1. Use Naïve Bayes Classifier to classify the objects [25]

We conducted a survey to collect people’s daily diets and try to build a model to predict whether their diets result in healthy conditions or not. The final results could be Yes, No, Unsure

|  |  |  |  |
| --- | --- | --- | --- |
| **Breakfast** | **Lunch** | **Dinner** | **Healthy?** |
| Ham | Carnivorous | Beef | Y |
| Milk | Carnivorous | Beef | N |
| Bread | Veggie | Pork | U |
| Bread | Veggie | Veggie | Y |
| Ham | Veggie | Veggie | Y |
| Bread | Carnivorous | Beef | N |
| Ham | Veggie | Pork | N |
| Milk | Veggie | Pork | U |
| Milk | Carnivorous | Veggie | U |
| Ham | Carnivorous | Pork | ? |

Manually solve this problem by calculations, do not use R

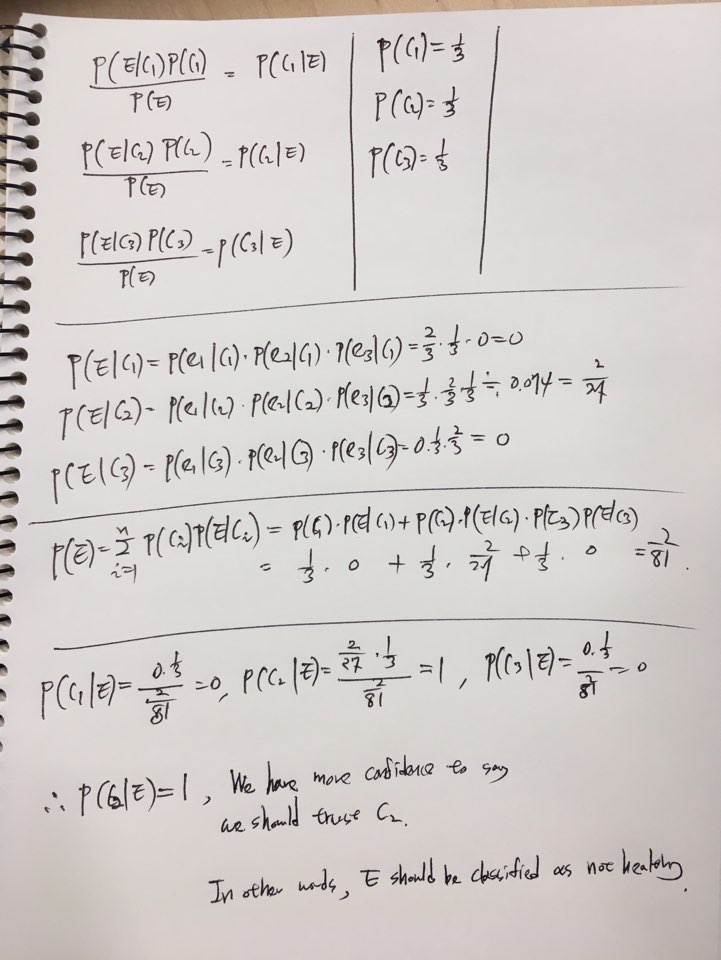
Show your method step by step

Class:

