

STOR 390: HOMEWORK 4

Riley Harper

March 18, 2024

This homework is designed to give you practice fitting a logistic regression and working with statistical/philosophical measures of fairness. We will work with the `titanic` dataset which we have previously seen in class in connection to decision trees.

Below I will preprocess the data precisely as we did in class. You can simply refer to `data_train` as your training data and `data_test` as your testing data.

#this is all of the preprocessing done for the decision trees lecture.

```
path <- 'https://raw.githubusercontent.com/guru99-edu/R-Programming/master/titanic_data.csv'
titanic <- read.csv(path)
head(titanic)
```

```
##   x pclass survived                name      sex
## 1 1      1        1      Allen, Miss. Elisabeth Walton female
## 2 2      1        1      Allison, Master. Hudson Trevor  male
## 3 3      1        0      Allison, Miss. Helen Loraine female
## 4 4      1        0      Allison, Mr. Hudson Joshua Creighton  male
## 5 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6 6      1        1      Anderson, Mr. Harry      male
##      age sibsp parch ticket      fare      cabin embarked
## 1      29      0      0 24160 211.3375      B5      S
## 2 0.9167      1      2 113781 151.55 C22 C26      S
## 3      2      1      2 113781 151.55 C22 C26      S
## 4      30      1      2 113781 151.55 C22 C26      S
## 5      25      1      2 113781 151.55 C22 C26      S
## 6      48      0      0 19952 26.55      E12      S
##                home.dest
## 1                St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                New York, NY
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
```

```
#replace ? with NA
replace_question_mark <- function(x) {
  if (is.character(x)) {
    x <- na_if(x, "?")
  }
  return(x)
}

titanic <- titanic %>%
  mutate_all(replace_question_mark)

set.seed(678)
shuffle_index <- sample(1:nrow(titanic))
head(shuffle_index)
```

```
## [1] 57 774 796 1044 681 920
```

```
titanic <- titanic[shuffle_index, ]
head(titanic)
```

```
##      x pclass survived      name
## 57      57      1      1  Carter, Mr. William Ernest
## 774    774      3      0    Dimic, Mr. Jovan
## 796    796      3      0    Emir, Mr. Farred Chehab
## 1044 1044      3      1  Murphy, Miss. Margaret Jane
## 681    681      3      0    Boulos, Mr. Hanna
## 920    920      3      0 Katavelas, Mr. Vassilios ('Catavelas Vassilios')
##      sex age sibsp parch ticket fare cabin embarked home.dest
## 57   male  36     1     2 113760   120 B96 B98      S Bryn Mawr, PA
## 774   male  42     0     0 315088  8.6625 <NA>      S      <NA>
## 796   male <NA>     0     0  2631  7.225 <NA>      C      <NA>
## 1044 female <NA>     1     0 367230  15.5 <NA>      Q      <NA>
## 681   male <NA>     0     0  2664  7.225 <NA>      C      Syria
## 920   male 18.5     0     0  2682  7.2292 <NA>      C      <NA>
```

```
library(dplyr)
# Drop variables
clean_titanic <- titanic %>%
  select(-c(home.dest, cabin, name, x, ticket)) %>%
  #Convert to factor level
  mutate(pclass = factor(pclass, levels = c(1, 2, 3), labels = c('Upper', 'Middle', 'Lower')),
         survived = factor(survived, levels = c(0, 1), labels = c('No', 'Yes'))) %>%
  na.omit()
#previously were characters
clean_titanic$age <- as.numeric(clean_titanic$age)
clean_titanic$fare <- as.numeric(clean_titanic$fare)
glimpse(clean_titanic)
```

```
## Rows: 1,043
## Columns: 8
## $ pclass    <fct> Upper, Lower, Lower, Middle, Lower, Middle, Lower, Lower, Upp~
## $ survived  <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, No, Yes, N~
## $ sex       <chr> "male", "male", "male", "male", "female", "female", "male", "~
## $ age       <dbl> 36.0, 42.0, 18.5, 44.0, 19.0, 26.0, 23.0, 28.5, 64.0, 36.5, 4~
## $ sibsp     <int> 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0~
## $ parch     <int> 2, 0, 0, 0, 0, 1, 0, 0, 2, 2, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ fare      <dbl> 120.0000, 8.6625, 7.2292, 13.0000, 16.1000, 26.0000, 7.8542, ~
## $ embarked  <chr> "S", "S", "C", "S", "S", "S", "S", "S", "S", "C", "S", "S", "S", "~
```

```
create_train_test <- function(data, size = 0.8, train = TRUE) {
  n_row = nrow(data)
  total_row = size * n_row
  train_sample <- 1: total_row
  if (train == TRUE) {
    return (data[train_sample, ])
  } else {
    return (data[-train_sample, ])
  }
}
data_train <- create_train_test(clean_titanic, 0.8, train = TRUE)
data_test <- create_train_test(clean_titanic, 0.8, train = FALSE)
```

