Domanda 7

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

What is the correct code to prune the following tree according to the one-standard-deviation rule?

```
library(dplyr)
library(rpart)
path <- 'https://raw.githubusercontent.com/guru99-edu/R-Programming/master/titanic_data.csv'
titanic <-read.csv(path)
head(titanic)
cleantitanic <- titanic %>%
  select(-c(home.dest, cabin, name, x, ticket)) %>%
  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')),
         age = as.numeric(age), fare = as.numeric(fare))
model <- rpart(survived ~ ., cleantitanic, cp =0)
printcp(model)
```

Scegli un'alternativa:

a.

```
min.cp.table <- model$cptable[which.min(model$cptable[,"xerror"]),]
min.cp_sd <- min.cp.table["xerror"]+min.cp.table["xstd"]
cptable <- model$cptable[model$cptable[,"xerror"]>min.cp_sd,]
best.cp <- cptable[which.min(cptable[,"nsplit"]),]
tree.pruned <- prune(model, cp=best.cp["CP"])
```

b.

```
min.cp.table <- model$cptable[which.min(model$cptable[,"CP"]),]
tree.pruned <- prune(model, cp=min.cp.table["CP"])
```

c.

```
min.cp.table <- model$cptable[which.min(model$cptable[,"xerror"]),]
min.cp <- min.cp.table["xerror"]
cptable <- model$cptable[model$cptable[,"xerror"]<=min.cp,]
tree.pruned <- prune(model, cp=cptable["CP"])
```

d.

```
min.cp.table <- model$cptable[which.min(model$cptable[,"xerror"]),]
min.cp_sd <- min.cp.table["xerror"]+min.cp.table["xstd"]
cptable <- model$cptable[model$cptable[,"xerror"]<=min.cp_sd,]
best.cp <- cptable[which.min(cptable[,"nsplit"]),]
tree.pruned <- prune(model, cp=best.cp["CP"])
```

✓

La risposta corretta è:

```
min.cp.table <- model$cptable[which.min(model$cptable[,"xerror"]),]
min.cp_sd <- min.cp.table["xerror"]+min.cp.table["xstd"]
cptable <- model$cptable[model$cptable[,"xerror"]<=min.cp_sd,]
best.cp <- cptable[which.min(cptable[,"nsplit"]),]
tree.pruned <- prune(model, cp=best.cp["CP"])
```

Domanda 5

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

The probability to be admitted to a college is modelled using a logistic regression. Two explanatory variables are used: the Grade Point Average (GPA - continuous) and the rank of the previous school (rank - factor).

```
library(arm)
df <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
fit <- glm(admit ~ gpa+as.factor(rank), family="binomial", data = df)
```

```
coef.df <- coefficients(fit); gpa.seq <- seq(0,4,by = 0.1)
```

The correct commands to plot the logit of the model varying `gpa` and the 4 `rank` levels are:

Scegli un'alternativa:

a. None of the options

b.

```
plot(gpa.seq, cbind(1,gpa.seq,1,0,0)%*%coef.df, type = "l", ylab="", ylim = c(-6,1))
lines(gpa.seq, cbind(1,gpa.seq,0,1,0)%*%coef.df, type = "l")
lines(gpa.seq, cbind(1,gpa.seq,0,0,1)%*%coef.df, type = "l")
```

c.

```
plot(gpa.seq, cbind(1,gpa.seq,0,0,0)%*%coef.df, type = "l", ylab="", ylim = c(-6,1))
lines(gpa.seq, cbind(1,gpa.seq,1,0,0)%*%coef.df, type = "l")
lines(gpa.seq, cbind(1,gpa.seq,0,1,0)%*%coef.df, type = "l")
lines(gpa.seq, cbind(1,gpa.seq,0,0,1)%*%coef.df, type = "l")
```

