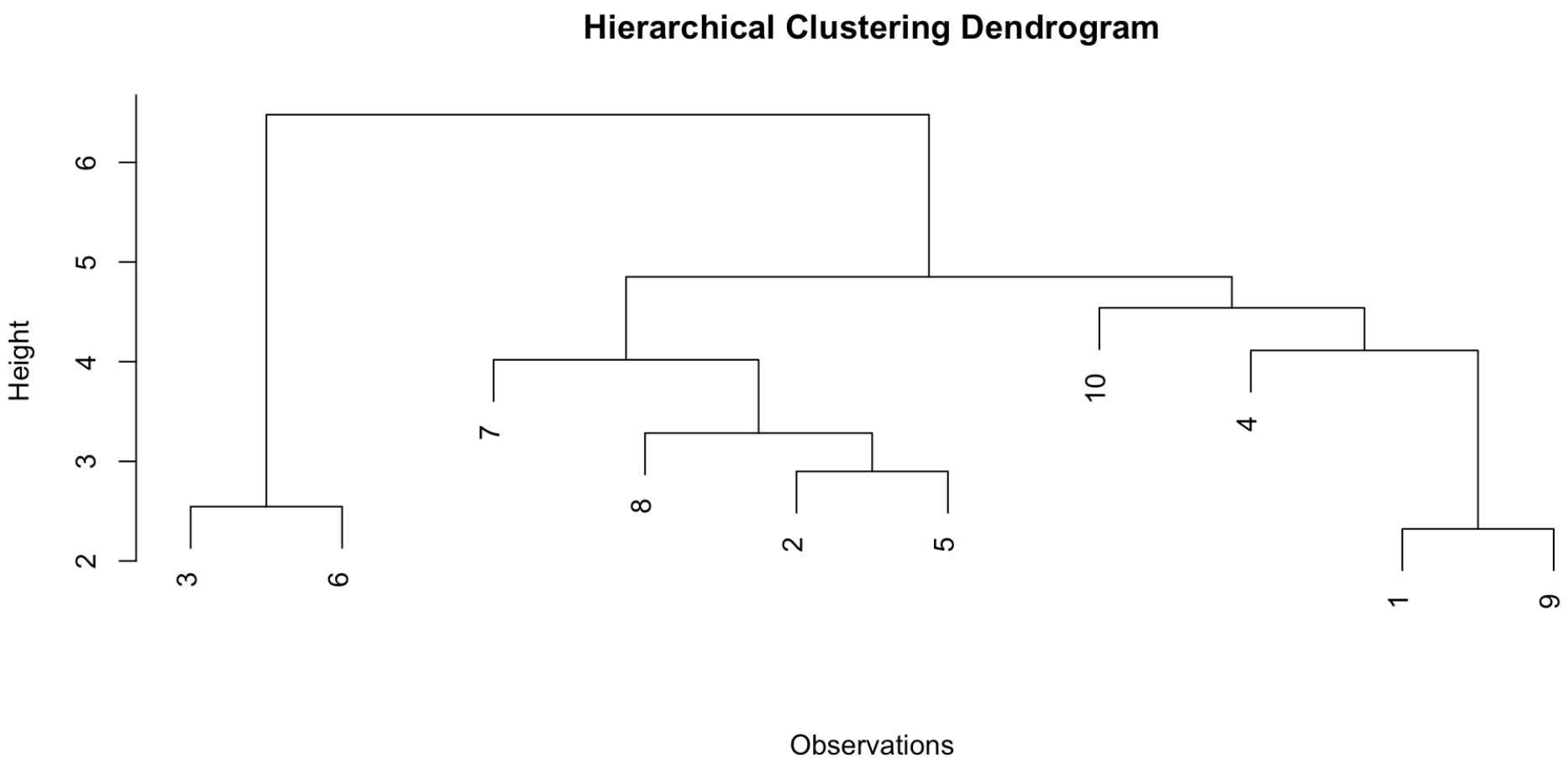
$$\text{cluster}(i) = \text{argmin}_k ||x_i - \widehat{\overline{x}}_k||^2$$

- $||\cdot||$ is some norm and $\text{cluster}(i)$ denotes which cluster x_i belongs to



