STAT5003

Week 13 : Review and Final Exam

Dr. Justin Wishart

THE UNIVERSITY OF SYDNEY

Review

Review: Methods we have learnt

- Regression
  - Multiple linear regression
  - Univariate smoothing (nonlinear) regression
- Clustering and higher dimensional viz
  - Hierarchical clustering
  - K-means clustering
  - PCA
  - t-SNE
  - MDS
- Classification
  - Logistic regression
  - LDA
  - kNN
  - SVM
  - Random forests
  - Boosting trees.

Multiple 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

Local regression (smoothing)

A typical model in this case is

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

- The function f is some smooth function (differentiable.)

Density estimation

- Maximum Likelihood approach

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

- Reformulate as

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

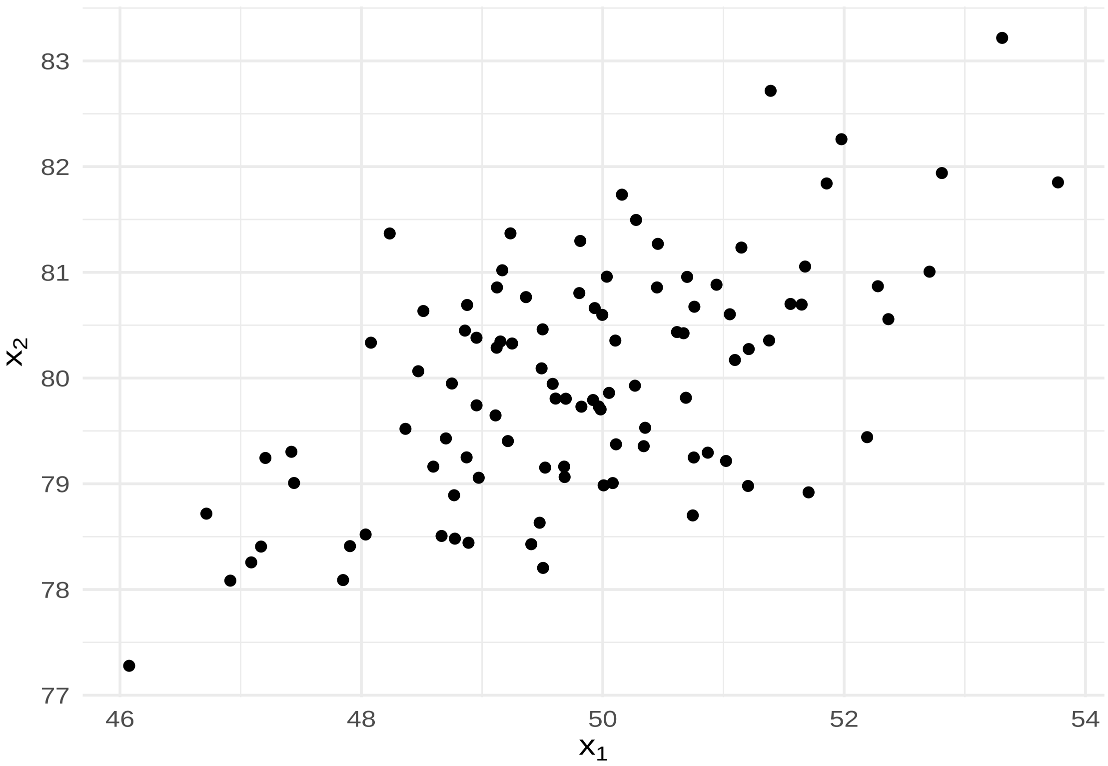
Kernel density estimation

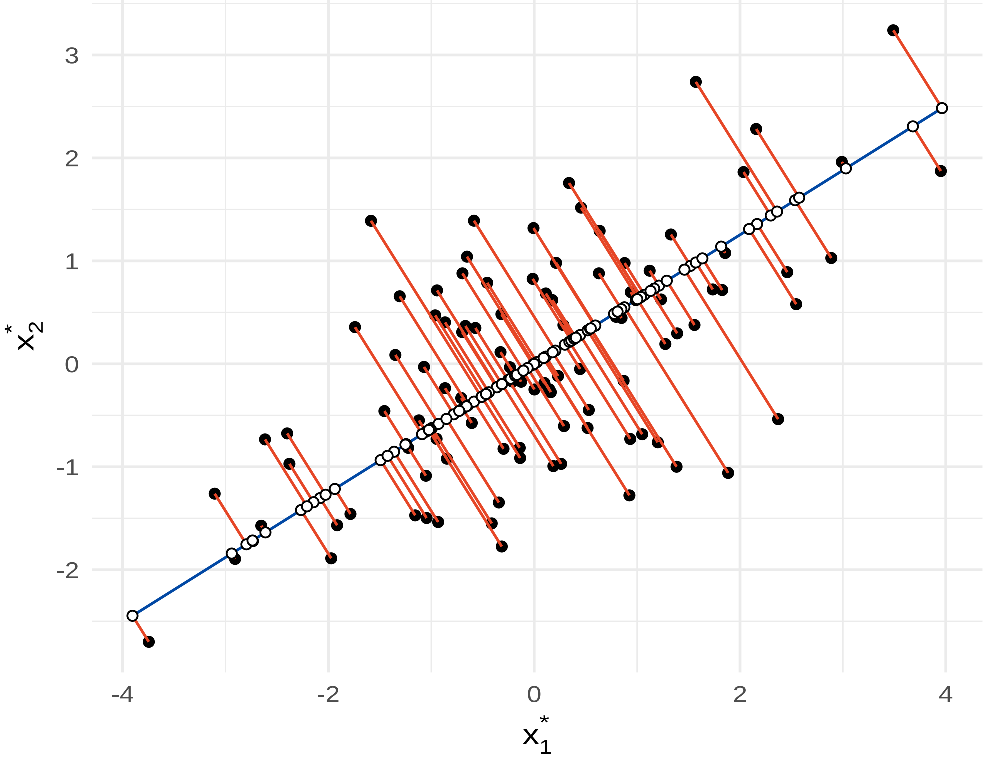
- Smooths the data with a chosen hyperparameter (bandwidth) to estimate the density.

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

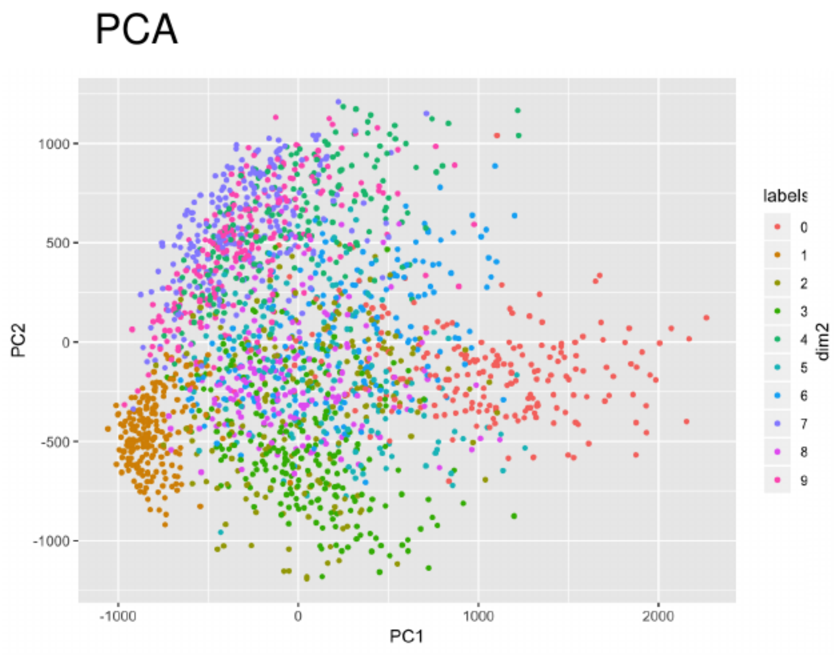
Principal Components Analysis (PCA)