✓

d.

```
plot(gpa.seq, invlogit(cbind(1,gpa.seq,0,0,0)%*%coef.df), type = "l", ylab="", ylim = c(0,1))
lines(gpa.seq, invlogit(cbind(1,gpa.seq,1,0,0)%*%coef.df), type = "l")
lines(gpa.seq, invlogit(cbind(1,gpa.seq,0,1,0)%*%coef.df), type = "l")
lines(gpa.seq, invlogit(cbind(1,gpa.seq,0,0,1)%*%coef.df), type = "l")
```

La risposta corretta è:

```
plot(gpa.seq, cbind(1,gpa.seq,0,0,0)%*%coef.df, type = "l", ylab="", ylim = c(-6,1))
lines(gpa.seq, cbind(1,gpa.seq,1,0,0)%*%coef.df, type = "l")
lines(gpa.seq, cbind(1,gpa.seq,0,1,0)%*%coef.df, type = "l")
lines(gpa.seq, cbind(1,gpa.seq,0,0,1)%*%coef.df, type = "l")
```

Domanda 1

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

Which of the following statements is false?

Scegli un'alternativa:

- a. A Random Forest is an example of Bagging using decision trees as the base models.
- b. Reducing the number of iterations could control for over-fitting in Boosting algorithms.
- c. Stacking/Blending involves fitting models and then using the predictions of those models as features in subsequent models along with the rest of the data.
- d. Bagging results in a high variance model because we train lots of models independently and take the average of their results. ✓

La risposta corretta è: Bagging results in a high variance model because we train lots of models independently and take the average of their results.

Domanda 2

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

Which of the following sentences describes cross-validation

Scegli un'alternativa:

- a. Split the data randomly into two groups and use one group to train algorithm and the rest for evaluating it (validation)
- b. Split the data randomly into K groups and use only $K-1$ groups to train the algorithm and the remaining for validation
- c. Split the data randomly into K groups and use $K-1$ groups to train the algorithm and the remaining for evaluation. Repeat this leaving out each time one of the group. Combine the evaluation obtained on the group left out in the K repetitions. ✓
- d. The same data used to train the algorithm are used for its evaluation

La risposta corretta è: Split the data randomly into K groups and use $K-1$ groups to train the algorithm and the remaining for evaluation. Repeat this leaving out each time one of the group. Combine the evaluation obtained on the group left out in the K repetitions.

Domanda 3

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

Assume that a logistic regression model is used for the probability of buying a new policy. The coefficient for the variable GENDER (which takes on the value 1 if Male and 0 otherwise) is estimated to be equal to 2 and is statistically significant.

Which of the following statement is false:

Scegli un'alternativa:

- a. Males customers are more likely to buy the policy
- b. The logarithm of the ratio between the probabilities of buying the policy for Male and Female is 2
- c. The probability of buying the policy for Male is more than 7 times larger than the same probability for Females.
- d. The ratio between the probabilities of buying the policy for Male and Female is 2 ✓

La risposta corretta è: The ratio between the probabilities of buying the policy for Male and Female is 2

Domanda 4

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

To predict the number of online purchased items Y a GLM is used. Assume that one among the observed variables is TIME which indicates for how long the user has been connected to the website. What do you think is the most appropriate way to use this information?

Scegli un'alternativa:

- a. Consider the ratio between the number of purchased items and the Time (Y/TIME) and use a logistic regression model for its prediction
- b. Insert the variable among the possible predictors.
- c. Use a GLM with Poisson link and insert the logarithm of TIME as an offset ✓
- d. Consider the ratio between the number of purchased items and the TIME (Y/TIME) and use a linear model for its prediction

La risposta corretta è: Use a GLM with Poisson link and insert the logarithm of TIME as an offset