* C1: Healthy=Y
* C2: Healthy=N
* C3: Healthy=U

E

* e1: Breakfast = Ham
* e2: Lunch = Carnivorous
* e3: Dinner = Pork



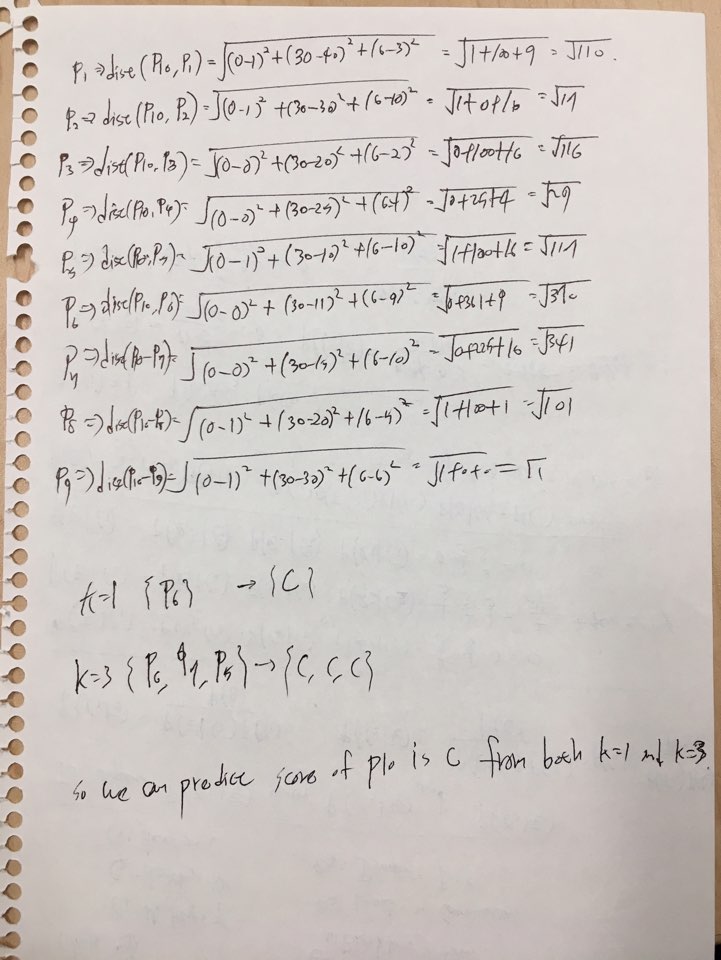
1. Use KNN Classifier to classify the objects [25]

We use students’ information to predict his score at the class

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Gender** | **Hrs/week on learning** | **Hrs/week on game playing** | **Score** |
| P1 | M | 40 | 3 | A |
| P2 | M | 30 | 10 | B |
| P3 | F | 20 | 2 | B |
| P4 | F | 25 | 4 | A |
| P5 | M | 10 | 10 | C |
| P6 | F | 11 | 9 | C |
| P7 | F | 15 | 10 | C |
| P8 | M | 20 | 5 | A |
| P9 | M | 30 | 6 | B |
| P10 | F | 30 | 6 | ? |

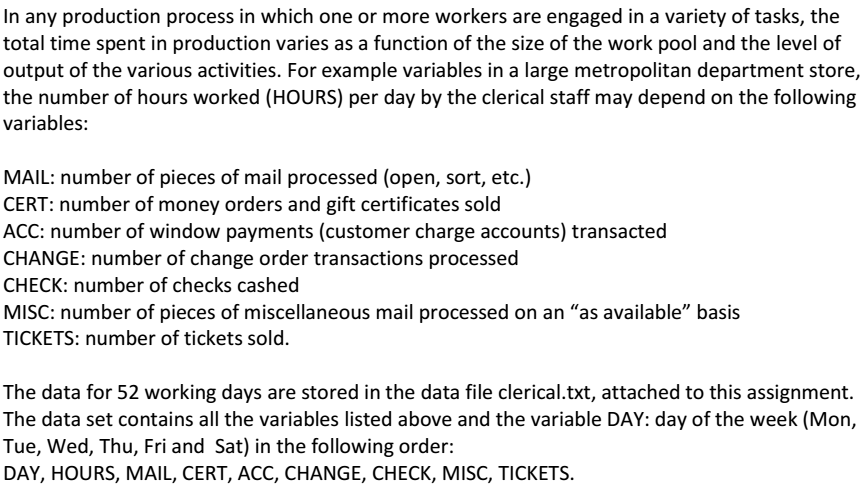
Use Manhattan distance as the distance measure, and try K = 1 and 3

