Quiz 3

Modern Data Mining/ Linda Zhao

April 26, 2022

Contents

This is an open book, 30-minute quiz. But we will keep the submission window up to 40 minutes. Choose the correct answer(s). There might be more than one right answers in some questions. No calculations are needed.

On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others. Our goal is to predict survived passengers based on their information.

The `titanic` dataset contains information of 705 real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including:

- `Survived`: Survival indicator (0 = No; 1 = Yes)
- `Pclass`: Passenger class (1 = 1st, Upper; 2 = 2nd, Middle; 3 = 3rd, Low)
- `Sex`
- `Age`
- `SibSp`: The number of siblings/spouses aboard
- `Parch`: The number of parents/children aboard
- `LogFare`: The log of passenger fare
- `Embarked`: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

```
url <- 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
titanic <- read_csv(url) %>%
    select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked) %>%
    drop_na() %>% mutate(across(c(Pclass, Sex, Embarked), factor)) %>%
    filter(Fare > 0) %>% mutate(LogFare = log(Fare)) %>% select(-Fare)
head(titanic)
```

```
## # A tibble: 6 x 8
##   Survived Pclass Sex      Age SibSp Parch Embarked LogFare
##      <dbl> <fct>  <fct>  <dbl> <dbl> <dbl> <fct>      <dbl>
## 1        0 3      male      22     1     0 S           1.98
## 2        1 1      female    38     1     0 C           4.27
## 3        1 3      female    26     0     0 S           2.07
## 4        1 1      female    35     1     0 S           3.97
## 5        0 3      male      35     0     0 S           2.09
## 6        0 1      male      54     0     0 S           3.95
```

1 Exploratory Data Analysis (2 points)

Q1. (1 point) We first explore how passengers' survival is related to their fare.

```
ggplot(titanic, aes(x=LogFare, y=Survived)) +
  geom_jitter(height = .05, aes(color = factor(Survived)))
```

True or false? Using `geom_jitter()` would be better than `geom_point()` since many passengers have similar value of LogFare and Survived.

