Instructions

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

QUESTION 1

Indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method, for each of the following. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

[2 marks]

better performance

(b) The number of predictors p is extremely large, and the number of observations n is small.

[2 marks]

worse performance

(c) The relationship between the predictors and response is highly non-linear.

[2 marks]

better performance

(d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high.

[2 marks]

worse performance

[Total: 8 marks]

— END OF QUESTION 1 —

QUESTION 2

Answer the questions on the following data:

Obs	X_1	X_2	X_3	Y	distance
1	1	1	1	violet	$\sqrt{3}$
2	0	-2	-2	orange	$\sqrt{8}$
3	-1	0	-1	orange	$\sqrt{2}$
4	0	-1	2	violet	$\sqrt{5}$
5	2	1	1	orange	$\sqrt{6}$
6	1	3	0	violet	$\sqrt{10}$
7	-2	-3	0	orange	$\sqrt{13}$

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$. Write it into the table.

[3 marks]

(b) What is our prediction with $K = 1$? Why?

[2 marks]

orange

(c) What is our prediction with $K = 3$?

[2 marks]

violet

(d) What is our prediction with $K = 5$?

[2 marks]

orange

[Total: 9 marks]

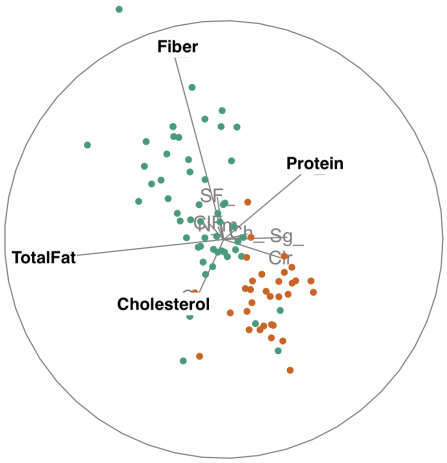
— END OF QUESTION 2 —

QUESTION 3

(a) In the projection from a tour of the chocolates data below, the main pattern in the data that can be seen is a different between dark (green) and milk (orange) chocolates. Which of the following variables can be seen to contribute most to this pattern? (Circle one)

[3 marks]

Fiber TotalFat Cholesterol Protein

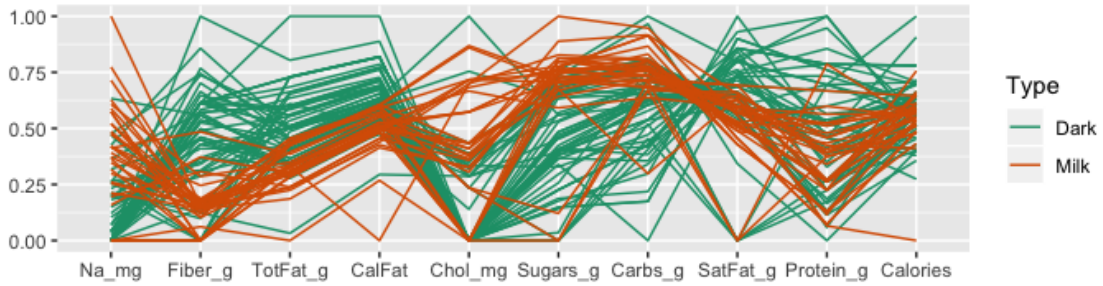


Fiber

(b) From the parallel coordinate plot below, which variables are most important for distinguishing between dark (green) and milk (orange) chocolates? (Circle them)

[3 marks]

Na Fiber TotalFat CalFat Chol Sugars Carbs SatFat Protein Calories



All except for Calories, Protein and Cholesterol

(c) Would it be appropriate to say that both milk and dark chocolates have a similar amount of calories, based on the parallel coordinate plot? **Yes** or **No**.