Present your calculations step by step. Do not use R



1. Logistic regressions [50]

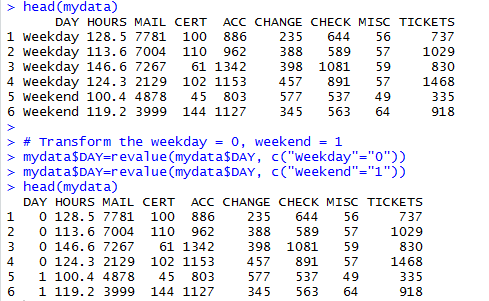
We use the data “HW6\_clerical\_Q2.txt”, this is the data set we used for in-class practice. The background of this data can be described as follows:



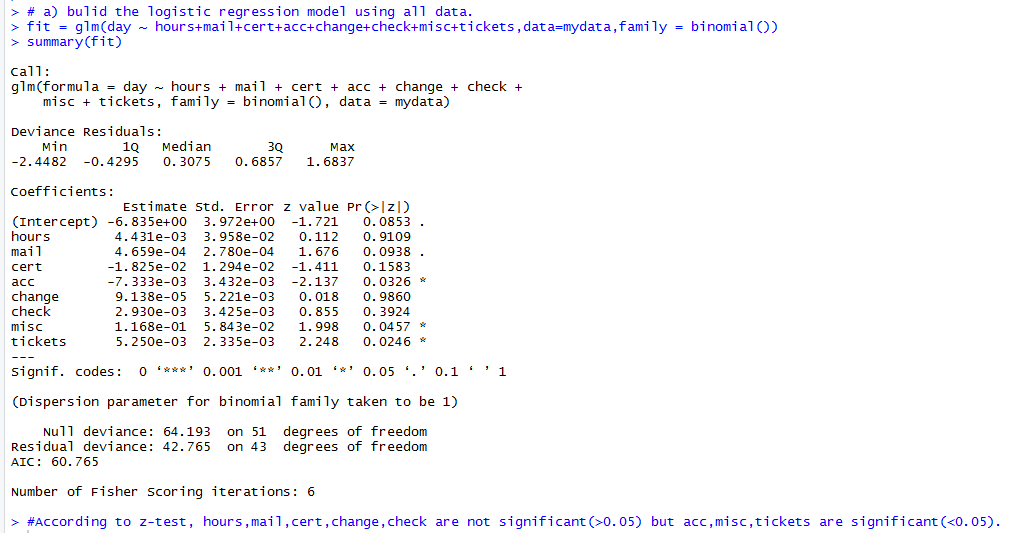
Okay, first, transform the original values in DAY to binary – weekend (Friday to Sunday) and weekday (Monday to Thursday). In this case, let’s have a practice of logistic regression by using DAY as the response variable, and all of the other variables (including hours) as the independent variables. Note: do not split data for this practice

