

## Instructions

There are 8 questions worth a total of 100 marks. You should attempt them all.

### QUESTION 1

- (a) True or false. The training error of the K Nearest neighbours (KNN) classifier with  $K = 1$  is zero. Explain your answer.

[2 marks]

True. Each point is its own neighbor, so 1-NN classifier achieves perfect classification on training data.

- (b) True or false. The depth of an estimated decision tree can be larger than the number of training examples used to create the tree. Explain your answer.

[2 marks]

False. Each split of the tree must correspond to at least one training example, therefore, if there are  $n$  training examples, a path in the tree can have length at most  $n$

- (c) Give one similarity and one difference between feature selection and principal component analysis (PCA).

[2 marks]

They both allow to reduce the dimensionality (similarity). Feature selection select variables while PCA is a feature extraction method which produces new variables (difference).

- (d) Briefly explain why minimizing the training mean squared error (MSE) can lead to overfitting. Explain why adding a penalty term allows us to control the bias and variance tradeoff of our error estimation.

[2 marks]

The training MSE is a biased estimate of the expected test MSE since we use the training data both for fitting and validation. Increasing flexibility will decrease the train MSE, while the expected test MSE will increase if the model is too flexible (high variance). Adding a penalty term allows to penalize the complexity of the model, and by selecting the right value of the tuning parameter, we can control model fit and model complexity, or equivalently bias and variance.

- (e) Suppose you apply clustering on a data set with 5 observations, explain if K-means or hierarchical clustering can produce the following three clusters:  $C_1 = \{1, 3, 4\}$ ,  $C_2 = \{2, 4\}$  and  $C_3 = \{5\}$ .

[2 marks]

No. Both K-means and hierarchical clustering cannot produce overlapping/fuzzy clusters.

- (f) True or false. It is possible to produce a nonlinear regression fit using linear regression. Explain your answer.

[2 marks]

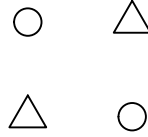
TRUE. We can for example produce nonlinear transformation of the predictors, and fit a linear regression model.

[Total: 12 marks]

— END OF QUESTION 1 —

## QUESTION 2

(a) Consider the following dataset:



Which classifiers will achieve zero training error on this data set? (i) logistic regression, (ii) SVM (with polynomial kernel of degree 2), (iii) 3-NN classifier.

[2 marks]

(ii) will achieve zero training error since the decision boundary will be an ellipse. (i) is linear and the majority vote of (iii) will not work.

(b) Consider the following classification problem. We first choose the class  $Y \sim \text{Bernoulli}(\frac{1}{2})$ , which is class 1 with probability  $\frac{1}{2}$ . If  $Y = 1$ , then  $X \sim \text{Bernoulli}(p)$ ; otherwise  $X \sim \text{Bernoulli}(q)$ . Assume that  $p > q$ .

(i) What is the Bayes optimal classifier  $C^*(X)$ ?

[3 marks]

$$C^*(x) = \arg \max_y P(Y = y|X = x) \quad (1)$$

$$= \arg \max_y \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \quad (2)$$

$$= \arg \max_y P(X = x|Y = y)P(Y = y) \quad (3)$$

$$= \arg \max_y P(X = x|Y = y) \quad (4)$$

- $C^*(1) = \arg \max_y P(X = 1|Y = y) = 1$  since  $p = P(X = 1|Y = 1) > P(X = 1|Y = 0) = q$ .
- $C^*(0) = \arg \max_y P(X = 0|Y = y) = 0$  since  $1 - p = P(X = 0|Y = 1) < P(X = 0|Y = 0) = 1 - q$ .

(ii) What is the Bayes error rate, i.e.  $P(Y \neq C^*(X))$ ?

[2 marks]

$$P(Y \neq C^*(X)) = P(Y \neq X) \quad (5)$$

$$= P(Y = 1)P(X = 0|Y = 1) + P(Y = 0)P(X = 1|Y = 0) \quad (6)$$

$$= \frac{1}{2}(1 - p) + \frac{1}{2}q \quad (7)$$

- (c) True or false. The Bayes error rate is the lowest possible error rate, and can never be zero. Explain your answer.

[2 marks]

FALSE. Although the Bayes error rate is the lowest possible error rate, its value can be zero when the classes are perfectly separable.

