STA130H1S – Winter 2021

Week 10 Problem Set - Sample Answers

S. Caetano and N. Moon

Instructions

How do I hand in these problems for the 11:59 a.m. ET, March 25 deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link:https://q.utoronto.ca/courses/206597/assignments/574836) by 11:59 a.m. ET, on March 25. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Question 1

Using data from the Gallup World Poll (and the World Happiness Report), we are interested in predicting which factors influence life expectancy around the world. These data are in the file `happinessdata_2017.csv`.

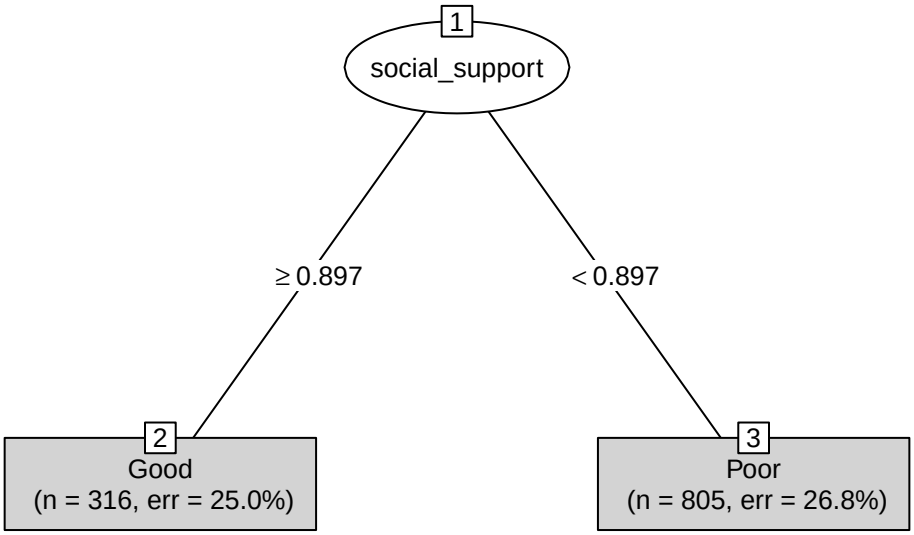
```
happiness2017 <- read_csv("happinessdata_2017.csv")
```

(a) Begin by creating a new variable called `life_exp_category` which takes the value "Good" for countries with a life expectancy higher than 65 years, and "Poor" otherwise.

```
life_exp_category <- happiness2017 %>%
  mutate(life_exp = case_when(life_exp > 65~ "Good",
                              life_exp <= 65 ~ "Poor"))
```

