

ETC3250

Practice exam 2016

Instructions

- In the actual exam, there are 11 questions worth a total of 100 marks. You should attempt them all.
- Open book
- You can use the approved calculator

This practice exam has a range of questions that show possible range of topics that are examined. They come from last year's exam and practice exam, and there are a few additional new questions. Some of these questions are more open ended than the questions on the actual exam, there are fewer, "what is XXX, write a few sentences explaining XXX" on the actual exam.

QUESTION 1

- (a) Briefly explain why best subset selection is not computationally feasible when the number of predictors p is large.

[2 marks]

With best subset selection, we need to consider $\binom{p}{s}$ for $s = 1, \dots, p$. The number of possible combinations increases exponential with p . If p is large, this is not computationally feasible.

- (b) True or false. By constraining the L_1 norm of the vector of coefficients, it is possible to obtain sparse estimate for the coefficients. Explain your answer.

TRUE. This can be explained geometrically. The L_1 ball has corners which encourage sparse solutions.

- (c) When is Multidimensional scaling (MDS) is equivalent to Principal Component Analysis (PCA)?

When the pairwise distance matrix uses Euclidean distances, and no transformations of the distances are applied.

- (d) What is the difference between 2-fold cross-validation and 10-fold cross-validation? Which one would you use in practice?

[2 marks]

The number of folds K control the bias and variance tradeoff in the cross-validation estimate. A large value decrease the variance but increase the bias, and vice versa. 10-fold cross-validation is often used in practice. However, 2-fold cross-validation can also be used with a large dataset.

- (e) Write the formula to compute the prediction of a K -nn classifier for the new data point x_0 .

[2 marks]

$$\frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbf{I}(y_i = j)$$

[Total: 6 marks]

— END OF QUESTION 1 —

QUESTION 2

Suppose you estimate the coefficients of a linear regression by solving the following optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

where $\lambda \geq 0$ is a **tuning parameter**.

- (a) Briefly explain how λ affects the bias and variance tradeoff of your estimate $\hat{\boldsymbol{\beta}}$.

[2 marks]

$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'y$. In the limit, when $\lambda = \infty$, the variance is zero, and the bias is very large. The other extreme is when $\lambda = 0$, which reduce to the least square estimate, with small bias and high variance. By using a value of λ between these two extremes, we can control the bias and variance tradeoff.

- (b) Explain why it is important to standardize the predictors in this case

[2 marks]

If we do not apply standardization, the predictors with the largest variance will dominate the penalty term.

- (c) Sparsity is very useful in high-dimensional regression. Explain why.

A lot of zero coefficients provides better interpretability of the final model. Possibly better predictions due to the lower variance. Finally, the best on sparsity principle.

- (d) Does the previous optimization allows sparse estimates? Explain your answer.

Although it shrinks the estimate towards zero, it does not provide sparse estimate, except when $\lambda = \infty$ where all the coefficients are zero which is not a useful solution in general. To obtain sparse estimate, we need to consider L_q norm with $q \leq 1$.

[Total: 4 marks]

— END OF QUESTION 2 —

QUESTION 3

This question is about bootstrapping.

- (a) Give an example of algorithm that uses bootstrapping and why.

Random forests. It allows to build multiple trees and average the results to reduce the variance.

- (b) Bootstrapping can be used to estimate the sampling distribution of a statistic. Explain this procedure.

- Sample with replacement the i.i.d. observations
- For each bootstrap sample, compute the statistic
- The distribution of these bootstrap statistics is an estimate of the sampling distribution of the statistic.

- (c) Can we use the previous bootstrap procedure when the data is a time series? Explain.

No. Time series data do not satisfy the i.i.d assumption required by the bootstrap procedure. It is possible to use another bootstrap procedure for time series, called block bootstrap for example.

[Total: 0 marks]

— END OF QUESTION 3 —

QUESTION 4

- (a) Briefly explain why the K-means algorithm is guaranteed to decrease the value of the objective at each step.

The sum of the squared euclidean distance between the observations in a cluster is equal to two times the sum of the squared euclidean distance between each observation and the centroid. The fact that K-means assign observations to the closest centroid, it is guaranteed that the sum of the squared euclidean distance will decrease.

- (b) True or false. For any starting values of the assignment of the observations, the K-means algorithm will always converge to the same solution.

FALSE. The K-means algorithms converges to a local optimum.

- (c) For the following data.

	X1	X2	X3	X4
A	-1.02	0.27	-0.81	-0.34
B	-0.61	0.97	0.76	0.71
C	0.70	-0.38	0.88	-0.32
D	-0.82	0.48	-0.71	-0.98
E	-0.72	0.97	-0.33	0.04

	A	B	C	D	E
A	0.00	2.06	2.50	0.71	0.98
B	2.06	0.00	2.15	2.30	1.28
C	2.50	2.15	0.00	2.45	2.33
D	0.71	2.30	2.45	0.00	1.20
E	0.98	1.28	2.33	1.20	0.00

and the associated distance matrix:

- (i) Compute the Euclidean distance between observations A and E.

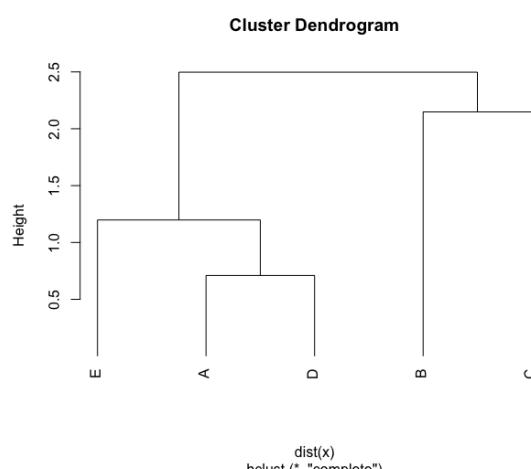
[2 marks]

0.977

- (ii) The dendrogram shows hierarchical clustering with complete linkage. Which two points were fused at the first step of hierarchical clustering with complete linkage?