- (d) True or false. Let  $Y$  be a binary response and  $P(X) = \text{Probability}(Y = 1|X)$ . If  $P(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$ , the log-odds are nonlinear in  $X$ . Prove your answer.

[2 marks]

FALSE.  $\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X$

- (e) In the quadratic discriminant analysis (QDA) model, the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the case where there is only one feature, and  $K$  classes. Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

[3 marks]

The Bayes' classifier maximizes the posterior conditional probabilities. In the QDA model, the discriminant functions are given by  $\delta_k(x) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) - \frac{1}{2\sigma_k^2}(x - \mu_k)^2$ , which are quadratic in  $x$ . (We have seen in class that  $\delta_k(x)$  is linear in  $x$  when  $\sigma_1 = \dots = \sigma_K = \sigma$ ).

[Total: 14 marks]

— END OF QUESTION 2 —

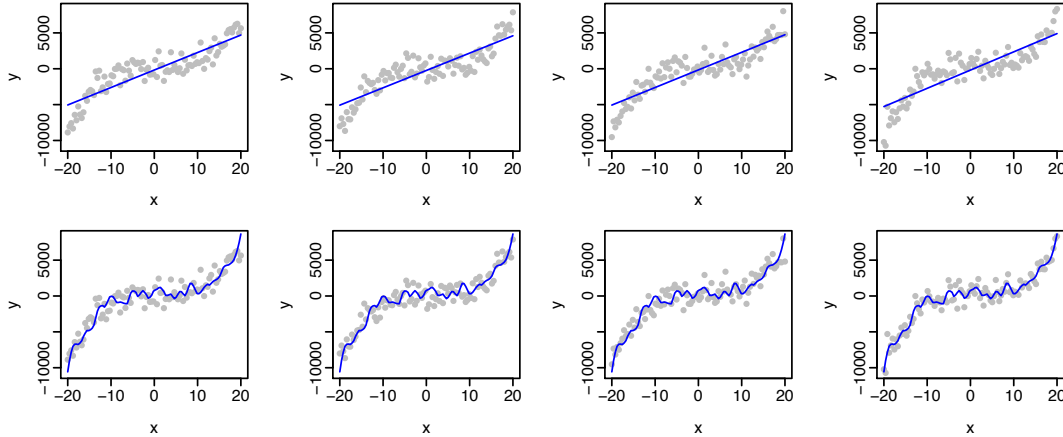
### QUESTION 3

Consider the following data generating process:

$$y = x^3 - 2x^2 + 1.5x + \varepsilon, \quad (8)$$

where  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  and  $\sigma = 1000$ .

- (a) The following Figure shows four samples with 100 observations drawn i.i.d. from (8), together with mean regression fitted values for two different models (in each row).



Describe the difference between the two models in terms of model complexity.

[2 marks]

We expect the students to talk about underfitting (left) and overfitting (right).

- (b) If we want to minimize expected out-of-sample squared errors, what is the optimal prediction  $\hat{y}_*$  for  $x_* = 10$ ? Explain your answer.

[2 marks]

The optimal prediction is given by the conditional mean:  $\hat{y}_* = 815$

- (c) What is the expected out-of-sample squared errors of the optimal prediction at  $x_* = 10$ ?

[2 marks]

$\sigma^2$ . Note that it does not depend on  $x$  since the variance of the noise is constant.

- (d) True or false.  $\hat{f}(x) = x^3$  will have minimum expected out-of-sample squared errors for all  $x$ . Explain your answer.

[2 marks]

Students should recognise that the optimal fit is given by  $f(x) = x^3 - 2x^2 + 1.5x$ , which has minimum expected out-of-sample squared errors.