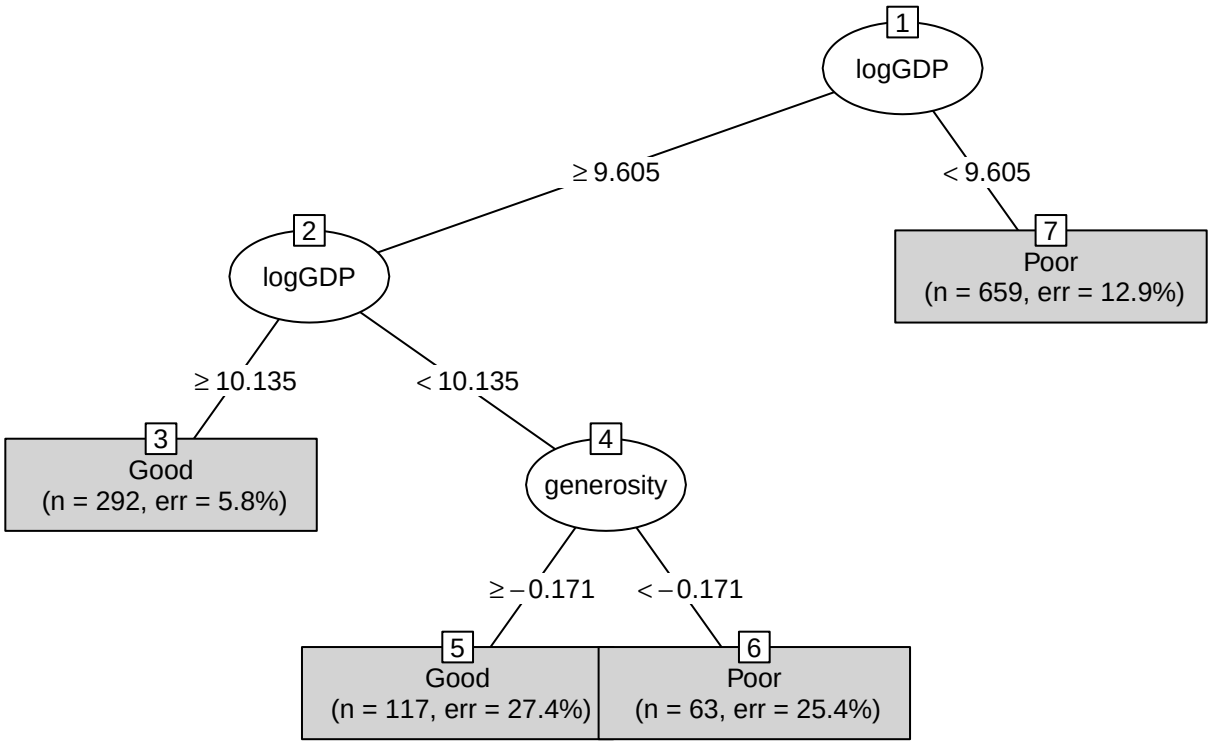
(b) Divide the data into training (80%) and testing (20%) datasets. Build a classification tree using the training data to predict which countries have `Good` vs `Poor` life expectancy, using only the `social_support` variable as a predictor. Use the last 3 digits of your student ID number for the random seed.

```
life_exp_category <- life_exp_category %>% rowid_to_column()

n <- nrow(life_exp_category)

set.seed(351)

train_ids <- sample(1:n, size = round(0.8*n))

#taking all the observations in the training sample (the 80% of data)
train <- life_exp_category %>%
  filter(rowid %in% train_ids)

#taking all the observations not in the training sample (the remaining 20%)
test <- life_exp_category %>%
  filter(!(rowid %in% train_ids))

tree1 <- rpart(life_exp ~ social_support,
               data = train)

plot(as.party(tree1), type = "simple",
     gp = gpar(cex = 0.8))
```



(c) Use the same training dataset created in (b) to build a second classification tree to predict which countries have good vs poor life expectancy, using `logGDP`, `social_support`, `freedom`, and `generosity` as potential predictors.

```
tree2 <- rpart(life_exp ~ logGDP + social_support + freedom + generosity,
               data = train)

plot(as.party(tree2), type = "simple",
     gp = gpar(cex = 0.8))
```



(d) Use the testing dataset you created in (b) to calculate the confusion matrix for the trees you built in (b) and (c). Report the sensitivity (true positive rate), specificity (true negative rate) and accuracy for each of the trees. Here you will treat "Good" life expectancy as a positive response/prediction.

Tree1:

```
tree_pred1 <- predict(tree1, newdata = test, type = "class")

t1.test <- table(tree_pred1, test$life_exp)
t1.test
```

```
##
## tree_pred1 Good Poor
##       Good   60   23
##       Poor   52  145
```

Overall accuracy:

```
(60+145)/(60+145+52+23)
```

```
## [1] 0.7321429
```

False-positive rate:

```
(52)/(145+52)
```

```
## [1] 0.2639594
```

False-negative rate:

```
(23)/(23+60)
```

```
## [1] 0.2771084
```

Tree2:

```
tree_pred2 <- predict(tree2, newdata = test, type = "class")

t2.test <- table(tree_pred2, test$life_exp)
t2.test
```

```
##
## tree_pred2 Good Poor
##       Good   82   20
##       Poor   30  148
```

Overall accuracy:

```
(82+148)/(82+148+20+30)
```

```
## [1] 0.8214286
```

False-positive rate:

```
(20)/(20+148)
```

```
## [1] 0.1190476
```

False-negative rate:

```
(30)/(30+82)
```

```
## [1] 0.2678571
```

(e) Fill in the following table using the tree you constructed in part (c). Does the fact that some of the values are missing (NA) prevent you from making predictions for the life expectancy category for these observations?