[2 marks]

A, D



- (iii) What is the intercluster distance (linkage) values between the new cluster and the remaining four points? (There are three.)

[2 marks]

AD-B = 2.30, AD-C=2.45, AD-E=1.20

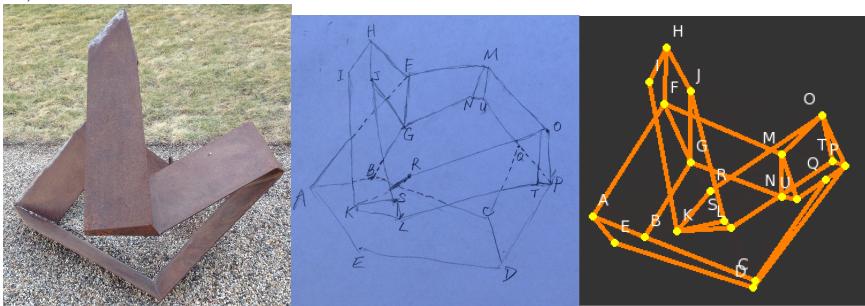
[Total: 6 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about dimension reduction.

You are asked to curate an art exhibit, containing 15 identical sculptures, that has one rule: the sculptures must all be oriented and resting on the ground uniquely. As you walk around the current exhibit you notice that there are several pairs where the sculptures are actually identically placed. To solve the problem you decide to make a virtual model by measuring the distances between each pair of vertices, and using multi-dimensional scaling to produce a 3D layout. Below are some photos of the exhibit, an initial sketch and the virtual model obtained.



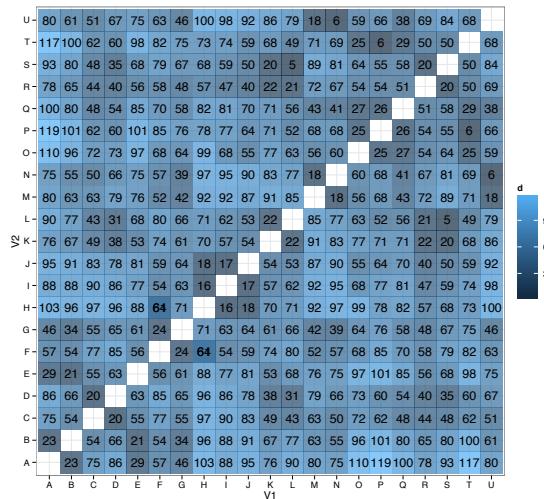
- (a) Sketch the three point shape that would perfectly minimise the MDS classical stress function of this distance matrix.

[2 marks]

$$\begin{bmatrix} 0 & 3 & 4 \\ 3 & 0 & 5 \\ 4 & 5 & 0 \end{bmatrix}$$

It is the right triangle with sides 3, 4, 5

- (b) Below is a display of the inter-vertex distance matrix for one sculpture.



The eigenvalues are:

```
> dm_pca$sdev^2
[1] 5.614e+03 4.284e+03 2.654e+03 7.438e+02 4.318e+02 1.349e+02 1.157e+02
[8] 7.505e+01 4.341e+01 2.743e+01 2.479e+01 1.757e+01 1.699e+01 1.299e+01
[15] 1.058e+01 9.287e+00 5.996e+00 1.907e+00 1.634e+00 9.388e-01 2.307e-29
```

and the coefficients of the first eigenvector are:

A	B	C	D	E	F	G	H	I	J	K	L	M
0.35	0.33	0.02	-0.09	0.23	0.15	0.18	-0.22	-0.19	-0.26	-0.15	-0.24	0.18
N	O	P	Q	R	S	T	U					
0.19	-0.20	-0.29	-0.12	-0.17	-0.26	-0.30	0.18					

- (i) What does this tell you about the main axis of variation of the sculpture?

[2 marks]

There are groups of vertices: (A, B, E) have high positive coefficients, (H, J, L, P, S, T) have high negative coefficients, and the remaining have little contribution to this main direction.

- (ii) Sketch the scree plot. How many principal components would the scree plot suggest should be used to summarise the distances?

[2 marks]

4 or 6

- (iii) What proportion of total variation do three principal components explain?

[2 marks]

0.882

- (iv) Explain how multidimensional scaling is related to principal components, in the context of this problem?

[2 marks]

We don't have an actual data matrix in this problem, simply the distance matrix. PCA will tell us how the variation decomposes among the vertices, but it does not give us a configuration of points in low-dimensions that can be used to construct the sculpture virtually. If we had data, we could predict the data values on the first three principal components, and these would be the 3D layout that would best match the full distance matrix, and would be used as the virtual model.

[Total: 10 marks]

— END OF QUESTION 5 —

QUESTION 6

A principal component analysis is conducted on a subset of Mexico City data, health and pollution where missing values have been imputed using regression methods. There are five variables in the subset: deaths (number of deaths each day), temp_mean (average temperature), humidity, NOX (nitrogen oxide, pollutant), O3 (ozone, pollutant).

```
> mexico.pca1 <- prcomp(mexico[,c("deaths", "temp_mean",
  "humidity", "NOX", "O3")], scale=T, retx=T)
```

```
> mexico.pca1
```

Standard deviations:

```
[1] 1.37 1.27 0.86 0.64 0.62
```

Rotation:

	PC1	PC2	PC3	PC4	PC5
deaths	-0.30	0.573	-0.471	-0.597	0.077
temp_mean	-0.33	-0.578	0.338	-0.631	0.213
humidity	0.64	-0.074	0.003	-0.468	-0.608
NOX	-0.25	0.510	0.766	-0.028	-0.300
O3	-0.58	-0.268	-0.279	0.161	-0.699

```
> summary(mexico.pca1)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.367	1.267	0.856	0.6362	0.624
Proportion of Variance	0.374	0.321	0.146	0.0809	0.078
Cumulative Proportion	0.374	0.695	0.841	0.9220	1.000

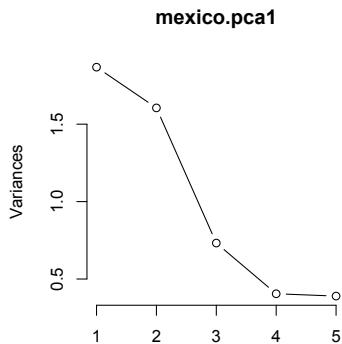
- (a) Compute the total variance.

[1 marks]

5

- (b) Make a sketch of the scree plot. Label your axes.

[2 marks]



- (c) The PCA was conducted on the correlation matrix. Why do you think that this was necessary?
[1 marks]

The variables were recorded in different units.

- (d) What proportion of variance is explained by four PCs?

[1 marks]

0.922

- (e) With the help of the summary statistics below, compute the principal component score for this data value: $deaths = 296$, $temp_mean = 12$, $humidity = 53$, $NOX = 0.028$, $O3 = 0.026$.

[2 marks]