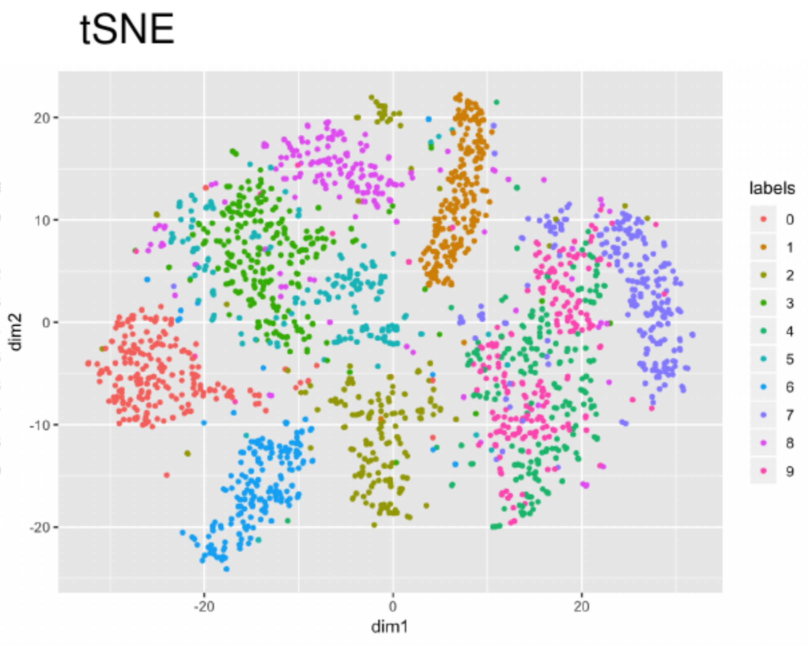
- Find linear combinations of variables that maximise the variability





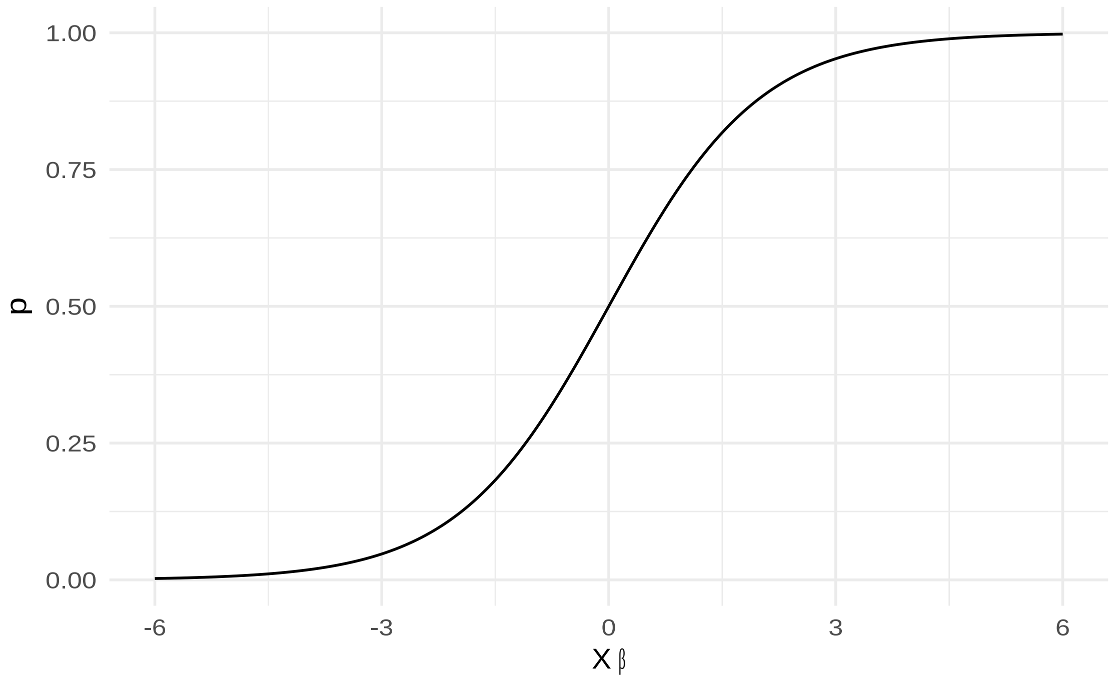
PCA and t-SNE


PCA
1000
500
0
-500
-1000
PC2
-1000
0
1000
2000
PC1
labels
0
1
2
3
4
5
6
7
8
9


tSNE
20
10
0
-10
-20
dim2
-20
0
20
dim1
labels
0
1
2
3
4
5
6
7
8
9

Logistic Regression model

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



Linear Discriminant Analysis (LDA)