- (e) Suppose we have generated a training sample of  $n$  pairs  $\mathbf{x}_i, y_i$  drawn i.i.d. from (8), where  $\mathbf{x}_i = (x_i, x_i^2, x_i^3)$ . We also generate a test sample of  $n$  pairs  $\mathbf{x}_i, y'_i$  drawn i.i.d. from  $y'_i = x_i^3 - 2x_i^2 + 1.5x_i + \varepsilon'_i$  where  $\varepsilon'_i$  and  $\varepsilon_i$  are independent but identically distributed. Note that the training and test sample have the same predictors but different responses. Let  $\hat{y}_i$  be the fitted values from linear regression (with intercept) applied on the training sample  $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ . We know that

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (y'_i - \hat{y}_i)^2 \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right] + \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{y}_i, y_i).$$

In our scenario, if  $n = 1000$ , prove step by step that

$$\frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{y}_i, y_i) = 8000.$$

[3 marks]

$$\frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{y}_i, y_i) = \frac{2}{n} \text{Trace}(\text{Cov}(\hat{y}, y)) = \frac{2}{n} \text{Trace}(E[\hat{y}y'] - E[\hat{y}]E[y']) = \frac{2}{n} \text{Trace}(E[\hat{y}y']) \quad (9)$$

$$= \frac{2}{n} \text{Trace}(E[\hat{y}y']) = \frac{2}{n} \sigma^2 \text{Trace}(H) = \frac{2}{n} \sigma^2 (p + 1) \quad (10)$$

The result follows from the fact that  $n = 1000$ ,  $p = 3$  and  $\sigma = 1000$ .

[Total: 11 marks]

— END OF QUESTION 3 —

#### QUESTION 4

Let  $\mathbf{y} = (y_1, \dots, y_n)' \in \mathbb{R}^n$ ,  $\mathbf{X}' = [\mathbf{x}_1 \ \dots \ \mathbf{x}_n] \in \mathbb{R}^{p \times n}$  with  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})' \in \mathbb{R}^p$ , and consider the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)' \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I})$ . Given a realization of  $\mathbf{y}$  and  $\mathbf{X}$ , we would like to estimate the coefficients  $\boldsymbol{\beta}$ .

- (a) When the number of predictors  $p$  is large, briefly explain why best subset selection is not computationally feasible.

[1 marks]

The number of possible combinations of the predictors increases exponentially with  $p$  + with a large number of combinations, we are more likely to overfit the data

Suppose we estimate the coefficients  $\boldsymbol{\beta}$  by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^p [(1 - \alpha)\beta_j^2 + \alpha|\beta_j|] \right\} \quad (11)$$

where  $\lambda \geq 0$ ,  $\alpha \in \{0, 1\}$  and  $\|\cdot\|_2$  is the  $L_2$  norm.

- (b) Assume  $\mathbf{X}$  is an  $n \times n$  identity matrix, and  $p = n$ .

- (i) Write the solution  $\hat{\boldsymbol{\beta}}$  when  $\lambda = 0$ .

[1 marks]

$$\hat{\boldsymbol{\beta}} = \mathbf{y}$$

- (ii) Calculate the bias and estimation variance of the estimator from (i).

[1 marks]

$$\operatorname{Bias}(\hat{\boldsymbol{\beta}}) = E[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} = \mathbf{0}, \text{ and } \operatorname{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 \mathbf{I}$$

- (iii) Write the solution  $\hat{\boldsymbol{\beta}}$  when  $\alpha = 0$ .

[1 marks]

$$\hat{\boldsymbol{\beta}} = \mathbf{y}/(1 + \lambda). \text{ A shrinkage is applied to all observations } \mathbf{y}.$$

- (iv) Calculate the bias and variance of the estimator from (iii).

[1 marks]

$$E[\hat{\boldsymbol{\beta}}] = \frac{\boldsymbol{\beta}}{1 + \lambda} \implies \operatorname{Bias}(\hat{\boldsymbol{\beta}}) = \frac{\boldsymbol{\beta}}{1 + \lambda} - \boldsymbol{\beta}, \text{ and } \operatorname{Var}[\hat{\boldsymbol{\beta}}] = \frac{\sigma^2}{(1 + \lambda)^2} \mathbf{I}$$

- (v) Compare the results of the bias and variance that you obtained in (ii) and (iv) and provide an explanation.

[1 marks]

The least squares estimator is unbiased. However, the ridge regression estimator (with tuning parameter  $\lambda \neq 0$ ) is biased but has lower variance, due to the bias-variance trade-off.

- (c) Briefly explain the difference between ridge regression and principal component regression (PCR). [2 marks]

PCR computes the first  $k \leq p$  principal components and apply OLS. In ridge regression, a shrinkage is applied to all principal components with a higher shrinkage for the components with smaller eigenvalues, i.e. smaller variance.