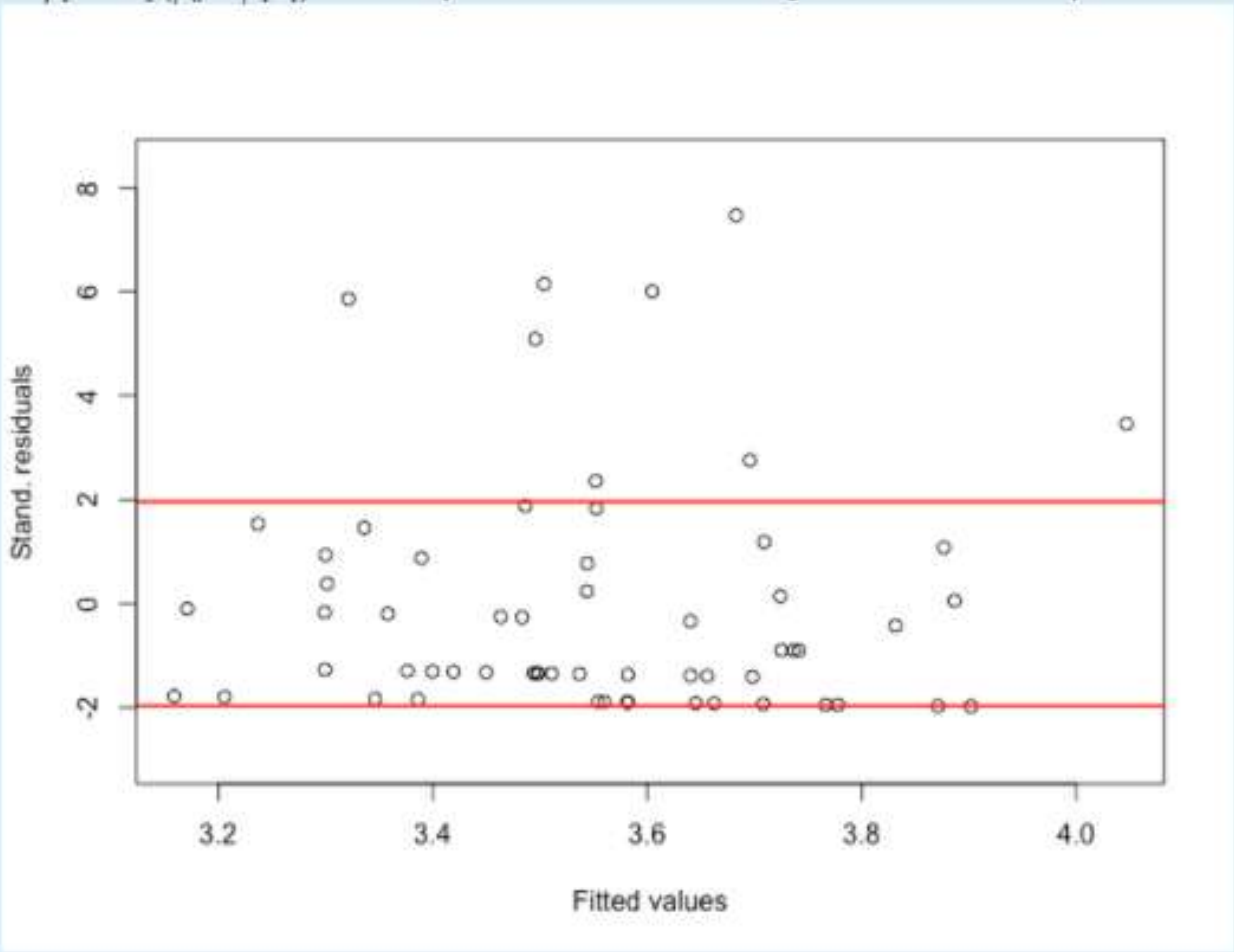
Domanda 6

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Contrassegna domanda

Assume that a Poisson regression for the number of daily phone calls $y_1, \ldots, y_n$ registered for n days is assumed, and that the only covariate x_i is the global external temperature in celsius degrees at the i-th day, thus $\mu_i = \exp(\beta_0 + \beta_1 x_i)$. What can you conclude from the following standardized residuals plot?



Scegli un'alternativa:

- a. A quadratic term should be added to the model.
- b. The residuals are randomly distributed between -1.96 and 1.96, thus there is not clue of poor fitting.
- c. The coefficient β_1 is not statistically significant.
- d. There is possible overdispersion in the data. ✓

La risposta corretta è: There is possible overdispersion in the data.