**Please refer to comments for explanation of output!**

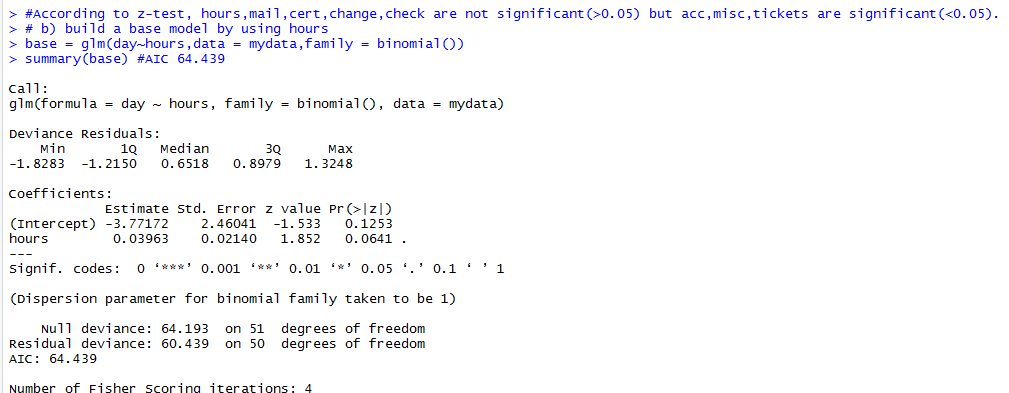


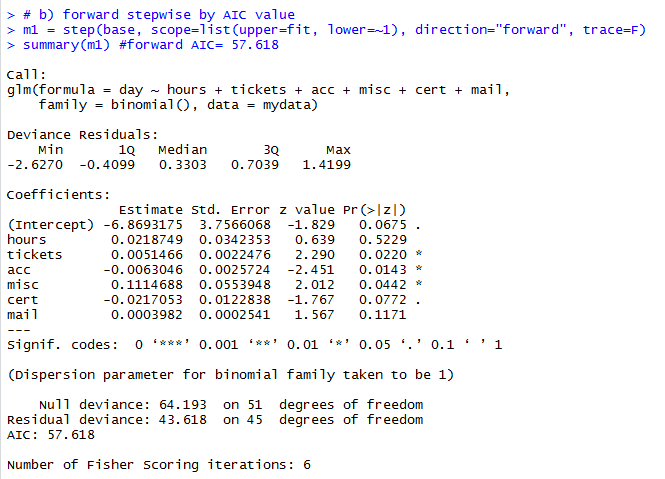


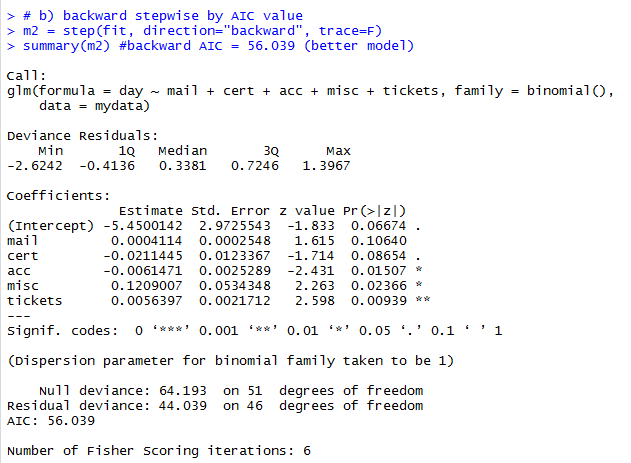
a). [10] Build the logistic regression model, Run it once, do not perform model selection. Show your outputs. Are all of the x variables significant?



b). [20] Build a base model by using hours as the single independent variable. Then use the stepwise forward and backward selections respectively based on the step() function. Compare the best models found by stepwise forward and backward selection, by using the AIC value?