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

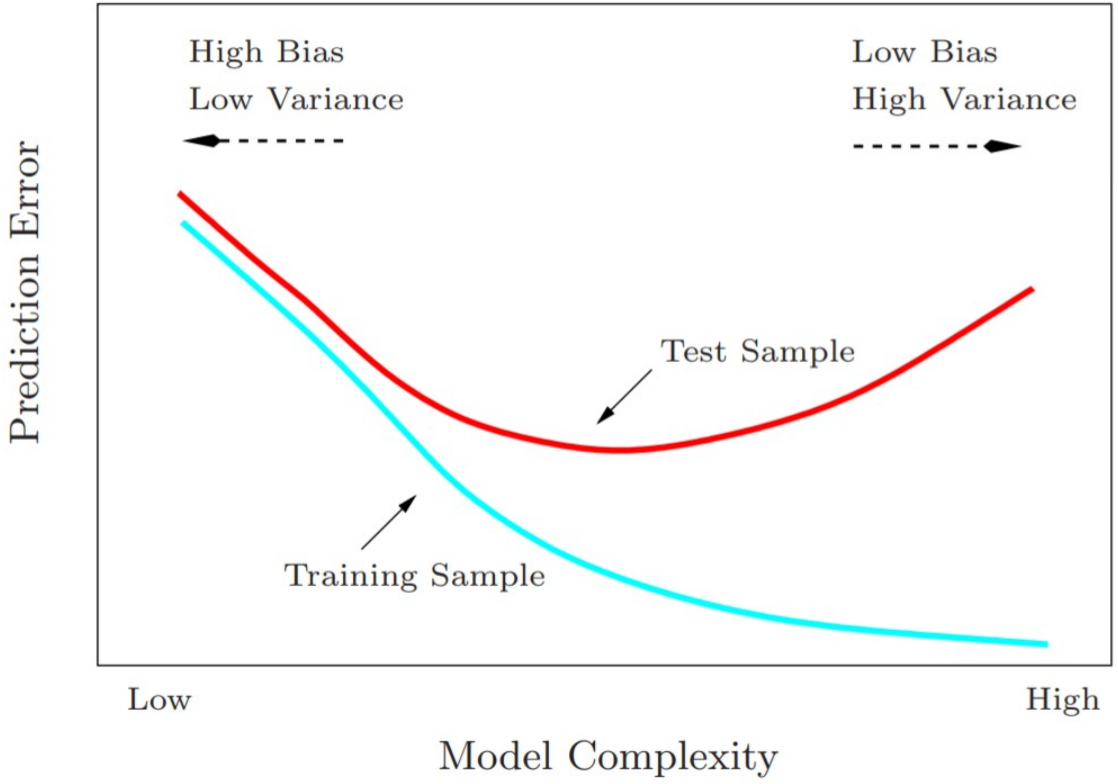
Posterior: The probability of classifying observation to group k given it has features x

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

- $f_k(x) = P(X = x|Y = k)$ is the density function for feature x given it's in group k