```
> apply(mexico[,c("deaths", "temp_mean", "humidity", "NOX", "O3")], 2, mean)
   deaths temp_mean humidity      NOX       O3
 273.951     16.767    52.969     0.055     0.022
> apply(mexico[,c("deaths", "temp_mean", "humidity", "NOX", "O3")], 2, sd)
   deaths temp_mean humidity      NOX       O3
 28.990      2.652    14.544     0.024     0.009
```

- (f) Interpret the first principal component.

[2 marks]

The first PC is a combination of all of the variables, but mostly a contrast between humidity and O3. If humidity is high O3 tends to be low.

- (g) How many principal components would you use to summarize the variation of this data? Why?

[2 marks]

Scree plot really suggests 4. Even two PCs explains a lot of the variation, though. Definitely need two, they explain different things in the variables. Coefficients of third PC contradicts the second, deaths and NOX have opposite signs, although it might be interpreted as only NOX. Similarly for PC4. Given that the eigenvalue for PC 3 is so much smaller, and not a lot more variation is explained, I'd go with two rather than 4.

[Total: 11 marks]

— END OF QUESTION 6 —

QUESTION 7

(a) Select the propositions that are true, and briefly explain your answer:

[4 marks]

- The test error can be smaller than the training error. **TRUE**
- The model is underfitting when the training error is very large. **TRUE**
- Increasing the number of neighbours in the K-nearest neighbours algorithm will increase the flexibility of the model. **FALSE**
- Using cross-validation for model selection will necessarily provide models with better prediction accuracy compared to models selected by AIC and BIC. **FALSE**

(b) Consider a simple classification procedure applied to a two-class dataset with 5000 predictors and 50 samples:

Step 1. Find the 100 predictors having the largest correlation with the class labels.

Step 2. Apply logistic regression using only these 100 predictors .

How do we estimate the test set performance of this classification procedure?

[2 marks]

Split into training and testing. Not forgetting to include step 1 in the procedure.

Can we use cross-validation? If yes, describe the step-by-step cross-validation procedure.

[3 marks]

Yes. A typical cross-validation procedure, but check if step 1 has also been considered in the cv procedure.

(c) If we have n data points, what is the probability that a given data point does not appear in a bootstrap sample?

[3 marks]

$$(1 - \frac{1}{n})^n$$

[Total: 12 marks]

— END OF QUESTION 7 —

QUESTION 8

- (a) Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = \text{GPA} \times \text{IQ}$, and $X_5 = \text{GPA} \times \text{Gender}$. The response Y is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

where e is a random error term, and we obtain estimates $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

Which answer is correct, and why?

[4 marks]

- (a) For a fixed value of IQ and GPA, males earn more on average than females.

False

- (b) For a fixed value of IQ and GPA, females earn more on average than males.

True, if $\text{GPA} < 3.5$, otherwise false.

- (c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

True. Once $\text{GPA} > 3.5$

- (d) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

False

- (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

[2 marks]

137.1

- (c) True or false: Since the coefficient for the $\text{GPA} \times \text{IQ}$ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

[3 marks]

This is true. The coefficient is small but the magnitude of values large, so the effect is small. In contrast, the coefficient for $\text{GPA} \times \text{Gender}$ is large, and this is where the interaction does have an effect on salary.

- (d) Suppose that the true relationship between starting salary and IQ is nonlinear for fixed values of GPA and Gender. You wish to compare the cubic model (including X_2 , X_2^2 and X_2^3) to the linear model (including X_2 but not the quadratic or cubic terms). You split the available data into two forming a training set and test set, and you estimate both models using only the training data. Which of the two models would you expect to have larger mean squared error computed on the training data, or is there not enough information to say? Which of the two models would you expect to have larger mean squared error computed on the test data, or is there not enough information to say? Justify your answer.

[4 marks]

The MSE for the linear model should be higher than the more complicated model. Adding more terms reduces the error term. It may be hard to say which has the bigger error for the test data. Errors should be higher on the test data anyway, but because the model was fitted to the training set, it is possible to have larger error with the more complicated model because the training data was overfitted.

- (e) You need to decide whether to add the quadratic and cubic terms to the model. Explain how you would do this using a test set, using cross-validation, and using the AIC. Comment on which of these three methods you would prefer and why.

[3 marks]

It is always good to have the training/test set because it is clear what the model is fitted on, and how it performs on the new data. However, if this approach is used, you really need to have training, validation and test sets, so that you have a set of data for prediction that was not used at all in the model building. The drawback is that you have one slice of the pie and all inference is depending on the training test split.

Cross-validation should give a better handle on error with future data, because it guards against odd sampling effects. You could use cross-validation on training and validation sets, to decide whether the cubic model gets substantial reduction in test error enough to use that model. Then you would need to fit the model with the training set, and finally report error for the test set.

AIC is useful for model selection. The linear vs cubic models are nested, which is the appropriate situation for comparing AIC values. You can still work with a training set of data, and use AIC to select the model. It should help to prevent overfitting.

[Total: 16 marks]

— END OF QUESTION 8 —

QUESTION 9

The OECD PISA Results in Focus report describes the survey as “the world’s global metric for quality, equity and efficiency in school education”. The goal of the PISA survey is to assess the workforce readiness of 15-year old students. Nearly 500,000 students were tested across 65 countries and economies. Students were examined on how well they can apply the knowledge they learned in school to applications outside of school. The reported scores range between 0–1000. Information about the students, parents, and schools is also collected. The students completed a questionnaire providing information about themselves, their homes, their schools, and a variety of psychological views regarding factors they believe affect their performance in school. School principals responded to a questionnaire covering their school system and learning experiences for their students. In some countries, parents completed a questionnaire requesting information about their perceptions regarding the school system, expectations for their child, and their involvement in their child’s schooling.

You have a sample of this data for five countries: Australia, Germany, Japan, Jordan and South Korea. The variables are responses of the students to questions about how they use the internet, for purposes that include “One player games”, “Collaborative games”, “Use email”, “Chat on line”, “Social networks”, “Browse the Internet for fun”, “Read news”, “Obtain practical information”, “Download music”, “Upload content”, “Internet for school”, “Email students”, “Email teachers”, “Download from School”, “Announcements”, “Homework”, “Share school material”:

- 0 Never or hardly ever
- 1 Once or twice a month
- 2 Once or twice a week
- 3 Almost every day
- 4 Every day

In addition, you have information about gender, and schoolid. Missing values are codes as NA. For this question we will focus just on this subset of the variables: “One player games”, “Collaborative games”, “Download music”, “Upload content”. This means that there are 5 variables, and 37904 student responses, with the following breakdown by country:

Country	Count
Australia	14481
Germany	5001
Japan	6351
Jordan	7038
South Korea	5033

- (a) This is a rough summary of the data, including summaries of missing values by cases and variables.

	name	Gender	One.player.games	Collaborative.games	Download.music
Australia	:14481	Female:18515	Min. :0.00	Min. :0.00	Min. :0.00
Germany	: 5001	Male :19389	1st Qu.:0.00	1st Qu.:0.00	1st Qu.:1.00
Japan	: 6351		Median :1.00	Median :0.00	Median :2.00
Jordan	: 7038		Mean :1.27	Mean :1.01	Mean :1.93
South Korea	: 5033		3rd Qu.:2.00	3rd Qu.:2.00	3rd Qu.:3.00
			Max. :4.00	Max. :4.00	Max. :4.00
			NA's :2567	NA's :2612	NA's :2613

Cases with missings	Count
0	37533
1	34
2	12
3	325

(a) What percentage of the internet usage variables had missing values?

[2 marks]

100%

(b) How many cases (students) did not complete any of these survey questions at all?

[2 marks]

325

(c) Write a paragraph on how you would handle the missing values in an analysis of the data.

[2 marks]

(For this question I am really looking to see how you would plan out and approach the problem, based on what you know about missing values. These few sentences are a start, but whatever you write here needs to be in your own words.)

Working only with the current data: For the students with missings on all three variables it is very difficult to do anything. It would be best to remove these. For the students that have missings on 1 (and 2) variables, it would be useful to impute these values. Based on their answers for the non-missing values, and stratified by the country and gender, use the mode of the variable from students that did answer.

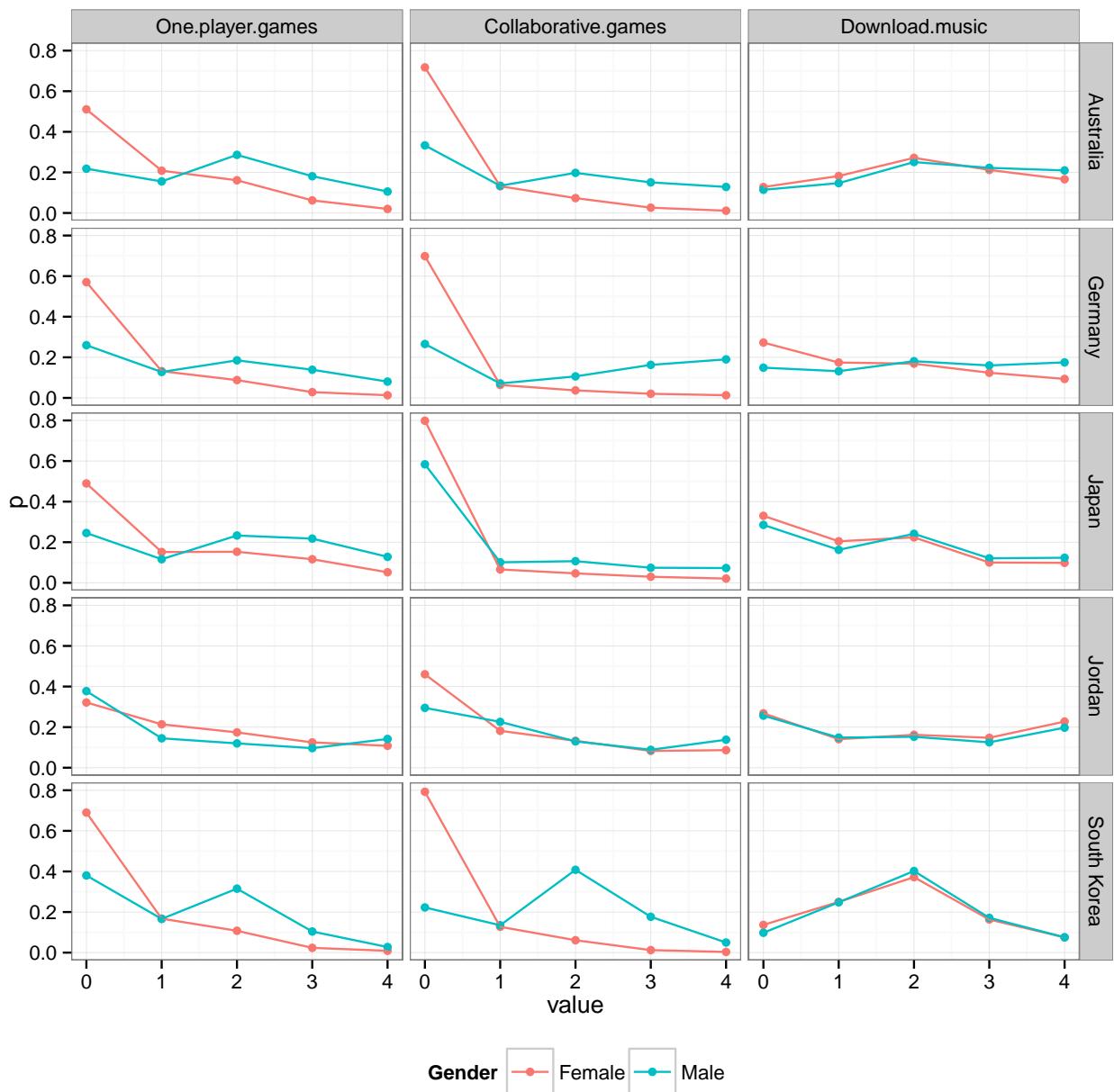
To impute values for students with missing on all three we would need to get additional data, which is available in the full study, so that we have a reasonable number of students that are similar on the complete data, who's values we can aggregate to get substitutes for the missing students data.

(b) We want to look at how students report using the internet, and in particular whether there are differences between countries and gender. To organize the data we need to rearrange it and summarise in order to make some plots. The following code was used to (partially) do this:

```
> internet.m.tb <- internet[,c(1,3,4,5,12,13)] %>%
+   gather(key=usage, value=value, -name, -Gender) %>%
+   group_by(name, Gender, usage, value) %>%
+   tally(sort=TRUE)
> head(internet.m.tb)
Source: local data frame [6 x 5]
Groups: name, Gender, usage [1]

      name Gender      usage value     n
      (fctr) (fctr)    (fctr) (int) (int)
1 Australia Female One.player.games     0  3613
2 Australia Female One.player.games     1  1474
3 Australia Female One.player.games     2  1140
4 Australia Female One.player.games     3   439
5 Australia Female One.player.games    NA   268
6 Australia Female One.player.games     4   141
...
> internet.m.tb.n <- summarise(group_by(internet.m.tb, name, Gender, usage), tot=sum(n))
> internet.m.tb <- merge(internet.m.tb, internet.m.tb.n)
> internet.m.tb.p <- summarise(group_by(internet.m.tb, name, Gender, usage, value), p=n/tot)
> head(internet.m.tb.p)
Source: local data frame [6 x 5]
Groups: name, Gender, usage [1]

      name Gender      usage value     p
      (fctr) (fctr)    (fctr) (int) (dbl)
1 Australia Female One.player.games     0 0.51067138
2 Australia Female One.player.games     1 0.20833922
3 Australia Female One.player.games     2 0.16113074
4 Australia Female One.player.games     3 0.06204947
5 Australia Female One.player.games     4 0.01992933
```



```

6 Australia Female One.player.games      NA 0.03787986
...
qplot(value, p, data=internet.m.tb.p, geom=c("point", "line"), color=Gender) +
  facet_grid(name~usage) +
  theme_bw() + theme(legend.position="bottom")

```

The plot shows the fully rearranged and summarised data of proportions of the level of usage, separately by country and purpose, colored by gender.

- (a) What does the code block (shown below from the full set of commands) do?

[2 marks]