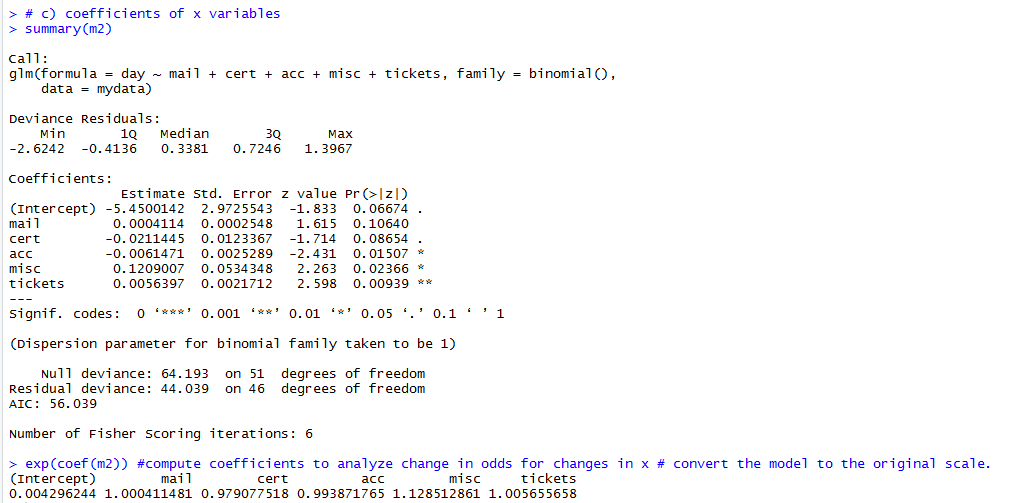


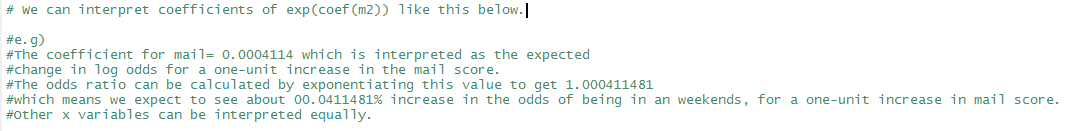


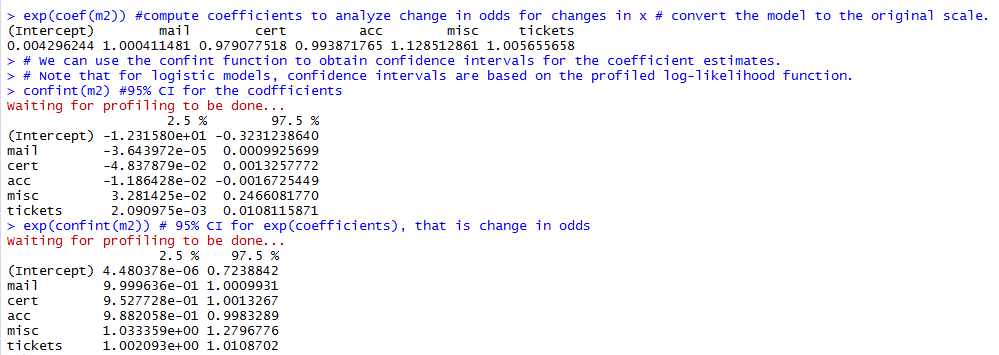


Therefore, m2 is best model in comparison with m1 and base model.

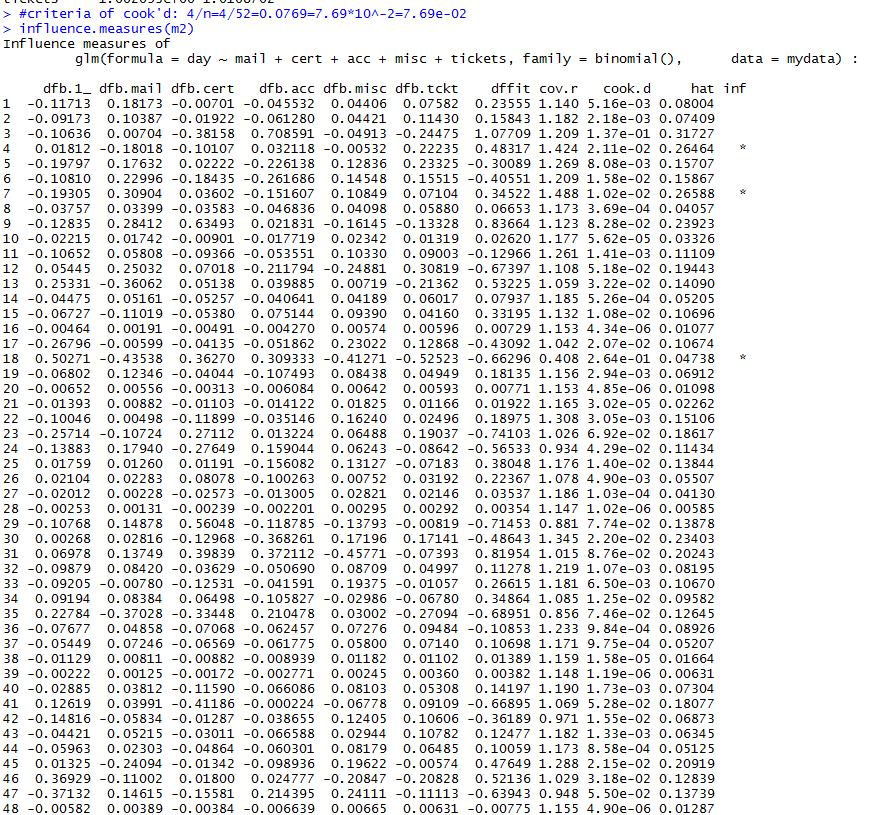
c). [10] Interpret the coefficient of the x variables in your best model







d). [10] Double check whether there are influential points by using cook’s distance as the criterion. Note: use influence.measures to produce cook’s distance