- (d) What does "sparse coefficients" mean? Which values of  $\alpha$  and  $\lambda$  provide "sparse coefficients"?

[1 marks]

Sparse coefficients means many coefficients are zero.  $\alpha = 1$  and  $\lambda \neq 0$

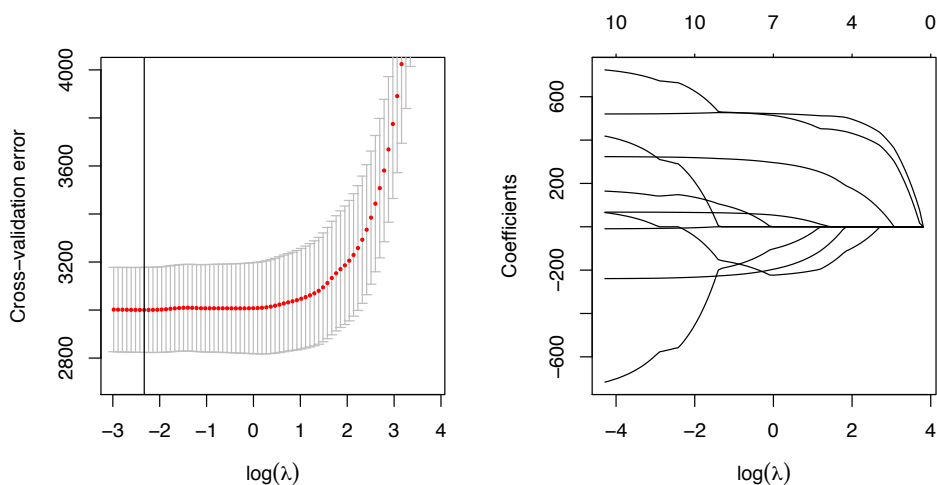
(e) In the Figure below, the left and right panels shows the cross-validation error and the regularization path as a function of the regularisation parameter  $\lambda$  (in log scale), respectively. The vertical line in the left panel shows the minimum value of the cross-validation error.

(i) Which value of the parameter  $\lambda$  will be selected if you apply the one-standard error rule? [2 marks]

$$\lambda \approx e^2$$

(ii) Which value of  $\alpha \in \{0, 1\}$  has been used to produce the plot in the right panel? Explain your answer. [2 marks]

$\alpha = 1$  since we produce sparse estimates.



[Total: 13 marks]

— END OF QUESTION 4 —

**QUESTION 5**

- (a) For the following dataset, which classifier has larger Leave-One-Out Cross validation error: (a) 1-NN classifier, (b) 3-NN classifier?

+	+	-	-
	-		-
+	+	-	-

[2 marks]

1-NN: 5/10, 3-NN: 1/10

- (b) True or false. 10-fold cross-validation is always better than using a validation set. Explain your answer.

[3 marks]

False. The student should talk about the bias-variance tradeoff in validation.

- (c) If we have  $n$  data points, derive the probability that the  $j$ th data point does not appear in a bootstrap sample?

[3 marks]

In bootstrap, we sample with replacement so each observation in the bootstrap sample has the same  $1/n$  (independent) chance of equaling the  $j$ th observation. Applying the product rule for  $n$  observations gives us  $(1 - \frac{1}{n})^n$

- (d) Consider a simple classification procedure applied to a two-class dataset with 2000 predictors and 100 samples:

- Step 1. Find the 2 predictors having the largest correlation with the response
- Step 2. Fit a logistic regression model using only these 2 predictors

We estimate the performance of this classification procedure on new samples using the following algorithm: (1) apply the classification procedure to multiple bootstrap samples, (2) for each classifier estimated using a bootstrap sample, predict the original dataset, and (3) average the classification error for all the bootstrap samples.

- (i) Using your answer in (c), explain why this algorithm will provide a bad estimate of the classification performance on new samples.

[2 marks]

The problem is the training set (bootstrap sample) and the validation set (the initial dataset) will have observations in common, roughly 63%.

- (ii) Modify this algorithm and show that it will provide a better estimate of the classification performance on new samples.