```
> internet.m.tb.n <- summarise(group_by(internet.m.tb, name,
                                    Gender, usage), tot=sum(n))
> internet.m.tb <- merge(internet.m.tb, internet.m.tb.n)
```

It computes the total count separately for country, gender and usage variable, over the different levels of each usage variable, e.g. sums 3613, 1474, 1140, 439, 268, 141 for Australia, Female, One.player.games. This total number is then merged back onto the original table, so that the total count (7075) are repeated:

- (b) Explain (in a few sentences) what the “wide” and “long” forms of the data are according to “tidy”, and how this can help when working with large data.

[2 marks]

The long form of the data has data represented by identifiers/keys for each value. It is a very primal form of a data set that allows it to be molded and reshaped, and aggregated in many ways. Having data in long form provides flexibility, once the data is reshaped, for conducting all sorts of analyses. The wide form is typically required for particular analyses.

- (c) From the plot, would you say there is a different pattern of one player gaming by boys and girls in Australia? And give your reasons.

[2 marks]

Yes, it looks so. There are twice as many girls who report that they do not play one player games. And for playing at least a few times each week, about 10% more boys than girls report that they do.

- (d) From the plot, which country and what type of usage exhibits the biggest difference between boys and girls? Describe the difference.

[2 marks]

South Korea, collaborative games seems to have the biggest difference. The majority of girls report no playing, whereas boys say they tend to play several times a week.

- (e) When you reduce the counts to proportions there are various choices in how to compute the proportion, and this changes the way the results can be compared and contrasted. Explain how proportions were calculated here, and why you think this was a reasonable way to calculate them in order to answer the two questions above.

[2 marks]

Proportion was calculated within each country, gender, usage variable, that is, usage level conditional on these three variables. This allowed direct comparison of the levels of usage by gender and country. To be complete we would want to look at the marginal distributions, proportion by country, proportion of girls, proportion by country and girls, ... and conditional distributions like proportion of girls conditional on country,

- (f) One cognitive phenomenon to avoid when making plots is “change blindness”. Explain what this means. Does faceting of plots potentially induce change blindness? if so, how?

[2 marks]

Change blindness means a switch in attention that there is actually a break in a view of an object. You get this effect in the presence of a strobe light, where lights go on and off and the scene changed. It happens as you stand on a street corner and a bus goes past obscuring the view across the street. To be able to digest changes that happened while obscuring by remembering what everything looked like before. Faceted plots have a gap between the plots, and this interrupts the perception of structure, so the different facets do suffer from change blindness, to some extent. You need to shift your visual focus from one plot to another, which creates the disruption. Primary comparisons are best to show in the same plot, and faceting helps to make secondary (less important) comparisons.

(g) How were “NA” or missings handled?

[2 marks]

the number of missings was incorporated in the calculations of the proportions but this category was not shown in the data plots.

[Total: 20 marks]

— END OF QUESTION 9 —

QUESTION 10

The OECD PISA Results in Focus report describes the survey as “the world’s global metric for quality, equity and efficiency in school education”. The goal of the PISA survey is to assess the workforce readiness of 15-year old students. Nearly 500,000 students were tested across 65 countries and economies. Students were examined on how well they can apply the knowledge they learned in school to applications outside of school. The reported scores range between 0–1000. Information about the students, parents, and schools is also collected. The students completed a questionnaire providing information about themselves, their homes, their schools, and a variety of psychological views regarding factors they believe affect their performance in school. School principals responded to a questionnaire covering their school system and learning experiences for their students. In some countries, parents completed a questionnaire requesting information about their perceptions regarding the school system, expectations for their child, and their involvement in their child’s schooling.

You have a sample of this data for five countries: Australia, Germany, Japan, Jordan and South Korea. The variables are responses of the students to questions about how they use the internet, for purposes that include “One player games”, “Collaborative games”, “Use email”, “Chat on line”, “Social networks”, “Browse the Internet for fun”, “Read news”, “Obtain practical information”, “Download music”, “Upload content”, “Internet for school”, “Email students”, “Email teachers”, “Download from School”, “Announcements”, “Homework”, “Share school material”:

- 0 Never or hardly ever
- 1 Once or twice a month
- 2 Once or twice a week
- 3 Almost every day
- 4 Every day

In addition, you have information about gender, and schoolid. Missing values are codes as NA. For this question we will focus just on this subset of the variables: “Use email”, “Chat on line”, “Social networks”, “Browse the Internet for fun”. This means that there are 6 variables, and 37904 student responses, with the following breakdown by country:

Country	Count
Australia	14481
Germany	5001
Japan	6351
Jordan	7038
South Korea	5033

This is a rough summary of the data, including summaries of missing values by cases and variables.

Use.email	Chat.on.line	Social.networks	Browse.the.Internet.for.fun
Min. :0.000	Min. :0.000	Min. :0.00	Min. :0.000
1st Qu.:1.000	1st Qu.:0.000	1st Qu.:1.00	1st Qu.:2.000
Median :2.000	Median :2.000	Median :3.00	Median :3.000
Mean :1.908	Mean :1.734	Mean :2.43	Mean :2.539
3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:4.00	3rd Qu.:4.000
Max. :4.000	Max. :4.000	Max. :4.00	Max. :4.000
NA's :2691	NA's :2716	NA's :2613	NA's :2600

Cases with missings	Count
0	37525
1	35
2	8
3	6
4	330

(a) What percentage of the internet usage variables had missing values?

[1 marks]

100%

(b) How many cases (students) did not complete any of these survey questions at all?

[1 marks]

330

(c) Write a paragraph on how you would handle the missing values in an analysis of the data.

[2 marks]

Cases where the students didn't respond to any of the survey questions are probably best removed, because we don't have enough other information to reliably impute their likely responses. Possibly this needs to be done for students who missed two questions. For the students who provided some responses we can use these responses, along with the demographic variables to impute the missings. Because it is count data, it is perhaps best to use the mode of the complete cases for the same strata of student as the imputed value. (I don't expect this complete an answer from the students, but some reasoning along these lines would get full marks.)

We want to look at how students report using the internet, and in particular whether there are differences between countries and gender. To organize the data we need to rearrange it and summarise in order to make some plots. The following code was used to (partially) do this:

```
> library(dplyr)
> library(tidyr)
> head(internet[,c(1,3,6,7,8,9)])
  name Gender Use.email Chat.on.line Social.networks Browse.the.Internet.for.fun
1 Australia Female      2          1            3                  1
2 Australia Female      1          2            4                  3
3 Australia Female      3          3            3                  3
4 Australia   Male     2          4            4                  4
5 Australia   Male     2          3            3                  3
6 Australia Female      0          2            4                  3
...
> internet.m.tb <- internet[,c(1,3,6,7,8,9)] %>%
  gather(key=usage, value=value, -name, -Gender) %>%
  group_by(name, Gender, usage, value) %>%
  tally(sort=TRUE)
> head(internet.m.tb)
Source: local data frame [6 x 5]
Groups: name, Gender, usage [1]