[2 marks]

Yes

[Total: 8 marks]

— END OF QUESTION 3 —

QUESTION 4

This is a summary of the principal component analysis for the dark chocolates in the data. Standardised nutritional variables are used. There are 56 observations. The last row of the table is the cumulative proportion of variance.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Calories	0.26	-0.38	-0.21	0.39	-0.61	0.15	-0.06	-0.28	0.34	0.05
CalFat	0.44	-0.12	-0.06	-0.03	-0.08	0.17	-0.13	0.35	-0.50	0.60
TotFat_g	0.45	-0.08	-0.06	0.02	-0.03	0.08	-0.17	0.23	-0.29	-0.78
SatFat_g	0.30	-0.40	-0.05	0.09	0.45	-0.59	-0.13	0.18	0.36	0.09
Chol_mg	-0.22	-0.26	-0.55	0.08	0.50	0.56	-0.12	-0.02	0.06	-0.01
Na_mg	-0.14	0.45	-0.11	0.79	0.06	-0.16	-0.29	0.14	-0.12	0.03
Carbs_g	-0.38	-0.27	0.06	0.14	-0.23	0.05	0.37	0.74	0.11	-0.08
Fiber_g	0.06	-0.24	0.78	0.31	0.27	0.38	-0.04	-0.10	0.03	0.00
Sugars_g	-0.34	-0.50	-0.03	0.16	-0.01	-0.33	0.10	-0.33	-0.61	-0.06
Protein_g	0.36	0.17	-0.15	0.26	0.22	0.03	0.83	-0.13	-0.06	-0.00
Variance	4.68	1.27	1.11	0.95	0.75	?0.53	0.35	0.23	0.09	0.04
Cum. Prop. Var.	0.47	0.59	?0.71	?0.80	0.88	0.93	0.96	0.99	1.00	1.00

(a) Fill in the values where there are question marks.

[3 marks]

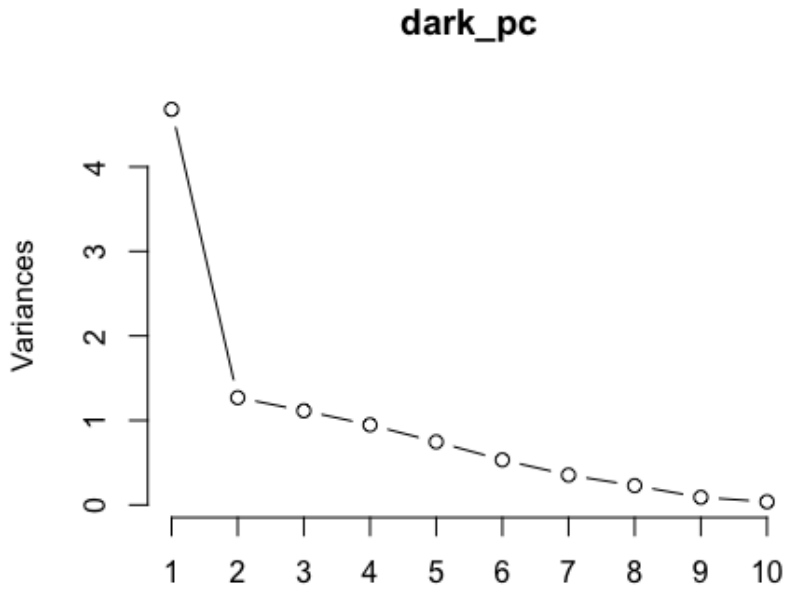
(b) Compute the total variance?

[2 marks]

10

(c) Make a scree plot for the results.

[4 marks]



(d) How many principal components would you suggest be used to reduce the dimensionality of this data? Justify your answer.