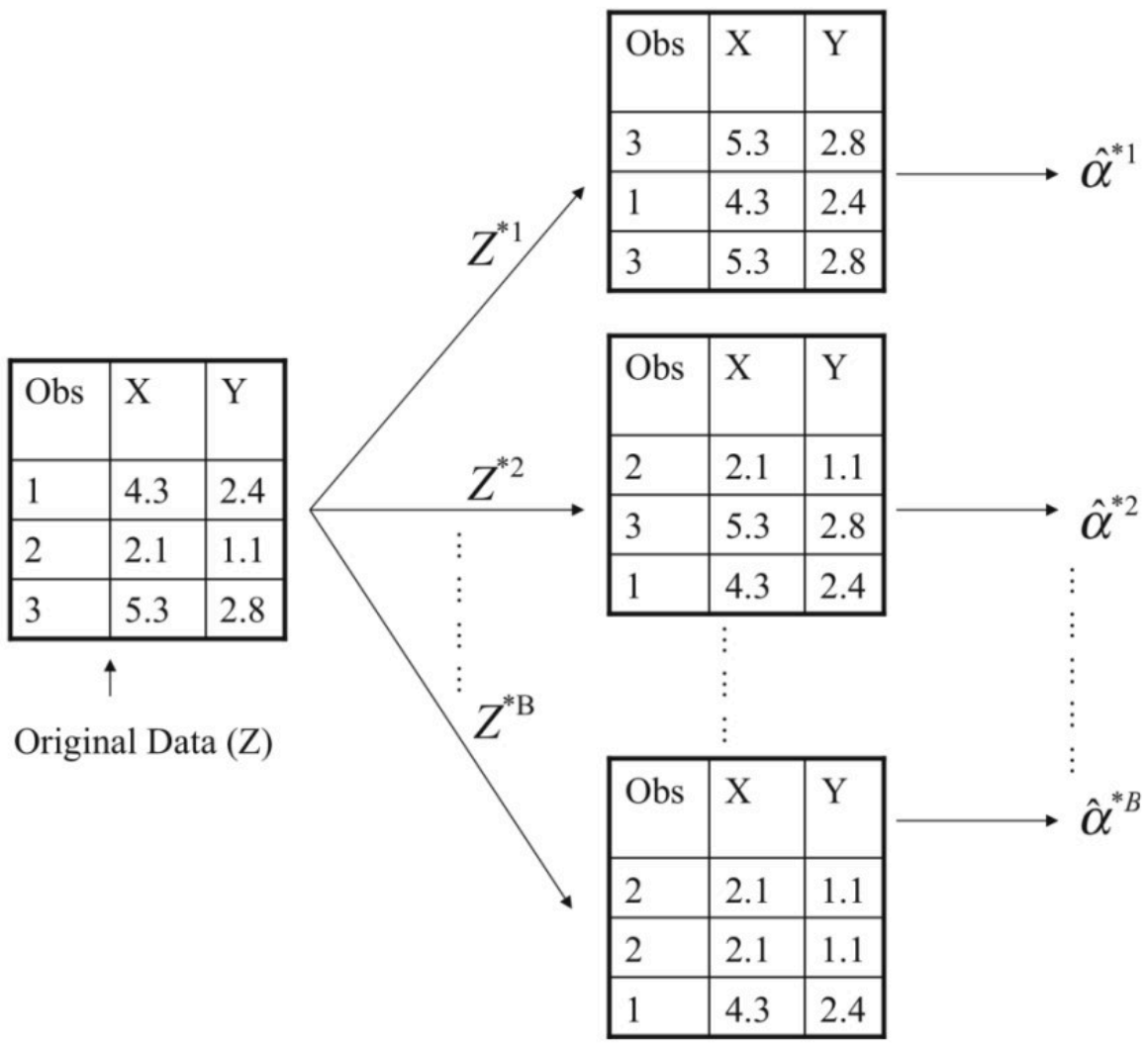
Clustering Methods

Hierarchical Clustering



Bootstrap

- Simulate related data (sampling with replacement) and examine statistical performance on all the re-sampled data.



Missing Data

- Remove missing data (complete cases)
- Single Imputation
- Multiple imputation
- Expert knowledge of reasons for missing data

Monte Carlo Methods

- Repeated simulation to estimate the full distribution and summary values
  - ➟ Assume $X \sim f$

$$\mathbb{E}[g(X)] = \int g(t) f(t)\, dt \approx \frac{1}{N} \sum_{i=1}^{N} g(X_i)$$

- Exploit law of large numbers
- Can sample from f if inverse of $F(x)$ exists
  - ➟ Can generate $X \sim f$ as: $X = F^{-1}(U)$
- Acceptance rejection method to handle more difficult distributions

Markov Chain Monte Carlo

- Big use in modelling Bayesian methods.
- Simulate a stochastic process (random variable that changes over time).
- Simulate new point based on the current point.
- Can estimate even more complex distributions that in Monte Carlo methods.