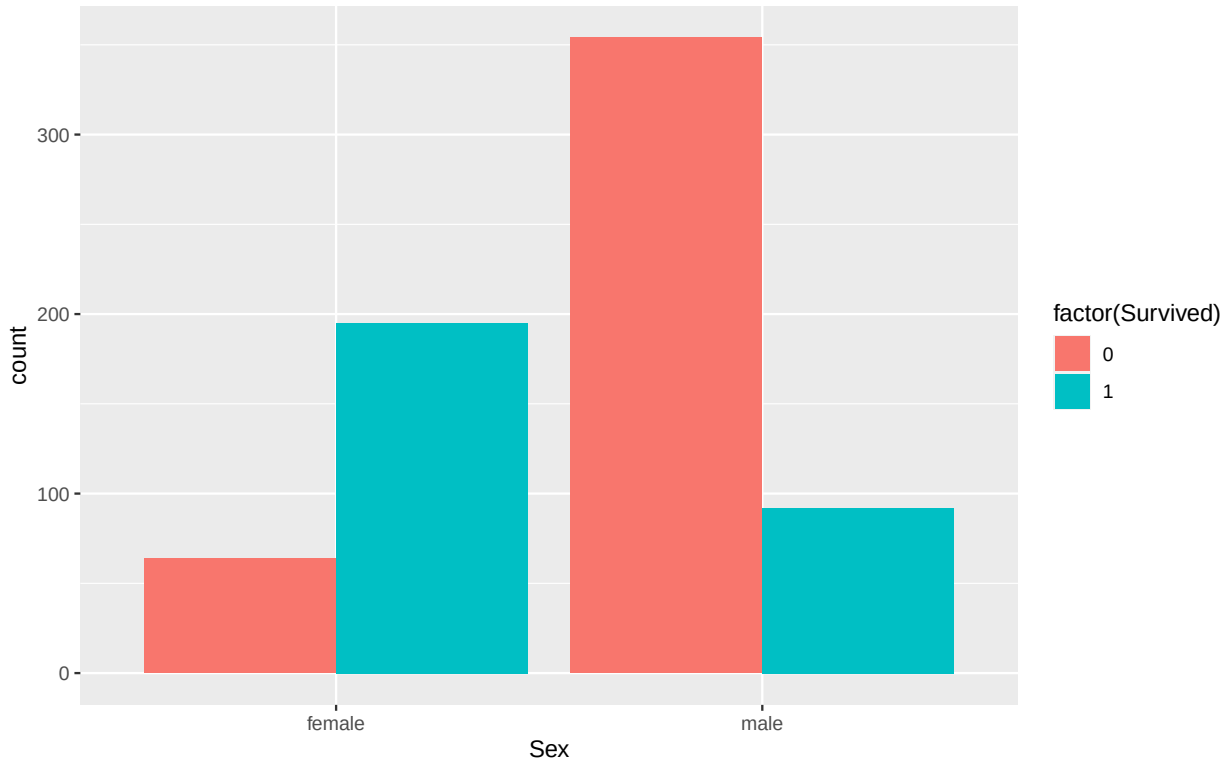
(A) TRUE

(B) FALSE

The answer is (A)

Q2. (1 point) We next focus on the relation between survival and sex.

```
ggplot(titanic) +
  geom_bar(aes(x = Sex, fill = factor(Survived)), position = "dodge")
```



Based on the above barplot, choose the correct answer(s).

(A) The survival rate of females is high than that of males.

(B) The survival rate of females is lower than that of males.

(C) More information is needed.

The answer is (A)

2 Logistic Regression (18 points)

2.1 `fit0`: `glm(Survived ~ LogFare)`

Q3. (4 points) We first investigate the relationship between the fare and the survival rate. We run a simple logistic regression of `Survived` on `LogFare`.

```
fit0 = glm(Survived ~ LogFare, data = titanic, family = binomial)
```

Which equation(s) correctly describe the model `fit0`?

(A) $\text{Survived} = \beta_0 + \beta_1 \times \text{LogFare}$

(B) $\mathbb{P}(\text{Survived} = 1 \mid \text{LogFare}) = \beta_0 + \beta_1 \times \text{LogFare}$

(C) $\log \frac{\mathbb{P}(\text{Survived}=1|\text{LogFare})}{\mathbb{P}(\text{Survived}=0|\text{LogFare})} = \beta_0 + \beta_1 \times \text{LogFare}$

(D) $\mathbb{P}(\text{Survived} = 1 \mid \text{LogFare}) = \frac{e^{\beta_0+\beta_1 \times \text{LogFare}}}{1+e^{\beta_0+\beta_1 \times \text{LogFare}}}$

The answers are (C) and (D)

Q4. (3 points) Here is the summary result for `fit0`.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.76	0.29	-9.48	0.00
LogFare	0.79	0.09	8.54	0.00

Choose the correct statement(s). The fitted coefficients in `fit0` are obtained by

(A) Maximizing the log likelihood $\mathcal{L}(\beta_0, \beta_1 \mid \text{data})$.

(B) Maximizing the probability $\mathbb{P}(\text{the outcome of data})$.

(C) Minimizing the cross entropy.

The answers are (A), (B) and (C)

Q5. (1 point) Based on `fit0`, we classify a passenger as "Survived" if $\mathbb{P}(\text{Survived} = 1 \mid \text{LogFare}) > 1/2$. What is the linear classification boundary?

(A) $-2.76 + 0.79 \times \text{LogFare} > \log(1) = 0$

(B) $-2.76 + 0.79 \times \text{LogFare} > 1/2$

The answer is (A)