1

Create a table reporting the proportion of people in the training set surviving the Titanic. Do the same for the testing set. Comment on whether the current training-testing partition looks suitable.

```
proportion_train <- round(table(data_train$survived) / nrow(data_train) * 100, 2)
proportion_test <- round(table(data_test$survived) / nrow(data_test) * 100, 2)
print("Proportion of people surviving the Titanic in the training set, ")
```

```
## [1] "Proportion of people surviving the Titanic in the training set, "
```

```
print(proportion_train)
```

```
##
##    No    Yes
## 60.19 39.81
```

```
print("Proportion of people surviving the Titanic in the testing set, ")
```

```
## [1] "Proportion of people surviving the Titanic in the testing set, "
```

```
print(proportion_test)
```

```
##
##    No    Yes
## 55.5 44.5
```

Since the proportions of the Yes, No data in training and testing sets are only ~5% different, there is an extremely close proportion of people surviving the Titanic in the training set when compared to the testing set.

2

Use the `glm` command to build a logistic regression on the training partition. `survived` should be your response variable and `pclass`, `sex`, `age`, `sibsp`, and `parch` should be your response variables.

```
logmod <- glm(survived ~ pclass + sex + age + sibsp + parch, data = data_train, family = binomial)
summary(logmod)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age + sibsp + parch,
##      family = binomial, data = data_train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.903165   0.409280   9.537  < 2e-16 ***
## pclassMiddle -1.291506   0.257421  -5.017 5.25e-07 ***
## pclassLower  -2.404084   0.262022  -9.175  < 2e-16 ***
## sexmale       -2.684206   0.200130 -13.412  < 2e-16 ***
## age           -0.036776   0.007494  -4.907 9.24e-07 ***
## sibsp         -0.395584   0.118587  -3.336 0.00085 ***
## parch          0.032494   0.111916   0.290 0.77155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1121.27  on 833  degrees of freedom
## Residual deviance:  757.87  on 827  degrees of freedom
## AIC: 771.87
##
## Number of Fisher Scoring iterations: 5
```

We would now like to test whether this classifier is *fair* across the sex subgroups. It was reported that women and children were prioritized on the life-boats and as a result survived the incident at a much higher rate. Let us see if our model is able to capture this fact.

3

Subset your test data into a male group and a female group. Then, use the `predict` function on the male testing group to come up with predicted probabilities of surviving the Titanic for each male in the testing set. Do the same for the female testing group.

```
male_test <- subset(data_test, sex == 'male')
male_predictions <- predict(logmod, newdata = male_test, type = "response")
```

```
female_test <- subset(data_test, sex == 'female')
female_predictions <- predict(logmod, newdata = female_test, type = "response")
```

4

Now recall that for this logistic *regression* to be a true classifier, we need to pair it with a decision boundary. Use an **if-else** statement to translate any predicted probability in the male group greater than 0.5 into **Yes** (as in Yes this individual is predicted to have survived). Likewise an predicted probability less than 0.5 should be translated into a **No**.

Do this for the female testing group as well, and then create a confusion matrix for each of the male and female test set predictions. You can use the **confusionMatrix** command as seen in class to expedite this process as well as provide you necessary metrics for the following questions.