	logGDP	social_support	freedom	generosity	Predicted life expectancy category
Obs 1	9.56	0.74	NA	-0.25	Poor
Obs 2	10.1	0.84	0.80	-0.219	Poor
Obs 3	11.2	0.88	0.77	0.1	Good

The missing value for freedom in Obs 1 did not prevent me from making a prediction for the life expectancy category, because the classification tree I constructed in part (c) predicts based on the predictors "logGDP" and "generosity". The reason for it excluding the other categories is that the other predictors do not "maintain" the purity of the tree.

(f) In most cases, two classification trees will make different predictions for some new observations. Using the classification trees you built in parts (b) and (c), fill in the table below with values which would lead the specified predictions.

logGDP	social_suppo	freedom	generosity	Pred life expectancy category based on (b)	Pred life expectancy category based on (c)
11	0.801	0.2	0.1	**Poor**	**Good**
9	0.987	0.2	0.1	**Good**	**Poor**

Question 2

Two classification trees were built to predict which individuals have a disease using different sets of potential predictors. We use each of these trees to predict disease status for 100 new individuals. Below are confusion matrices corresponding to these two classification trees.

Tree A

	Disease	No disease
Predict disease	35	20
Predict no disease	3	42

Tree B

	Disease	No disease
Predict disease	23	4
Predict no disease	15	58

a) Calculate the accuracy, false-positive rate, and false negative rate for each classification tree. Here, a "positive" result means we predict an individual has the disease and a "negative" result means we predict they do not.

Tree A

- Overall accuracy:

```
(35+20)/(35+20+3+42)
```

```
## [1] 0.55
```

- False-positive rate:

```
(20)/(42+20)
```

```
## [1] 0.3225806
```

- False-negative rate:

```
(3)/(35+3)
```

```
## [1] 0.07894737
```

Tree B

- Overall accuracy:

```
(23+58)/(23+58+4+15)
```

```
## [1] 0.81
```

- False-positive rate:

```
(4)/(4+58)
```

```
## [1] 0.06451613
```

- False-negative rate:

```
(15)/(15+23)
```

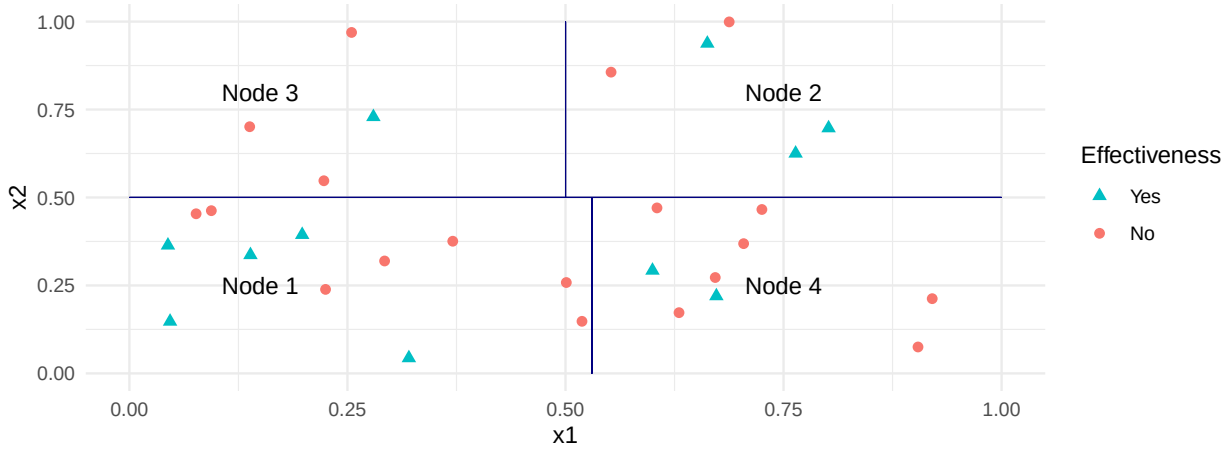
```
## [1] 0.3947368
```

b) Suppose the disease is very serious if untreated. Explain which classifier you would prefer to use.