2.2 `fit1`: full model

Q6. (1 point) We next run a logistic regression of `Survived` on all variables as `fit1`.

```
fit1 = glm(Survived ~ ., data = titanic, family = binomial)
print(xtable(summary(fit1), digits = 3))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.420	1.090	4.057	0.000
Pclass2	-1.278	0.403	-3.172	0.002
Pclass3	-2.489	0.506	-4.918	0.000
Sexmale	-2.641	0.225	-11.752	0.000
Age	-0.044	0.008	-5.261	0.000
SibSp	-0.372	0.142	-2.631	0.009
Parch	-0.058	0.137	-0.423	0.672
EmbarkedQ	-0.797	0.604	-1.321	0.187
EmbarkedS	-0.386	0.277	-1.391	0.164
LogFare	0.047	0.236	0.201	0.841

According to the summary table of `fit1`, which value would decrease by 0.37 if `SibSp` increases by one holding all other variables constant?

(A) The probability of survival

(B) The odds of survival

(C) The log odds of survival

The answer is (C)

Q7. (2 points)

Based on the summary table of `fit1`, choose the correct statement(s).

(A) The chance of survival is highest for people in `Pclass1`

(B) The chance of survival is higher for female than male, given all other variables are the same

The answer is (B)

Q8. (1 point) Based on the summary table of `fit1`, we fail to reject the hypothesis that the probabilities of survival across the 3 classes of `Pclass` are the same at 0.001 level because `Pclass2` is not significant at 0.001 level, controlling for all other variables. True or false?

(A) TRUE

(B) FALSE

The answer is (B)

Q9. (1 point) We are interested in testing the following hypothesis:

$$H_0 : \beta_{\text{Parch}} = \beta_{\text{Embarked}} = \beta_{\text{LogFare}} = 0.$$

True or false? We fail to reject the null hypothesis at 0.05 level because each of the p-value `Parch`, `Embarked` and `LogFare` is larger than 0.05.

(A) TRUE

(B) FALSE

The answer is (B)

Q10. (1 point) We build our first classifier using `fit1` model with threshold $1/2$. We can summarize how well this rule works by confusion matrix.

```
fit1_pred = ifelse(fit1$fitted > 1/2, 1, 0)
fit1_cfm = table(fit1_pred, titanic$Survived)
fit1_cfm
```

```
##
## fit1_pred   0   1
##         0 358  79
##         1  60 208
```

Here, the row indicates $\widehat{y}$ and the column indicates y. What is the sensitivity of this rule based on the confusion matrix?

(A) $358/(358+60)$

(B) $208/(79+208)$

(C) $60/(358+60)$

The answer is (B)

Q11. (1 point) Based on the confusion matrix above, what is the mis-classification error?

(A) $(60+79)/(358+60+79+208)$

(B) $(79+60)/(358+60+79+208)$

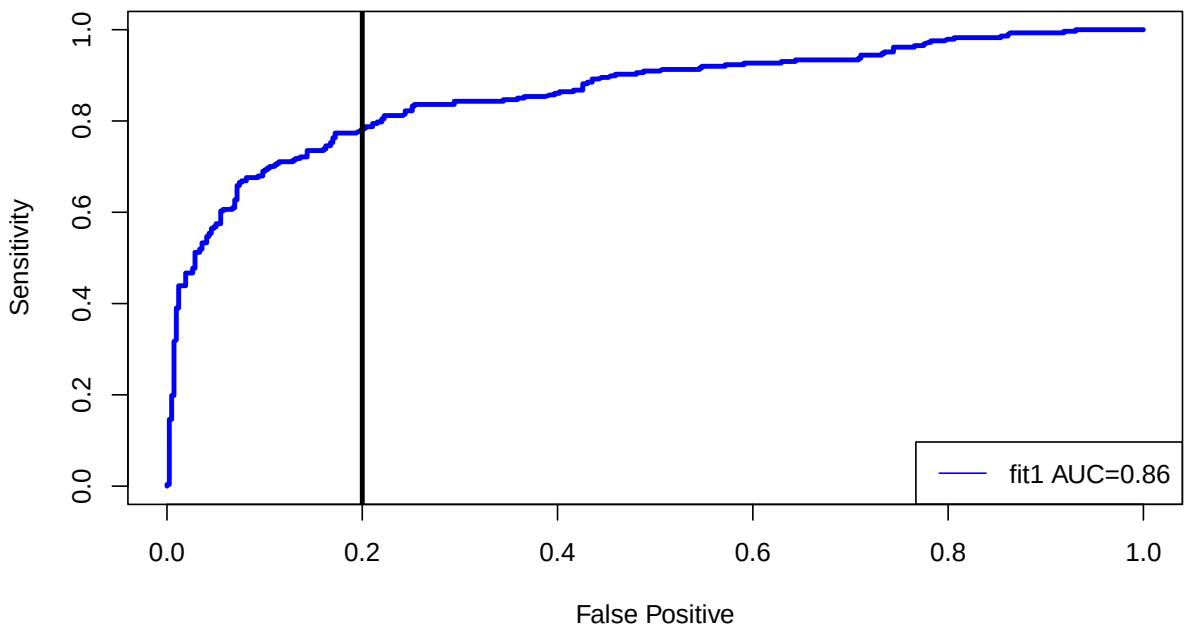
(C) $60/(358+60)+79/(79+208)$

The answer is (B)

Q12. (3 points) We will evaluate the performance of `fit1` using ROC curve.

