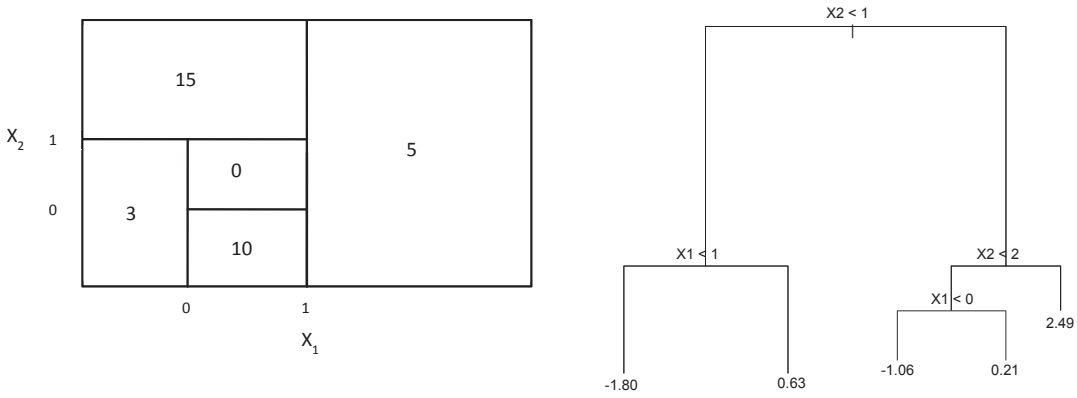
ST464/ST674
Assignment 4
Katarina Domijan
Due on Thursday 2nd May 1pm

- Do all questions. Submit questions: 1, 2, 4 and 5.
- Use R markdown to knit your results to a .pdf or .html file. Print and submit to your tutor's box.
- Place your name and student number under author in the YAML header. E.g.

```
---
title: "Assignment 4"
output: pdf_document
author: Jane Doe 1234567
---
```

- There will be a tutorial on Monday 29th April.

1. For the Boston data available in package MASS we wish to relate dis (weighted mean of distances to five Boston employment centres) to nox (nitrogen oxides concentration in parts per 10 million).
   (a) Fit a cubic polynomial to the data. Plot the data and the fit. Comment on the fit. Calculate the MSE.
   (b) Repeat (a), this time using a 10th degree polynomial. Compare the fits and the MSE. Use anova to compare the two fits and comment on your findings.
   (c) Describe how you might use cross-validation to select the optimal degree (say between 1 and 10).
   (d) Carry out the cross-validation procedure. What is the optimal degree?
   (e) Use bs() to fit a regression spline with 4 degrees of freedom. What are the knots used? Plot the data and the fit. Comment on the fit. Calculate the MSE.
   (f) Fit a curve using a smoothing spline with the automatically chosen amount of smoothing. Display the fit. Does the automatic λ give a good result?
   (g) Now use smoothing spline with a larger value of spar. Overlay both smoothing spline fits on the plot. Which looks better?
2. Using the Boston data, with dis as the response and predictors medv, age and nox.
   (a) Split the data into training 60% and test 40%. Using the training data, fit a generalised additive model (GAM). Use ns with 4 degrees of freedom for each predictor.
   (b) Use plot.gam to display the results. Does it appear if a linear term is appropriate for any of the predictors?
   (c) Simplify the model fit in part (a). Refit the model. Use anova to compare the two fits and comment on your results.
3. (a) Sketch the tree corresponding to the partition shown below. The numbers in the rectangles are the mean response for predictors in that rectangle.
   (b) For the tree shown on the right, draw the corresponding partition of predictor space. Indicate the mean for each region.



4. (a) For the training data in question 2, fit a tree model. Use dis as response, and predictors medv, age and nox. Draw the tree. Calculate the training and test MSE.

   (b) Use cv.tree to select a pruned tree. If pruning is required, fit and draw the pruned tree. Calculate the training and test MSE. Compare the results to those in (a).

   (c) Which fit is better, the (optionally pruned) tree or the GAM? Compare their performance on the test data.

5. For the data generated in question 6, Assignment 3:

```
set.seed(1)
```

```
x <- rnorm(100)
y <- 1 + .2*x+3*x^2+.6*x^3 + rnorm(100)
d <- data.frame(x=x,y=y)
```

   (a) Fit a regression model containing predictors X, X^2, $\ldots X^{10}$. Based on the output in summary() which terms are needed in the model?

   (b) Fit a ridge regression model using the glmnet function over a grid of values for λ ranging from 0.001 to 50. Plot coefficients vs penalty using the default plot method. Use the inbuilt function cv.glmnet to choose the tuning parameter λ . How do the coefficients at the optimal value of λ compare to the linear regression ones in (a)?

   (c) Repeat (b) for lasso regression instead of ridge.

   (d) Plot the data y vs x and superimpose the fitted models from linear regression, ridge and lasso with optimal values of lambda as chosen by cross-validation.

6. Titanic data from Assignment 3:

```
ttrain <- read.csv("yourfolder/ttrain.csv", header=T, row.names=1)
ttest <- read.csv("yourfolder/ttest.csv", header=T, row.names=1)
head(ttrain)
```

(a) For the training data, fit a tree model using all three predictors. Draw the tree. Interpret the model. For the training and test data what proportion of survivors are missclassified? What proportion of those who died are missclassified? What proportion of the predicted survivors actually survived? What is the overall error rate for the training data?

(b) Use cv.tree to select a pruned tree. If pruning is required, fit and draw the pruned tree.

(c) Fit a tree model using only Age and Class as predictors. Draw the tree. Interpret the model. Compare the test set results to (a).

(d) Fit a random forest model (using randomForest) using all three predictors and compare the test set results to (a) and (c). Which variables are important?

7. Heart data: binary outcome AHD for 303 patients who presented with chest pain. An outcome value of Yes indicates the presence of heart disease, while No means no heart disease.

   There are 13 predictors including Age, Sex, Chol (a cholesterol measurement), and other heart and lung function measurements.

```
heart <- read.csv("yourfolder/heart.csv", row.names=1)
head(heart)
heart <- na.omit(heart)
set.seed(2)
s <- sample(nrow(heart), 200)
heartTrain <- heart[s,]
heartTest <- heart[-s,]
```

   Fit a support vector machine with a radial kernel to this data. Use cross validation to tune the γ and cost parameters (see function tune() in e1071 library). How does your result (test error) compare to the test error in the notes (obtained using trees and random forrests)?