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
male_outcomes <- ifelse(male_predictions > 0.5, 'Yes', 'No')
male_test$survived <- factor(male_test$survived, levels = c('No', 'Yes'))
male_confusion <- confusionMatrix(factor(male_outcomes), male_test$survived)
print(male_confusion$table)
```

```
##           Reference
## Prediction No Yes
##           No  93  28
##           Yes   4   4
```

```
female_outcomes <- ifelse(female_predictions > 0.5, 'Yes', 'No')
female_test$survived <- factor(female_test$survived, levels = c('No', 'Yes'))
female_confusion <- confusionMatrix(factor(female_outcomes), female_test$survived)
print(female_confusion$table)
```

```
##           Reference
## Prediction No Yes
##           No   4   2
##           Yes 15  59
```

5

We can see that indeed, at least within the testing groups, women did seem to survive at a higher proportion than men (24.8% to 76.3% in the testing set). Print a summary of your trained model and interpret one of the fitted coefficients in light of the above disparity.

```
summary(logmod)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age + sibsp + parch,
##      family = binomial, data = data_train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.903165   0.409280   9.537 < 2e-16 ***
## pclassMiddle -1.291506   0.257421  -5.017 5.25e-07 ***
## pclassLower  -2.404084   0.262022  -9.175 < 2e-16 ***
## sexmale      -2.684206   0.200130 -13.412 < 2e-16 ***
## age          -0.036776   0.007494  -4.907 9.24e-07 ***
## sibsp        -0.395584   0.118587  -3.336 0.00085 ***
## parch         0.032494   0.111916   0.290 0.77155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1121.27  on 833  degrees of freedom
## Residual deviance:  757.87  on 827  degrees of freedom
## AIC: 771.87
##
## Number of Fisher Scoring iterations: 5
```

The coefficient for sexmale is -2.684206, meaning that being male decreases the log odds of surviving the Titanic compared to being female by -2.684206, holding all other variables constant. This is demonstrated in a more interpretable way by viewing the right column of the above confusion table which is the total amount who survived (TP + FN). When we divide the right column by the total number in each of the tables we get 24.8% and 76.25% for men and women, respectively.

6

Now let's see if our model is *fair* across this explanatory variable. Calculate five measures (as defined in class) in this question: the Overall accuracy rate ratio between females and males, the disparate impact between females and males, the statistical parity between females and males, and the predictive equality as well as equal opportunity between females and males (collectively these last two comprise equalized odds). Set a reasonable ϵ each time and then comment on which (if any) of these five criteria are met.

```
# Male
AccM <- male_confusion$overall['Accuracy']
ProportionPosMales <- sum(male_confusion$table[, "Yes"]) / sum(male_confusion$table)
TotMale <- sum(male_confusion$table)
FPR_Males <- male_confusion$table["Yes", "No"] / sum(male_confusion$table["No",])
TPR_Males <- male_confusion$table["Yes", "Yes"] / sum(male_confusion$table[, "Yes"])

# Female
AccF <- female_confusion$overall['Accuracy']
ProportionPosFemales <- sum(female_confusion$table[, "Yes"]) / sum(female_confusion$table)
```

```

TotFemale <- sum(female_confusion$table)
FPR_Females <- female_confusion$table["Yes","No"] / sum(female_confusion$table["No",])
TPR_Females <- female_confusion$table["Yes","Yes"] / sum(female_confusion$table[, "Yes"])

# Overall Accuracy Rate Ratio (OARR)
### The ratio of accuracy rates between two groups such that a value of 1 indicates perfect fairness.
OARR <- AccF / AccM

# Disparate Impact (DI)
### Measures the ratio of positive outcomes received by the unprivileged group to the privileged group.
DI <- ProportionPosMales / ProportionPosFemales

# Statistical Parity (SP)
### The difference in the rate of positive outcomes between two groups such that a value of 0 indicates fairness.
SP <- abs(ProportionPosMales - ProportionPosFemales)

# Predictive Equality (PE)
### The difference in False Positive Rates (FPR) between groups such that a value of 0 indicates fairness.
PE <- abs(FPR_Males - FPR_Females)

# Equal Opportunity (EO)
### The difference in True Positive Rates (TPR) between groups such that a value of 0 indicates fairness.
EO <- abs(TPR_Males - TPR_Females)

print(paste("OARR:", OARR))

## [1] "OARR: 1.04729381443299"

print(paste("Disparate Impact:", DI))

## [1] "Disparate Impact: 0.325327233447706"

print(paste("Statistical Parity:", SP))

## [1] "Statistical Parity: 0.514437984496124"

print(paste("Predictive Equality:", PE))

## [1] "Predictive Equality: 2.46694214876033"

print(paste("Equal Opportunity:", EO))