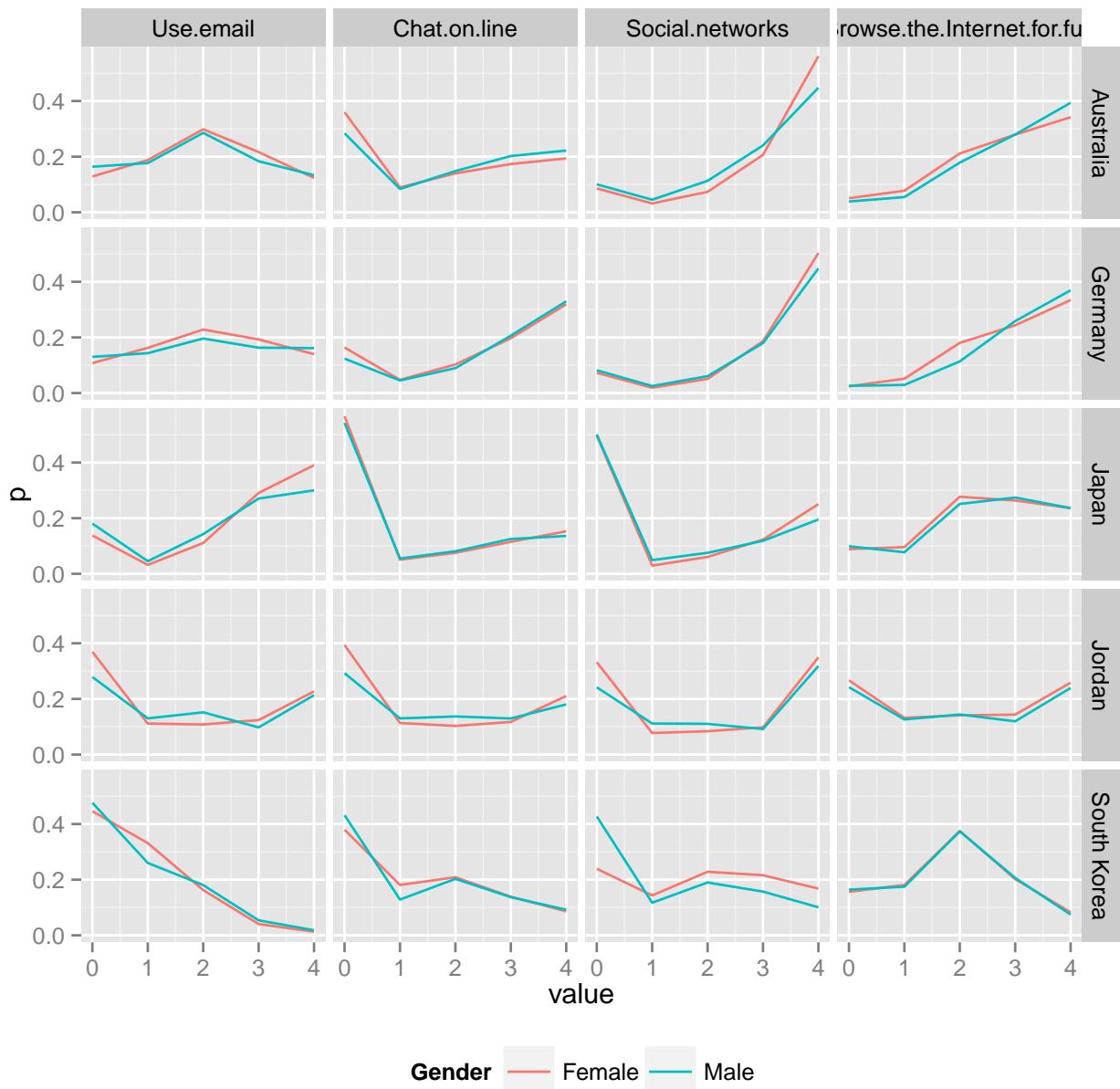
  name Gender     usage value     n
  (fctr) (fctr)   (fctr) (int) (int)
1 Australia Female Use.email      2  2112
2 Australia Female Use.email      3  1533
3 Australia Female Use.email      1  1329
4 Australia Female Use.email      0   911
5 Australia Female Use.email      4   875
6 Australia Female Use.email    NA   315
...
> internet.m.tb.n <- summarise(group_by(internet.m.tb, name, Gender, usage), tot=sum(n))
> internet.m.tb <- merge(internet.m.tb, internet.m.tb.n)
> internet.m.tb.p <- summarise(group_by(internet.m.tb, name, Gender, usage, value), p=n/tot)
> head(internet.m.tb.p)
Source: local data frame [6 x 5]
Groups: name, Gender, usage [1]
```

```

      name Gender      usage value      p
      (fctr) (fctr)    (fctr) (int)    (dbl)
1 Australia Female Use.email      0 0.12876325
2 Australia Female Use.email      1 0.18784452
3 Australia Female Use.email      2 0.29851590
4 Australia Female Use.email      3 0.21667845
5 Australia Female Use.email      4 0.12367491
6 Australia Female Use.email     NA 0.04452297
...
qplot(value, p, data=internet.m.tb.p, geom="line", color=Gender) +
  facet_grid(name~usage) +
  theme(legend.position="bottom")

```

The plot shows the fully rearranged and summarised data of proportions of the level of usage, separately by country and purpose, colored by gender.



- (d) What does the code block (shown below from the full set of commands) do?

[2 marks]