```
fit1.roc = roc(titanic$Survived, fit1$fitted)

plot(1-fit1.roc$specificities,
     fit1.roc$sensitivities, col="blue", lwd=3, type="l",
     xlab="False Positive",
     ylab="Sensitivity")
legend("bottomright",
       c(paste0("fit1 AUC=", round(fit1.roc$auc,2)) ),
       col=c("blue"),
       lty=1)
abline(v=.2, lwd = 3)
```



Based on the above plot, choose the correct answer(s).

(A) The ROC curve is drawn by computing the false positive rate and sensitivity with different thresholds.

(B) The classifier will predict the passengers' survival accurately in approximately 86% of the data because AUC is 0.86.

(C) For the thresholds with false positive less than 0.2, the sensitivity will be less than 0.8.

The answers are (A) and (C)

3 LASSO in Logistic Regression (2 points)

Q13. (1 point) We next run LASSO in logistic regression from the full model to select variables using `lambda.1se`.

```
set.seed(30301566)
X = model.matrix(Survived~., data=titanic)[,-1]
Y = pull(titanic, Survived)

fit2.cv = cv.glmnet(X, Y, alpha=1, family="binomial", nfolds = 10,
                    type.measure = "deviance")
# plot(fit3.cv)
coef.1se = coef(fit2.cv, s="lambda.1se")
coef.1se = coef.1se[which(coef.1se!=0),]
coef.1se
```

```
(Intercept)     Pclass3     Sexmale         Age       SibSp   EmbarkedS
      1.223      -1.012      -2.050      -0.012      -0.103      -0.155
    LogFare
      0.216
```

True or false? The selected variables could change if we don't set a seed.

(A) TRUE

(B) FALSE

The answer is (A)

Q14. (1 point) We refit the logistic regression with the variables chosen from LASSO as `fit2`.