[2 marks]

We should record the classification error for the  $i$ th observation only for the bootstrap samples that do not contain observation  $i$ .

**[Total: 12 marks]**

— END OF QUESTION 5 —



## QUESTION 6

- (a) Consider 3 observations  $x_1$ ,  $x_2$  and  $x_3$  where  $x_i \in \mathbb{R}^p$ ,  $i = 1, 2, 3$ . The pairwise Euclidean distance matrix is given below.

	$x_1$	$x_2$	$x_3$
$x_1$	0	10.296	30.265
$x_2$	10.296	0	22.847
$x_3$	30.265	22.847	0

- (i) Compute the  $K = 2$  clusters  $C_1$  and  $C_2$  that are solutions to the following optimization problem:

$$\text{minimize}_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$$

Explain how you computed it.

[2 marks]

$$C_1 = \{1, 2\} \text{ and } C_2 = \{3\}$$

- (ii) What is the value of the objective function for the clusters you computed in (i)?

[2 marks]

$$10.296^2/2$$

- (b) Let  $\{x_1, \dots, x_n\}$  be a set of points where  $x_i \in \mathbb{R}^p$ , and  $C_1, C_2, \dots, C_K$ , a set of  $K$  clusters. Prove that the total variation in the data set  $T$  is equal to the sum of the within-cluster variation  $W_K$  and the between-cluster variation  $B_K$ . In other words, prove that

$$T = W_K + B_K,$$

with

$$T = \sum_{i=1}^n \|x_i - \bar{x}\|_2^2,$$

$$W_K = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2,$$

$$B_K = \sum_{k=1}^K n_k \|\bar{x}_k - \bar{x}\|_2^2,$$

where  $\bar{x}$  is the overall average, i.e.  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , and  $\bar{x}_k$  is the average of points in cluster  $k$ , i.e.  $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ . (Hint 1: you need to start from the total variation. Hint 2: you need to add and subtract the cluster averages  $\bar{x}_k$ .)

[4 marks]

$$T = \sum_{i=1}^n \|x_i - \bar{x}\|_2^2 \quad (12)$$

$$= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}\|_2^2 \quad (13)$$

$$= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_k + \bar{x}_k - \bar{x}\|_2^2 \quad (14)$$

$$= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2 + \|\bar{x}_k - \bar{x}\|_2^2 + 2(x_i - \bar{x}_k)(\bar{x}_k - \bar{x}) \quad (15)$$

$$= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2 + \sum_{k=1}^K \sum_{i \in C_k} \|\bar{x}_k - \bar{x}\|_2^2 + \sum_{k=1}^K \sum_{i \in C_k} 2(x_i - \bar{x}_k)(\bar{x}_k - \bar{x}) \quad (16)$$

$$= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2 + \sum_{k=1}^K n_k \|\bar{x}_k - \bar{x}\|_2^2 = W_K + C_K, \quad (17)$$

where we used the fact that  $\sum_{i \in C_k} (x_i - \bar{x}_k) = 0$ .

(c) Suppose that we have five observations, for which we compute a dissimilarity matrix, given by

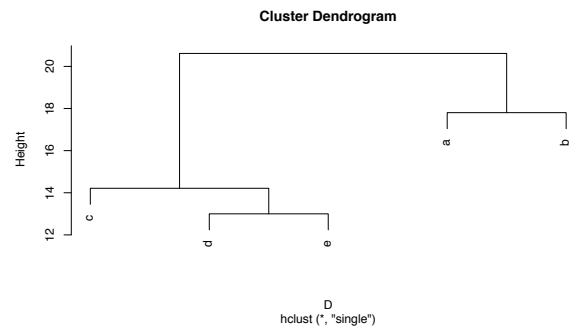
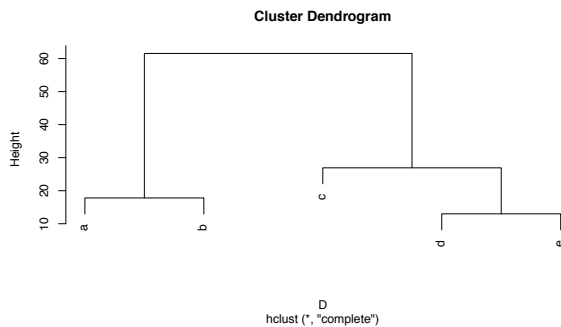
	a	b	c	d	e
a	0	17.804	20.616	44.045	33.121
b	17.804	0	36.359	61.555	50
c	20.616	36.359	0	26.926	14.213
d	44.045	61.555	26.926	0	13
e	33.121	50	14.213	13	0

On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these five observations using *complete linkage*. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram

[3 marks]

(d) Repeat (c), this time using *single linkage* clustering.