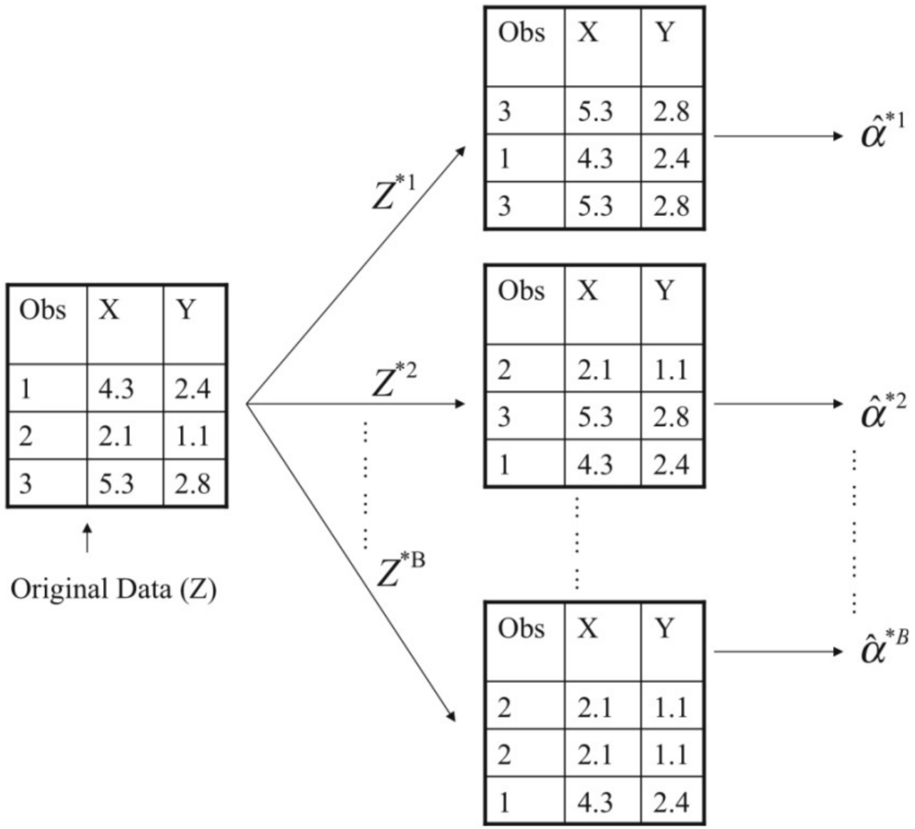
Cross validation

- 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.



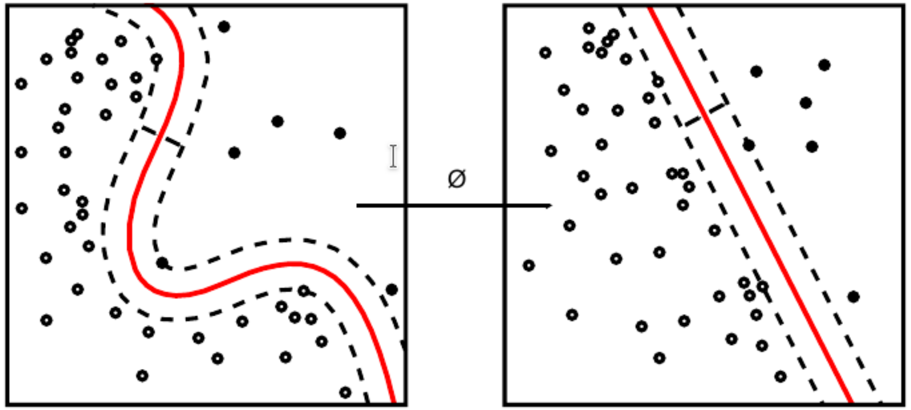
Bootstrap

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



Support Vector Machines (SVM)