[3 marks]

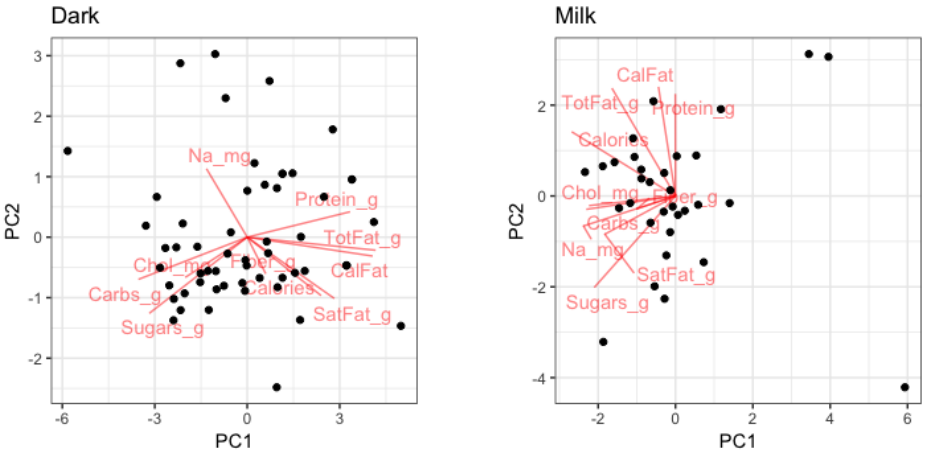
2, based on the scree plot, which would explain about 60% of the variation. The amount of variation explained by more PCs declines very gradually after this, so its hard to differentiate between choices more than 2.

(e) Interpret the first principal component. (What variables is it mostly composed of?)

[3 marks]

The first PC is a rough contrast of Sugars (inc Carbs) against Fats and Protein.

(f) Below are two biplots. One is summarising the PCA for the dark chocolates, and the second is computed on the PCA of the milk chocolates.



i. The two biplots different. What does this imply about the variance-covariance matrices for the two groups?

[3 marks]

The biplots indicate quite different variables contribute to the first two PCs. For dark chocolates PCs is a contrast of fats/protein against sugars, but for milk chocolates it is roughly a contribution of all variables. This indicates that the variance-covariance for each group is quite different.

ii. What is an obvious problem with PCA on the milk chocolates?

[3 marks]

There are some outliers!

(g) TRUE or FALSE. The variance-covariance matrix computed on all of the data is the typically the same as pooling (averaging) the variance-covariance matrices computed separately on each group.

[2 marks]

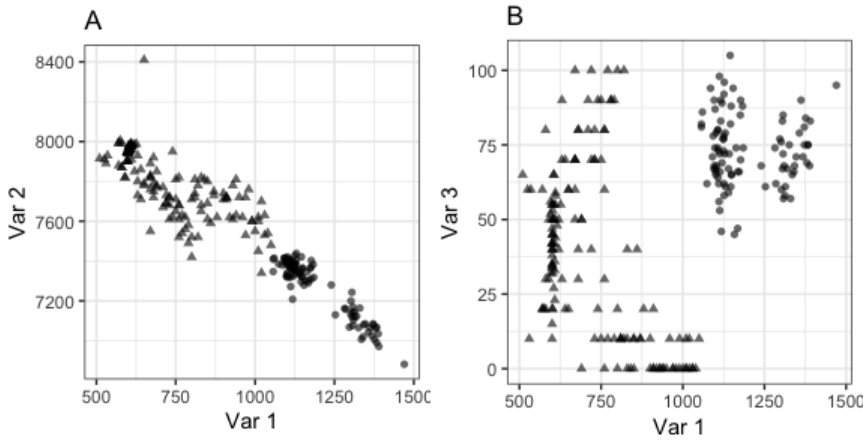
FALSE

[Total: 23 marks]

— END OF QUESTION 4 —

QUESTION 5

Both of the plots below show different views of the same data, with differences between the two groups (circles, triangles).



(a) If you were to choose two variables for splitting the two groups, which would you choose, Var 2 **or** Var 3, in association with Var 1? Explain.

[2 marks]

Var 3 because it contains a big gap between the two groups. It would be more difficult to model, though because it is non-linear. For that reason maybe Var 2 could be the preference, but it will give less reliable future predictions.