[2 marks]



[Total: 13 marks]

— END OF QUESTION 6 —

## QUESTION 7

We consider a data set which contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. Below is a principal component analysis (PCA) of this data set after centering and scaling each column.

	PC1	PC2	PC3	PC4
Murder	-0.536	0.418	-0.341	0.649
Assault	-0.583	0.188	-0.268	-0.743
UrbanPop	-0.278	-0.873	-0.378	0.134
Rape	-0.543	-0.167	0.818	0.089
Variance	2.480	0.990	?	0.173
Cum. Prop.	0.620	?	?	?

- (a) Complete the missing entry in the row "Variance" for PC3. Explain your answer.

[2 marks]

Since the data has been scaled and centered, the total variance is 4. Since the four PC explained the total variance:  $4 - 2.480 - 0.990 - 0.173 = 0.357$

- (b) What proportion of the total variance does the fourth principal component explain?

[2 marks]

0.04325

- (c) Complete the bottom line of the previous Table which gives the cumulative proportions of total variance explained.

[2 marks]

0.620 - 0.868 - 0.957 - 1

- (d) How many principal component directions would we need to explain at least 60% of the variance?

[2 marks]

1 principal component

- (e) Interpret the first principal component.

[2 marks]

Almost the same weight is given to Murder, Assault and Rape, with much less weight on UrbanPop. Hence this component roughly corresponds to a measure of overall rates of serious crimes.

- (f) Let  $\phi_1 \in \mathbb{R}^4$  and  $\phi_2 \in \mathbb{R}^4$  be the first two loading vectors. What is the value of  $\phi_1' \phi_2$ ? Explain your answer.

[2 marks]

$\phi_1' \phi_2 = 0$  since  $\phi_1$  and  $\phi_2$  are orthogonal to each other.

[Total: 12 marks]

— END OF QUESTION 7 —

## QUESTION 8

- (a) True or false. Regression trees can only model constant functions. Explain your answer.

[2 marks]

**FALSE.** Although regression trees assume a constant value in each partitions, the estimated function can be very flexible.