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

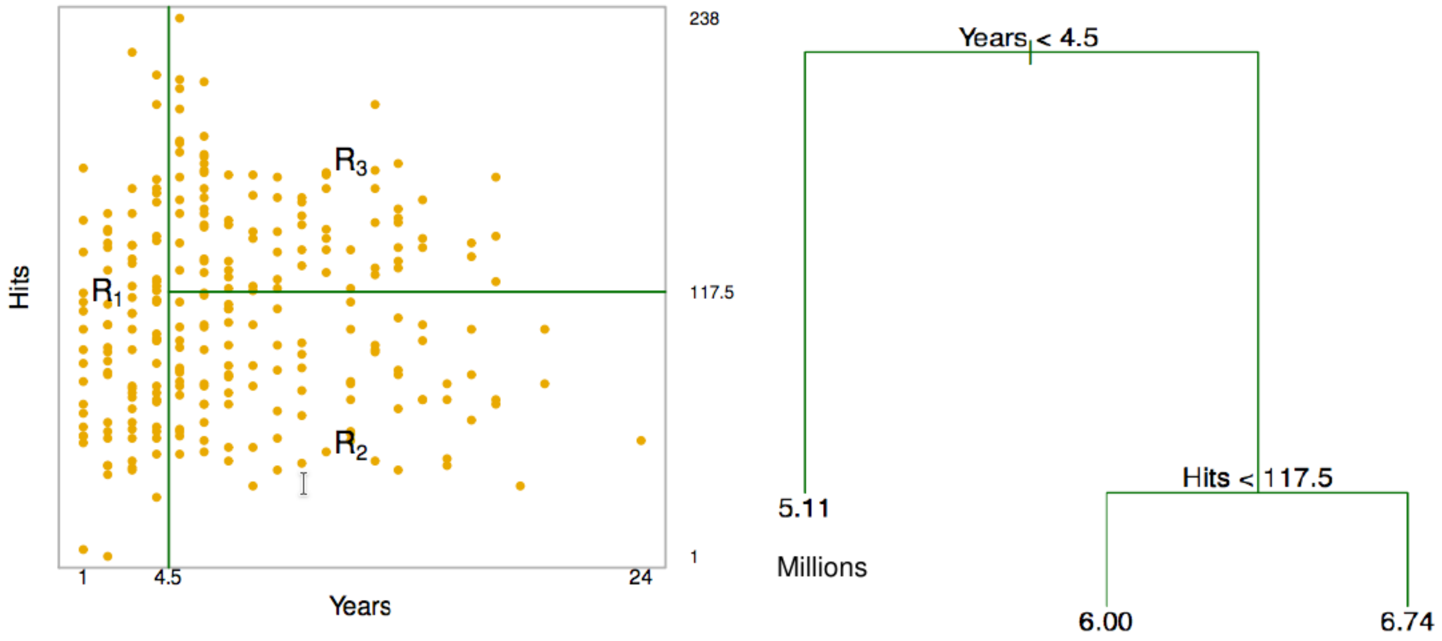


Missing Data

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

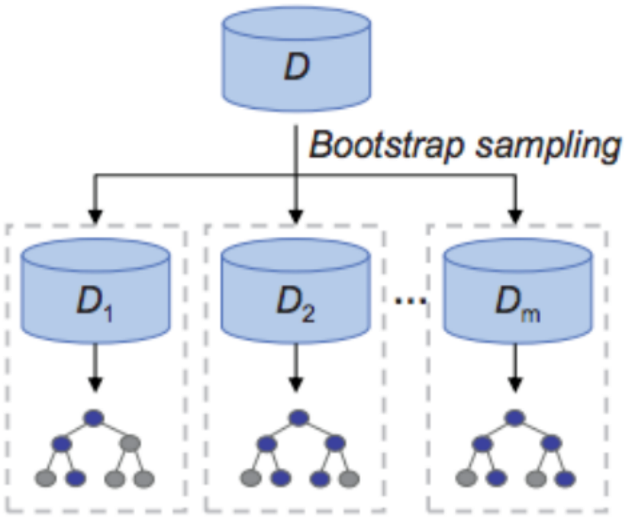
Basic decision trees

- 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.
- $\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}_{*b}(x)$
- Random forests do further subsampling of predictors at split to improve 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

Feature Selection

- Best subset selection.
- Forward selection.
- Backward selection.
- Choose model that minimises test error
  - Directly via test set
  - Indirectly via penalised criterion.

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.

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

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

Monte Carlo Methods

- Repeated simulation to estimate the full distribution and summary values.

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

- Exploits 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.
- Simulates a process (random variable that changes over time)
- Simulate new point based off the current point.
- Can estimate even more complex distributions that in Monte Carlo methods.

Methods and metrics to evaluate models

- Sensitivity and specificity
- Accuracy
- Residual sum of squares (for regression)
- ROC curves and AUC
- K-fold cross-validation

Exam Format

Exam format

- Two hour written exam (conducted online)
- 9 Multiple choice questions
  - Questions have **two correct answers**.
  - You need to select the correct answer(s) to get a mark.
- Some short answer question.
  - Some data context given
  - Some individual short answer subquestions given on each data context.
- Two longer answer questions.

Topics covered

- Everything in the lectures/labs from weeks 1 to 11
    - Except any topic that was marked as not examinable
- Writing code is not tested.
    - There will be some questions on interpretting R outputs.
- You should understand how the algorithms work and be able to sketch out the key steps in pseudo code.

Example multiple choice question

Which of the following method(s) is/are unsupervised learning methods?

A. K-means clustering B. Logistic regression C. Random forest D. Support vector machines

Example short answer question

a. Explain how the parameters are estimated in simple least squares regression.

b. Explain a scenario where simple linear regression is not appropriate.

c. Compute the predicted weight for a person that is 160cm tall and compute the residual of the first person in the table below.

$$\widehat{Y} = 50.412 + 0.0634X$$

Sample	X: Height (cm)	Y: Weight (kg)
1	160.0	60
2	170.2	77
3	172.0	62

Example long answer question

- Describe the Markov Chain Monte Carlo procedure. You may use pseudo code as part of your answer.