(b) A decision tree is fit to the data, using the `rpart` library. This is the tree:

```
n= 249

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 249 98 3 (0.39 0.61)
  2) Var1>=1.1e+03 98  0 2 (1.00 0.00) *
  3) Var1< 1.1e+03 151  0 3 (0.00 1.00) *
```

i. How many observations in the data? 249

[2 marks]

ii. What is the predicted value for the split at node 2? 2

[2 marks]

iii. What is the error for the model? 0

[2 marks]

iv. How many terminal nodes are there? 2

[2 marks]

(c) A random forest model is fit to the same data, and variable importance is calculated as follows:

	2	3	MeanDecreaseAccuracy
Var6	0.00	0.00	0.00
Var7	0.01	0.00	0.01
Var5	0.01	0.00	0.01
Var2	0.28	0.11	0.18
Var1	0.33	0.18	0.24
Var4	0.05	0.01	0.03
Var3	0.05	0.01	0.03

i. Which variables are the most important?

[2 marks]

Vars 1 and 2

ii. Var 3 in conjunction with Var 1 produce a big gap between the two groups (as seen from the plot in part a). Why doesn't Var 3 show up as being an important variable in the random forest model?

[3 marks]

Trees are greedy algorithms, and because it is possible to split the data perfectly with just Var 1 and nearly with Var 2, the trees don't pick up on Var 3 as being important, which follows through in the forest model.

(d) Sketch what you think the boundary between the two groups in Var 1 and Var 3 might be if a **radial** kernel SVM classifier is used.

[2 marks]

It will tightly wrap around the smaller group, circles.

[Total: 17 marks]

— END OF QUESTION 5 —

QUESTION 6

A (feed forward back propagation) neural network can be written as a nested regression model:

$$\hat{y} = g(\alpha_0 + \sum_{k=1}^{s}(\alpha_k f(\beta_{0k} + \sum_{j=1}^{p}\beta_{jk}x_j))) \quad (1)$$

Let $u = \beta_0 + \sum_{j=1}^{p}\beta_j x_j$ and f is a logistic function, $\frac{1}{1+e^{-u}}$.

The model was fitted to a data set with 2 variables, and 4 nodes in the hidden layer were used, yielding these coefficients:

	b0	b1	b2
node 1	-19.02	18.22	77.10
node 2	-2.34	18.65	-16.43
node 3	-67.63	40.33	69.44
node 4	-41.02	66.94	24.29

(a) What is the value of s in the fitted model? 4

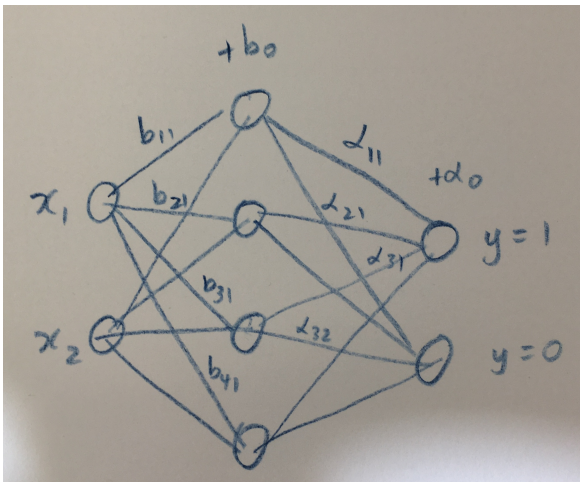
[2 marks]

(b) What is the value of p in the fitted model? 2

[2 marks]

(c) Make a sketch of the network diagram for this data.

[3 marks]



(d) Write out the equation for the logistic regression at the first node of the hidden layer.

[3 marks]

$\hat{s_1} = \frac{1}{1+e^{-(-19.02+18.22x_1+77.10x_2)}}$

(e) Generally, in relation to logistic regression, show that the logistic function

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

can be re-arranged into

$$\log_e \frac{f(x)}{1-f(x)} = \beta_0 + \beta_1 x$$

.