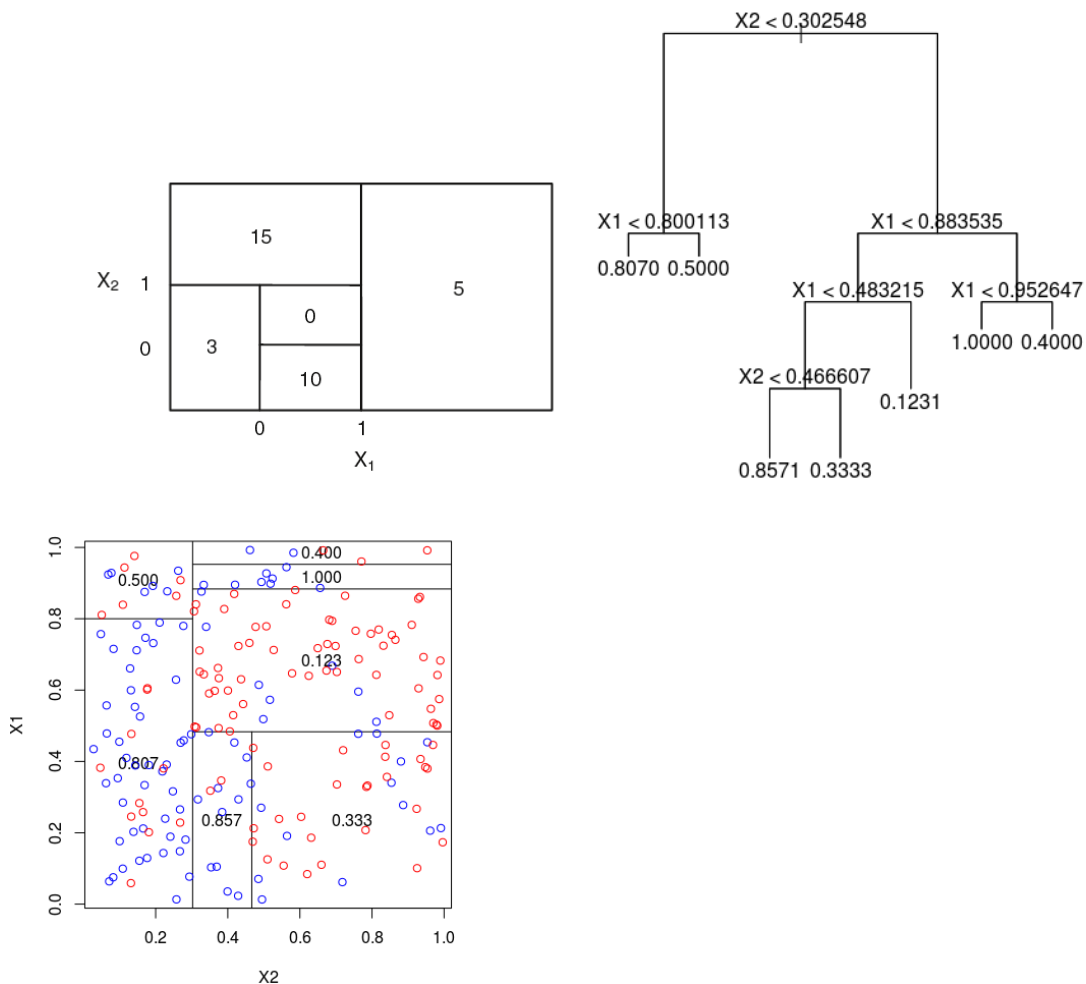
- (b) Suppose we have a sample of  $n$  pairs  $x_i, y_i$  drawn i.i.d. from  $y_i = f(x_i) + \varepsilon_i$ , where  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ ,  $f(x) = \sum_{j=1}^J c_j I(x \in R_j)$  with  $R_1; R_2; \dots; R_J$  being  $J$  partitions of the input space. What are the optimal values of  $c_1, \dots, c_J$  that minimize  $\sum_{i=1}^n (y_i - f(x_i))^2$ ? Give a proof of your answer.

[3 marks]

$$c_j = \frac{\sum_{i=1}^n y_i I(x_i \in R_j)}{\sum_{i=1}^n I(x_i \in R_j)} = \frac{1}{n_j} \sum_{x_i \in R_j} y_i, \text{ which is essentially the average response in partition } R_j.$$

- (c) Create a diagram similar to the left-hand panel of the following Figure, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

[3 marks]



- (d) Let  $Z_1, \dots, Z_B$  be a set of  $B$  i.i.d. (identically distributed, but not necessarily independent) random variables, each with variance  $\sigma^2$ , with positive pairwise correlation  $\rho$  ( $\rho > 0$ ), and  $\bar{Z} = \frac{1}{B} \sum_{b=1}^B Z_b$ .

(i) Prove that

$$\text{Var}(\bar{Z}) = \rho\sigma^2 + \sigma^2\frac{1-\rho}{B}. \quad (18)$$

[3 marks]

$$\text{Cor}(Z_i, Z_j) = \frac{\text{Cov}(Z_i, Z_j)}{\text{sd}(Z_i)\text{sd}(Z_j)} \implies \text{Cov}(Z_i, Z_j) = \text{Cor}(Z_i, Z_j)\sigma^2 = \rho\sigma^2. \quad \text{Var}(\bar{Z}) = \frac{1}{B^2}\text{var}\left(\sum_{b=1}^B Z_b\right) = \rho\sigma^2 + \sigma^2\frac{1-\rho}{B}.$$

(ii) Which algorithm discussed in the class exploits expression (18)? Briefly explain how the algorithm exploits it.

[2 marks]

The random forest algorithm uses this idea with  $B$  estimators. The algorithm tries to make the correlations between the estimators as small as possible by resampling and random feature selection in each split.

**[Total: 13 marks]**

— END OF QUESTION 8 —