I would rather choose tree A, because the False-negative rate is lower by approximately 32%, which would mean that fewer diseased individuals would be falsely predicted as not having the disease with this tree.

Question 3

Data was collected on 30 cancer patients to investigate the effectiveness (Yes/No) of a treatment. Two quantitative variables, $x_i \in (0, 1), i = 1, 2$, are considered to be important predictors of effectiveness. Suppose that the rectangles labelled as nodes in the scatterplot below represent nodes of a classification tree.



The diagram above is the geometric interpretation of a classification tree to predict drug effectiveness based on two predictors, x1 and x2. What is the predicted class of each node?

Node	Proportion of "Yes" values in each node	Prediction (assume we declare "effective" if more than 50% of the values are "Yes")
1	$5/12 = 0.4167$	Not effective
2	$2/9 = 0.222$	Not effective
3	$1/4 = 0.25$	Not effective
4	$3.5 = 0.6$	Effective

Part 2

The activity this week will focus on two areas, varying your register and describing classification trees. Prior to starting the assignment, it is highly recommended that you review the infographic available under Modules ⇒ Course Orientation ⇒ Writing skills resources.

There are three people that you can describe classification trees to:

- Using the 1,000 most commonly used words in the English language (see infographic for a list of the words and software to check if you used them). If you select this option, you are allowed to use words that are not part of the 1,000 most common words as long as it is a vocabulary word you are defining from the list. (e.g., you can use classification only if you are describing classification, and only when you are describing classification).
- A 10-year-old child
- A first-year undergraduate student who has not taken any statistical science courses, but that likes math

You must pick 2 of the people above to explain classification trees. You must incorporate at least 2 vocabulary words. For this assessment, you do NOT need to include an introduction or a conclusion. Rather, you should have two paragraphs, one for each scenario. The assignment cannot be more than 1 single spaced page in length. After one page the Teaching Assistant will stop reviewing your response.

Vocabulary

- Classification
- Prediction
- Predictor(s)
- Covariate(s)
- Independent variable(s)
- Dependent variable(s)
- Input(s)
- Output(s)
- Training set/sample
- Testing set/sample
- Fitting a model
- Confusion matrix
- Category
- Tree
- Terminal node
- Stopping rule
- Threshold
- True positive (sensitivity)
- True negative (specificity)
- False positive
- False negative

- Accuracy
- Classifier
- Node(s)
- Terminal Node
- Binary
- Split(ting)

Some things to keep in mind

- Try to not spend more than 20 minutes on your writing (plus the time to read the article).
- Aim for more than 200 but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).
- Be specific. A good principle when responding to a prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt).

Dear first-year undergraduate student who has not taken any statistical science courses, but that likes math, Classification trees are essentially decision trees. Based on data that is inputted, the tree will help guide you towards a decision or a prediction, and there should only be two possible predictions – positive or negative, which means it is a binary response. The tree is created by considering past or historical data linked to the possible predictions/outcomes. For instance, if there is a link between the amount of daily exercise an individual gets and that individual having heart issues, then the tree would be based on the amount of daily exercise the individual gets. And based on the amount, the tree would predict whether the individual has heart issues or not. But sometimes, of course, the trees can be wrong. They could give us a false positive, where the tree would predict that the individual has heart issues when, in reality, they do not. Or they could lead us to a false negative prediction, where the tree predicts that the individual does not have heart issues, when, in reality, they do.

Dear 10-year-old child, Classification trees help us make predictions based on people's answers to specific questions. It is kind of like a story game; you make choices and then those choices lead to a certain ending or conclusion. The story game might ask you a series of questions you say "yes" or "no" to and then they generate an ending for you based on those answers. For example, if you say "yes" to "Do you like animals?" and "no" to "Do you like the sea?", the story game might generate a story like "One day, your parents woke you up and surprised you with a road trip. They took you to a farm and there were so many cute and amazing farm animals!". Knowing that you like animals but do not like the sea, the story chooses to generate a story where you do not encounter underwater animals but instead animals on land. This is very similar to how classification trees help us make predictions. The trees give us a prediction based on our data or information.