STA130 – Winter 2024

Week 8 Problem Set

N. Moon and J. Speagle

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

Question 1

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. Replace the values of NULL below by your answers (use formulas to calculate your answers instead of typing the values in decimals, for example 1/3 instead of 0.333)

```
# Replace NULL by your answers here
# Use fractions (ex: 1 / (1+2), instead of 0.33)
Q1a_treeA_overall_accuracy <- (35 + 42) / (35 + 3 + 20 + 42)
Q1a_treeA_false_positive_rate <- 20 / (20 + 42)
Q1a_treeA_false_negative_rate <- 3 / (35 + 3)

Q1a_treeB_overall_accuracy <- (23 + 58) / (23 + 15 + 4 + 58)
Q1a_treeB_false_positive_rate <- 4 / (4 + 58)
Q1a_treeB_false_negative_rate <- 15 / (23 + 15)


# Print your results
Q1a_treeA_overall_accuracy
```

```
## [1] 0.77
```

```
Q1a_treeA_false_positive_rate
```

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

```
Q1a_treeA_false_negative_rate
```

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

```
Q1a_treeB_overall_accuracy
```

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

```
Q1a_treeB_false_positive_rate
```

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

```
Q1a_treeB_false_negative_rate
```

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

b) Suppose the disease is very serious if untreated. Explain which classifier you would prefer to use. You should make specific reference to the rates you calculated above in your answer (2-3 sentences).

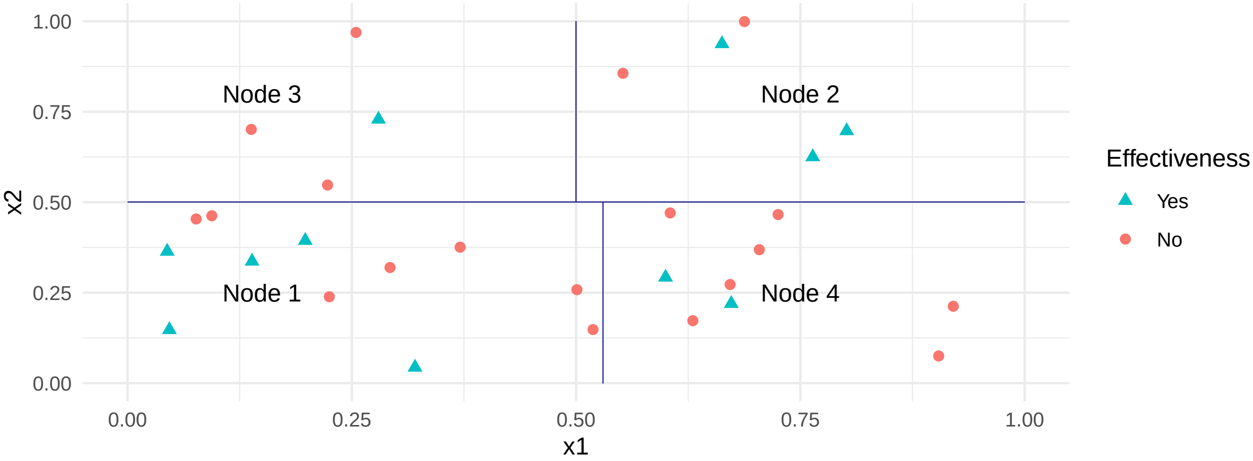
I would rather use Tree A. The false negative rate for tree A is 0.08 while the false negative rate for tree B is 0.39, so assuming there's 100 people who have the diseases then 8 people would die if we used tree a while 39 people would die if we used tree b.

c) Now suppose the treatment has very serious side effects. Explain which classifier you would prefer to use. You should make specific reference to the rates you calculated above in your answer (2-3 sentences).

If the disease is not very serious, I would rather use Tree B. If the false positive rate for tree A is 0.33, while the false positive rate for tree B is 0.06, so assuming there's 100 people who don't have the disesaes then 33 people would experience severe side effects if we used tree A while only 6 would if we use tree B.