Code:

#Logistic Regression

getwd()

setwd('C:/users/mg/Desktop/Data Analytics/HW/HW8')

mydata=read.table('HW6\_clerical\_Q2.txt',header=T)

#install.packages("dummies")

#install.packages("caret")

#install.packages("leaps")

#install.packages("nnet")

#install.packages("boot")

library(car)# for use N-fold cross validataion

library(nnet)

library(dummies)

library(plyr) #for using revalue weekdays

# Transform the original values in DAY to binary - weekend (Friday to Sunday) and weekday (Monday to Thursday)

mydata$DAY = revalue(mydata$DAY, c("S"="Weekend"))

mydata$DAY = revalue(mydata$DAY, c("F"="Weekend"))

mydata$DAY = revalue(mydata$DAY, c("M"="Weekday"))

mydata$DAY = revalue(mydata$DAY, c("T"="Weekday"))

mydata$DAY = revalue(mydata$DAY, c("W"="Weekday"))

mydata$DAY = revalue(mydata$DAY, c("Th"="Weekday"))

head(mydata)

# Transform the weekday = 0, weekend = 1

mydata$DAY=revalue(mydata$DAY, c("Weekday"="0"))

mydata$DAY=revalue(mydata$DAY, c("Weekend"="1"))

head(mydata)

# declare variables

day = mydata$DAY

hours = mydata$HOURS

mail = mydata$MAIL

cert = mydata$CERT

acc = mydata$ACC

change = mydata$CHANGE

check = mydata$CHECK

misc = mydata$MISC

tickets = mydata$TICKETS

library(dummies)

library(caret)

library(boot)

library(leaps)

# a) bulid the logistic regression model using all data.

fit = glm(day ~ hours+mail+cert+acc+change+check+misc+tickets,data=mydata,family = binomial())

summary(fit)

#According to z-test, hours,mail,cert,change,check are not significant(>0.05) but acc,misc,tickets are significant(<0.05).

# b) build a base model by using hours

base = glm(day~hours,data = mydata,family = binomial())

summary(base) #AIC 64.439

# b) forward stepwise by AIC value

m1 = step(base, scope=list(upper=fit, lower=~1), direction="forward", trace=F)

summary(m1) #forward AIC= 57.618

# b) backward stepwise by AIC value

m2 = step(fit, direction="backward", trace=F)

summary(m2) #backward AIC = 56.039 (better model)

# c) coefficients of x variables

summary(m2)

exp(coef(m2)) #compute coefficients to analyze change in odds for changes in x # convert the model to the original scale.

# We can interpret coefficients of exp(coef(m2)) like this below.

#e.g)

#The coefficient for mail= 0.0004114 which is interpreted as the expected

#change in log odds for a one-unit increase in the mail score.

#The odds ratio can be calculated by exponentiating this value to get 1.000411481

#which means we expect to see about 00.0411481% increase in the odds of being in an weekends, for a one-unit increase in mail score.

#Other x variables can be interpreted equally.

# We can use the confint function to obtain confidence intervals for the coefficient estimates.

# Note that for logistic models, confidence intervals are based on the profiled log-likelihood function.

confint(m2) #95% CI for the codfficients

exp(confint(m2)) # 95% CI for exp(coefficients), that is change in odds

#criteria of cook'd: 4/n=4/52=0.0769=7.69\*10^-2=7.69e-02

influence.measures(m2)

#index[3,9,18,29,31] are influence points