Local regression (smoothing)

A typical model in this case is

$$Y_i = f(x_i) + \varepsilon_i$$

* The function f is some smooth function (differentiable)

Density estimation

- Maximum Likelihood approach

$$f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta})$$

- For iid data, reformulate as

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \rightsquigarrow \ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log L(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$$

Kernel density estimation

- A kernel is a special type of probability density function (PDF) having the properties.
  - ➠ non-negative $K(x) \geq 0$, symmetric $K(-x) = K(x)$, unit measure $\int K(x)\,dx = 1$
- Smooth the data with a chosen hyperparameter (bandwidth) to estimate the density

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Example of Multiple Choice Questions

Example Multiple Choice Questions

3. Which of the following are supervised learning techniques?

   A. K-means clusters

   B. Random Forest

   C. Linear Discriminant Analysis

   D. Density estimation

4. Which of the following are characteristics of a kernel function (as used in density estimation)?

   A. a frequency function from a histogram

   B. a symmetric function

   C. a function ranging from -1 to 1

   D. a function that integrates to 1 over its support

Example Multiple Choice Questions

5. Which of the following practices may overestimate the test performance?

   A. Using PCA to construct new independent features from the original features. B. Imputing missing values using the mean calculated from the entire dataset

   C. Using 10-fold cross-validation to assess model performance. D. To address class imbalance, reporting Cohen's kappa from the test set as the overall performance metric.

Example Multiple Choice Questions

6. Which of the following statements about the support vector machine are correct?

A. SVM aims to find the hyperplane that maximises the margin between different classes.

B. SVMs can only be used for linear classification problems.

C. Changes in the position of the support vectors will not impact the decision boundary.

D. Increasing the value of C in the SVM's optimisation function will lead to an increase in the bias but the model will generalises better to the unseen data.

Example Multiple Choice Questions

7. Which of the following are indirect measures of the test error?
    a. $C_p = \frac{1}{n}\left(\text{RSS} + 2d\widehat{\sigma}^2\right)$
    b. $\text{RSS} = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$
    c. $\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\widehat{\sigma}^2\right)$
    d. $F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

where in the above:

$\widehat{Y_i}$ is the predicted response for the ith observation; d is the number of features in the model, not including the intercept;

Example of Extended Answer Questions

Your friend recently started an internship as a data analyst at a university student support unit. The team is interested in building a model to predict whether a student is at risk of dropping out, using available academic and engagement data collected from the learning management system.

Your friend explains the dataset consists of 5,000 observations, where each data point is represented as $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, 5000$. $\mathbf{x}_i$ contains 80 features, including number of logins, average time spent per week on the platform, assignment grades, and forum participation. $y_i = 1$ means the student eventually dropped out, while $y_i = 0$ means the student successfully completed the semester. Around 2% of students in the dataset dropped out.

Here's the modeling workflow your friend followed:

1. They noticed a few missing values in some features and filled them using the mean of each variable.
2. Then, they applied feature selection on the imputed full dataset, selecting the top 20 features most correlated with the target variable.
3. Next, they randomly split the data into 75% training and 25% testing sets.
4. Finally, they trained a SVM classifier and evaluated it using test set accuracy, achieving 92% accuracy.

Based on your understanding of statistical learning and good modeling practice, identify **three** problematic aspects of your friend's modeling workflow. For each issue: briefly explain why it is problematic, and suggest an alternative.

Example of Extended Answer Questions

Retirement problem

> You are planning your retirement and decide that you will retire with $1,000,000 invested in an index fund. During retirement you plan to withdraw $50,000 each year from your investment with the remaining money being invested in an index fund. Assume the index fund has an average return rate of 9% and a standard deviation of 15% (normally distributed). Assume you retire at 65 and will live until you are 100, and the CPI adjustment is 104% each year. Compute the probability that your investment will support your lifestyle until you die.

Your friend uses Monte Carlo to study your retirement plan and proposes the pseudo code to solve this problem. Evaluate this code and fill in <1> to <4>.

Pseudo Code

```
# Set initial parameters
initial_investment <- 1000000
annual_withdrawal <- 50000
mean_return <- 0.09
sd_return <- 0.15
cpi <- 1.04
years <- 35
n_sim <- <1>
success_count <- 0

# Start Monte Carlo simulation
for (sim in 1:n_sim) {
  investment <- initial_investment
  withdrawal <- annual_withdrawal

  for (year in 1:years) {
    # Simulate annual return from normal distribution
    annual_return <- random value from <2>

    # Update investment value
    investment <- investment * (1 + annual_return)
    investment <- investment - withdrawal
```

Write down an expression for <1> to <4> respectively.

Consultation Hours

- To be announced via Ed forum soon!
- Good luck (not that you need it :)