Question 2

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 for each node, assuming that we predict "effective" if more than 50% of the values in a given node are "Yes"? Replace NULL by your answers

```
# For the proportion of yes, specify your answer in fractional form
# (e.g. 1/3 instead of 0.333)
Q2_node1_proportion_Yes <- 5 / (5 + 7)
Q2_node2_proportion_Yes <- 3 / (2 + 3)
Q2_node3_proportion_Yes <- 1 / (1 + 3)
Q2_node4_proportion_Yes <- 2 / (2 + 7)

# For each prediction, put either "effective" or "not effective"
# (be sure to use exactly the same spelling and to use quotation marks)
Q2_node1_prediction <- "not effective"
Q2_node2_prediction <- "effective"
Q2_node3_prediction <- "not effective"
Q2_node4_prediction <- "not effective"
```

Question 3

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, and add this new variable to the `happinessdata_2017` tibble

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

Q2a_happiness2017 <- happiness2017 # DO NOT CHANGE THIS LINE OF CODE
```

(b) Below, a chunk of R code divides the happiness data into two parts. Build a classification tree to predict which countries have `Good` vs `Poor` life expectancy, using only the `social_support` variable as a predictor, based only on the `happiness_training` tibble. Plot the resulting tree using `plot(as.party())`.

```
# DO NOT CHANGE THIS CODE CHUNK
set.seed(130) # DO NOT CHANGE THIS LINE

n <- nrow(happiness2017); # number of obs in the full dataset
n
```

```
## [1] 1420
```

```
# Use sample_n to randomly select 80% of observations from happiness2017
happiness_training <- sample_n(happiness2017, size = round(0.8*n))
# With anti_join(), we're putting all observations from happiness2017 which are
# NOT in happiness_training in the new happiness_testing tibble
# In other words, each observation from happiness_training is in exactly one of
# the two tibbles happiness_training and happiness_testing
happiness_testing <- anti_join(happiness2017, happiness_training)
```

```
# Fit the tree based on training data
tree <- rpart(life_exp_category ~ social_support, data = happiness_training)
```

(c) [Optional] Now build a second classification tree, again using the `happiness_training` tibble, to predict which countries have good vs poor life expectancy, with `logGDP`, `social_support`, `freedom`, and `generosity` as potential predictors. Plot your tree using `plot(as.party())`.

```
# Fit the tree based on training data
```

(d) In this part, you will use the tree you built in (b) to make predictions for the observations in the `happiness_testing` tibble. You will report the sensitivity (true positive rate), specificity (true negative rate) and accuracy. Here you will treat "Good" life expectancy as a positive response/prediction. Replace NULL with your answers below

```
predictions <- predict(tree, newdata=happiness_testing, type="class")
table(predictions, happiness_testing$life_exp_category)
```

```
##
## predictions Good Poor
##        Good   58   21
##        Poor   61  142
```

```
# Type your answers here, as functions (e.g. 1/3 instead of 0.33)
Q3d_accuracy <- (58 + 142) / (58 + 21 + 61 + 142)
Q3d_true_positive_rate <- 58 / (58 + 61)
Q3d_true_negative_rate <- 142 / (21 + 142)

# Print your results
Q3d_accuracy
```

```
## [1] 0.7092199
```

```
Q3d_true_positive_rate
```

```
## [1] 0.487395
```

```
Q3d_true_negative_rate
```

```
## [1] 0.8711656
```

(e) [Optional] You'll now use code similar to what was given in (d) to make predictions for the observations in the `happiness_testing` tibble using the tree you fit in (c). You will report the sensitivity (true positive rate), specificity (true negative rate) and accuracy. Again, you will treat "Good" life expectancy as a positive response/prediction. Replace NULL with your answers below

```
# Type your answers here, as functions (e.g. 1/3 instead of 0.33)
Q3e_accuracy <- NULL
Q3e_true_positive_rate <- NULL
Q3e_true_negative_rate <- NULL

# Print your results
Q3e_accuracy
```

```
## NULL
```

```
Q3e_true_positive_rate
```

```
## NULL
```

```
Q3e_true_negative_rate
```

```
## NULL
```

(f) 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	
Obs 2	10.1	0.84	0.80	-0.219	
Obs 3	11.2	0.88	0.77	0.1	

```r
# Repace NULL by "Poor" or "Good" (make sure to use the correct spelling and use quotation marks)
Q3f_obs1_prediction <- "Poor"
Q3f_obs2_prediction <- "Poor"
Q3f_obs3_prediction <- "Good"
```

(g) [Optional] 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_support`	`freedom`	`generosity`	Pred life expectancy category based on (b)	Pred life expectancy category based on (c)
				Poor	**Good**
				Good	**Poor**