[4 marks]

$f(x) = \frac{1}{1/e^{\beta_0+\beta_1 x}+1}$
$\rightarrow 1/f(x) = 1/e^{\beta_0+\beta_1 x} + 1$
$\rightarrow 1/f(x) - 1 = 1/e^{\beta_0+\beta_1 x}$
$\rightarrow \frac{1}{1/f(x)-1} = e^{\beta_0+\beta_1 x}$
$\rightarrow \frac{f(x)}{1-f(x)} = e^{\beta_0+\beta_1 x}$
$\rightarrow \log_e \frac{f(x)}{1-f(x)} = \beta_0 + \beta_1 x$

[Total: 14 marks]

— END OF QUESTION 6 —

QUESTION 7

In ridge regression we minimise this function:

$$\underset{\beta}{\text{minimize}} \left(\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

where $\lambda \geq 0$ is a tuning parameter, and

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

(a) TRUE or FALSE. If $\lambda = 0$ the model fit equal is least squares. TRUE

[2 marks]

(b) If λ is very, very large what will β_j equal? 0

[2 marks]

(c) What would be the change in the formula that would change this to lasso?

[2 marks]

β_j^2 would be replaced with $|\beta_j|$

(d) Explain in two sentences how ridge regression effectively operates to enable model fitting with a large number of variables and few observations.

[3 marks]

Ridge regression forces the coefficients of some variables to 0, automatically dropping them from the model fit.

[Total: 9 marks]

— END OF QUESTION 7 —

QUESTION 8

(a) Match the linkage type to the explanation.

[4 marks]

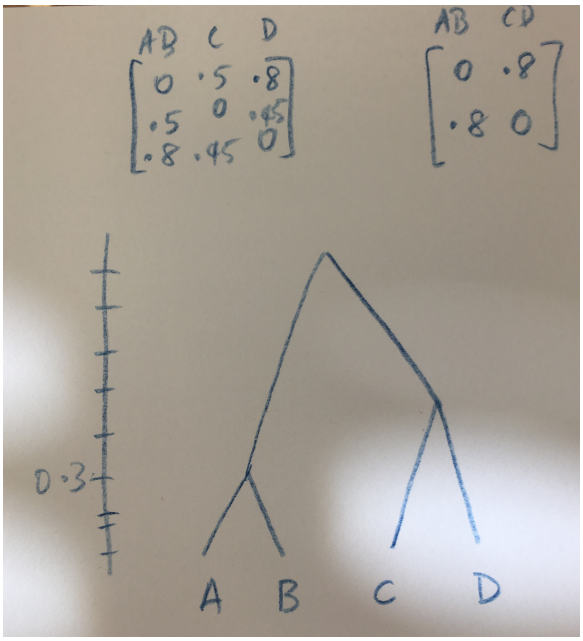
linkage	explanation
Complete	Maximal intercluster dissimilarity.
Single	Minimal intercluster dissimilarity.
Average	Mean intercluster dissimilarity.
Centroid	Dissimilarity between the measure of centres

Its already in correct order

(b) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four 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.

[4 marks]

$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$



(c) Does the following metric satisfy the definition of a distance metric? Justify your answer.

[4 marks]

$$d_{x,y} = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p|$$

There are four properties that need to be satisfied: (1) >0 (2) =0 indicates same point, (3) $d_{x,y} = d_{y,x}$, (4) $d_{x,y} < d_{x,z} + d_{y,z}$. 1, 2, 3 are all clearly satisfied. 4 is a bit harder to justify, but it is can be argued that it follows the triangle inequality, that $|x_j - y_j| < |x_j - z_j| + |y_j - z_j|, \forall j = 1, ..., p$ A proof would be made by squaring both sides.

[Total: 12 marks]

— END OF QUESTION 8 —