## [1] "Equal Opportunity: 0.842213114754098"

```

The Overall Accuracy Rate Ratio is close to 1, indicating that the accuracy of the model is almost equally good for both males and females. An OARR of 1 would indicate perfect fairness in terms of accuracy between the two groups. Our model meets this criterion reasonably well, suggesting that it is fair in terms of overall accuracy across genders.

Disparate Impact (DI) measures the ratio of positive outcomes received by the unprivileged group to the privileged group, with values close to 1 indicating fairness. A DI value of 0.325 is far from 1, suggesting a

significant disparity in the rate of positive outcomes, with males being much less likely to receive positive outcomes compared to females. This indicates a strong bias against males in the model's predictions.

Statistical Parity (SP) measures the difference in the rate of positive outcomes between two groups. A value of 0 would indicate perfect fairness, meaning both groups have the same rate of positive outcomes. A SP of 0.514 indicates a significant difference in the rate of positive outcomes in favor of females, suggesting the model significantly favors females in predicting survival on the Titanic. This criterion for fairness is not met.

Predictive Equality looks at the difference in false positive rates (FPR) between groups. A value of 0 indicates that both groups have equal FPRs-indicating fairness. A PE of 2.467 suggests a significant difference in FPRs, with females likely having a higher FPR than males. This indicates the model is more likely to incorrectly predict survival for females compared to males, suggesting unfairness in predictive equality. This criterion for fairness is not met.

Equal Opportunity measures the difference in true positive rates (TPR) between groups. A value of 0 indicates that both groups have equal TPRs, which would be considered fair. An EO of 0.842 indicates that females have a much higher TPR than males, meaning the model is more likely to correctly predict survival for females than for males. This shows a bias in favor of females for correctly predicting survivors, indicating the model does not meet this fairness criterion.

6.0.1

It is always important for us to interpret our results in light of the original data and the context of the analysis. In this case, it is relevant that we are analyzing a historical event post-facto and any disparities across demographics identified are unlikely to be replicated. So even though our model fails numerous of the statistical fairness criteria, I would argue we need not worry that our model could be misused to perpetuate discrimination in the future. After all, this model is likely not being used to prescribe a preferred method of treatment in the future.

7

Even so, provide a *philosophical* notion of justice or fairness that may have motivated the Titanic survivors to act as they did. Spell out what this philosophical notion or principle entails?

The actions of the Titanic survivors in prioritizing women and children for lifeboat spots can be philosophically grounded in the notion of care ethics. Normative ethical theories propose standards for what actions are morally right or wrong, as opposed to descriptive ethics, which simply describes how people behave or what moral beliefs they hold. The framework of care ethics emphasizes the moral importance of responding to the needs of individuals who are in a position of vulnerability or dependence. Care ethics is a normative ethical theory that highlights the significance of interpersonal relationships and the virtues of empathy, compassion, and caring. It suggests that moral actions stem from the understanding of and response to the needs of others, especially those who are dependent or vulnerable. In the context of the Titanic disaster, this principle can be seen in the collective decision to prioritize those perceived as most vulnerable (women and children) for evacuation. While this ethical principle provides a moral grounding for the actions taken during the Titanic disaster, it also highlights the complexity of making any sort of ethical decision in life-and-death situations. These decisions to protect women alongside children reflects the values and norms of the time-March 31, 1909-when many of the gender reforms were not yet in place. Although it still may be the case that women and children are prioritized if a similar situation were to occur today, I believe it is more likely than it was that the distribution of adult men and women saved is closer to equal than it was in 1909.

As discussed more recently in class, the veil of ignorance-positied by John Rawl in A Theory of Justice-suggests that we should imagine ourselves behind a "veil of ignorance" when making decisions related to social justice or resource allocation without knowledge of one's position within society. While behind the veil,

we know nothing of our ethnicity, social status, gender, or idea of how to lead a good life and instead will tend to help those who are most disadvantaged in case this becomes our place in society outside of the veil. Help comes in many forms, of course, such as resources, rights, and freedom which Rawls calls “Primary Goods.” Related to the Titanic, someone looking behind a veil of ignorance may seek to help woman more than men on-board, since although men ultimately had lower odds of survival, in the social paradigms of 1909 women were typically withheld many rights leaving them as the unprivileged class in the general sense, although statistically they were the privileged group in the scenario of Titanic survival.