```
fit2 = glm(Survived ~ Sex+Age+SibSp+LogFare+Pclass+Embarked,
           data = titanic, family = binomial)
```

True or false? All variables in `fit2` will be significant at 0.05 level since they are chosen from LASSO, i.e., the p-value of each variable will be less than 0.05.

(A) TRUE

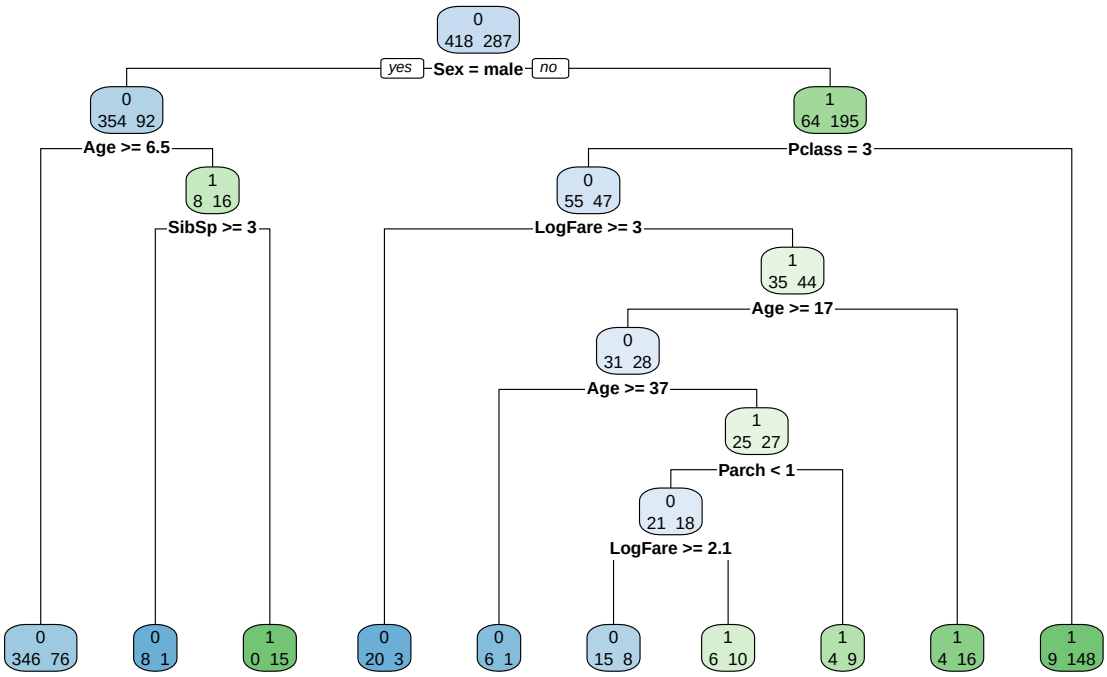
(B) FALSE

The answer is (B)

4 Tree and Random Forest (4 points)

Q15. (1 point) We fit a single tree model as `fit3` using `rpart()` package.

```
fit3 = rpart(Survived ~ ., data=titanic, method='class')
rpart.plot(fit3, extra=1, fallen.leaves=TRUE)
```



According to the fitted model, what is the predicted survival of the following passenger?

(Hint: the left branch is "yes" and the right branch is "no". In each node, the number on the left is the number of 0's (dead) and the number on the right is the number of 1's (survived). The predicted value is shown on the top of each node and is based on the majority vote.)

- `Pclass`: 3 (3rd, Low)
- `Sex`: female
- `Age`: 20
- `SibSp`: 2
- `Parch`: 2
- `LogFare`: 1
- `Embarked`: C (Cherbourg)

(A) 0

(B) 1

The answer is (B)

Q16. (1 point) Based on `fit3`, choose the correct answer(s).

(A) There are 10 different predicted probabilities of survival.

(B) There are 19 different predicted probabilities of survival.

The answer is (A)

Q17. (1 point) We build a random forest to predict survival using all the variables as `fit4`.

```
fit4 = randomForest(as.factor(Survived)~., data=titanic, mtry=3,
                    ntree=1000, importance = TRUE)
```

Choose the best description that describes how this model was built.

(A) Bagging 1000 bootstrap trees with 3 features randomly sampled for each tree.

(B) Randomly sampling 3 features and bagging 1000 bootstrap trees.

(C) Bagging 1000 bootstrap trees with each split of each tree based on the best variable chosen out 3 randomly sampled features.

The answer is (C)

Q18. (1 point) True or false? The training error of `fit4` (random forest) is always smaller than that of `fit3` (single tree).

(A) TRUE

(B) FALSE

The answer is (B)

5 Neural Network (4 points)

Q19. (4 points) We finally build a neural network to predict survival using all the variables.

```
model = keras_model_sequential() %>%
  layer_dense(units = 16, activation = "relu", input_shape = c(9)) %>%
  layer_dense(units = 8, activation = "relu") %>%
  layer_dense(units = 2, activation = "softmax")
```

Which statement(s) are correct?

(A) The model is same as logistic regression model.

(B) The model is built on one layer with 48 neurons.

(C) There are two layers with 16 and 8 neurons in each layer respectively. Each neuron is created by taking the linear combination of all the neurons from the previous layer then apply the relu function.

(D) There are two layers with 16 and 8 neurons in each layer respectively. Each neuron is created by taking the linear combination of all the neurons from the previous layer.

The answer is (C)