```
> internet.m.tb <- internet[,c(1,3,6,7,8,9)] %>%
  gather(key=usage, value=value, -name, -Gender) %>%
  group_by(name, Gender, usage, value) %>%
  tally(sort=TRUE)
```

This code takes the selected columns of the data, gathers it into long form, computes the count by country (name), gender and the type of internet use, and saves it into the data structure named ‘internet.m.tb’. (The use of the pipe command is efficient in not saving intermediate steps in the computation.)

- (e) Explain (in a few sentences) what the “Split-Apply-Combine” approach to working with large data is.

[2 marks]

When there are natural subsets in the data, upon which we want to compute different statistics, this approach will split the data into the subsets, operate on the chunks and join them back together when done. (It can be useful for large data also, and is how we would need to operate when data is stored in systems like hadoop.)

- (f) From the plot, which country and what type of usage exhibits the biggest difference between boys and girls? Describe the difference.

[2 marks]

South Korea, and social network use. Higher proportions of girls use social networks regularly.

- (g) Something that is missing from the display is a representation of the variation. Without this, it is hard to support statements about differences between the boys and girls internet usage. In a few sentences describe how you might calculate a measure that represents the variability estimates of proportions. (There is more than one way, but only discuss one.) And for a bonus point explain how you would add these to the plot using ggplot2 commands.

[2 marks]

A simple way would be to do a pointwise confidence interval using the standard error for proportions. A more computational approach would be to generate bootstrap confidence intervals or permutation intervals. Whatever way the confidence intervals were calculated these could be shown as a ribbon structure or vertical lines at discrete points (0, 1, 2, 3, 4). Bonus: You could use *geom_ribbon* or *geom_errorbar* to add them to the plot.

- (h) For ease of making comparisons, the “proximity principle” is important in constructing data plots. The arrangement used in this plot allows for the primary comparison of girls and boys usage. Suppose you really want to primarily compare usage between countries, and secondarily gender. How could you re-arrange the facets to facilitate that comparison?

[2 marks]

Interchange the colour (to be name) and the *facet_grid* element (to be Gender).

- (i) There is no indication of “NA” or missings in the plot. Why not? What happened to the missings?

[2 marks]

The missing were included in the summarisations, as a separate category. But they were dropped from the plot, because the “value” variable is an integer and by default NAs are simply not shown. If we had forced "value" to be treated as a factor, I think they would have been included.

[Total: 16 marks]

— END OF QUESTION 10 —

QUESTION 11

- (a) Entropy is commonly used to measure the impurity of a subset generated by a split. In the following data we want to use a one-sided purity measure based on entropy to measure the quality of the split, which is defined as this, for two classes (0,1):

$$\min\{(-\hat{p}_0^L \log \hat{p}_0^L - \hat{p}_1^L \log \hat{p}_1^L), (-\hat{p}_0^R \log \hat{p}_0^R - \hat{p}_1^R \log \hat{p}_1^R)\}$$

X_1	X_2	Class	
-5	1	0	(i) Calculate the impurity for a split on X_1 between 1 and 3. [2 marks]
-3	2	1	0, because the right bucket is pure.
0	-1	1	(ii) Would a split on X_2 between 1 and 2 yield a lower (or equal) impurity value? [2 marks]
1	0	1	Order the X_2 values: -1, 0, 1, 2, 7, 8, 10, and corresponding classes: 1, 1, 0, 1, 0, 0, 0; would give an impurity value bigger than 0.
3	7	0	
4	8	0	
6	10	0	

- (b) For a data set with 5 variables and 10000 cases, how many calculations of impurity would need to be made in order to decide on where the first split would be made? (Explain your working.)
[2 marks]

$$(10000-1)*5=49995$$

- (c) Look at the following tree fit:

```
> rpart(area~., data=subset(y, region==1, select=area:eicosenoic))
n= 323

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

1) root 323 120 South-Apulia (0.1734 0.0774 0.1115 0.6378)
  2) linoleic< 9.5e+02 119 63 Calabria (0.4706 0.2017 0.2521 0.0756)
    4) palmitoleic>=96 74 20 Calabria (0.7297 0.0135 0.1351 0.1216)
      8) arachidic< 80 67 13 Calabria (0.8060 0.0149 0.0448 0.1343)
        16) linolenic>=37 53 2 Calabria (0.9623 0.0000 0.0189 0.0189) *
        17) linolenic< 37 14 6 South-Apulia (0.2143 0.0714 0.1429 0.5714) *
      9) arachidic>=80 7 0 Sicily (0.0000 0.0000 1.0000 0.0000) *
    5) palmitoleic< 96 45 22 North-Apulia (0.0444 0.5111 0.4444 0.0000)
      10) stearic< 2.6e+02 23 2 North-Apulia (0.0000 0.9130 0.0870 0.0000) *
        11) stearic>=2.6e+02 22 4 Sicily (0.0909 0.0909 0.8182 0.0000) *
    3) linoleic>=9.5e+02 204 7 South-Apulia (0.0000 0.0049 0.0294 0.9657) *
```

- (i) How many cases are there in the terminal node labelled 11?

[2 marks]

22

- (ii) Calculate the error rate of the tree.

[2 marks]

$$21/323 = 0.065$$

[Total: 10 marks]

— END OF QUESTION 11 —

QUESTION 12

- (a) Linear and quadratic discriminant analysis: We fit a linear discriminant analysis model to the chocolates data with results as shown below.

```
> lda(Type~., data=subset(chocolates, select=Type:Protein), prior=c(0.5,0.5))
Call:
lda(Type ~ ., data = subset(chocolates, select = Type:Protein),
     prior = c(0.5, 0.5))
```

Prior probabilities of groups:

Dark Milk

0.5 0.5

Group means:

	Calories	CalFat	TotFat	SatFat	Chol	Na	Carbs	Fiber	Sugars	Protein
Dark	551	356	40	23	4.5	20	46	7.4	31	7.5
Milk	527	274	31	18	14.6	76	57	2.3	48	6.7

Coefficients of linear discriminants:

	LD1
Calories	-0.00045
CalFat	0.00129
TotFat	-0.06475
SatFat	-0.00721
Chol	0.02427
Na	0.01394
Carbs	-0.00382
Fiber	-0.17421
Sugars	0.01599
Protein	0.12223

- (a) Write down the classification rule.

[3 marks]

Classify the new observation to milk chocolate if $-0.00045\text{Calories} + 0.00129\text{CalFat} - 0.06475\text{TotFat} - 0.00721\text{SatFat} + 0.02427\text{Chol} + 0.01394\text{Na} - 0.00382\text{Carbs} - 0.17421\text{Fiber} + 0.01599\text{Sugars} + 0.12223 > 0$ else classify as dark chocolate.

- (b) Based on the plots of the data in the previous question would it be better to use LDA or QDA for this data, in order to satisfy the assumptions. Explain your answer.

[2 marks]

QDA, because the variances are not equal for the two groups. For example, Carbs has a smaller variance for milk chocolates than dark.

[Total: 5 marks]

— END OF QUESTION 12 —