Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Кафедра Штучного інтелекту

Звіт з лабораторної роботи №1

з курсу Машинне навчання

по темі “Знайомство з середою R, з поняттям навчальної та тестової вибірки, реалізувати процедури попередньої обробки”

Виконав:

студент гр. ІТШI 17-1

Гура А. О.

Перевірила:

проф. Кулішова Н.Є.

Харків 2019

1. Цель работы: Ознакомиться со средой R, с понятием обучающей и тестовой выборки, реализовать процедуры предварительной обработки

2. Ход работы:

2.1.Ознакомился со структурой UCI-репозитория.

2.2. Детально ознакомился с выборками IRIS, WINE, ZOO, а также выборками HEART DEASISES(содержит пропущенные значения) и AIRQUALITY (содержит пропущенные значения).

2.3.Ознакомился со средой, базовыми командами и функциями R.

2.4.Написал 3 процедуры предварительной обработки выборки (на R):

* процедуру случайного перемешиванию записей выборки и деления ее на обучающую и тестовую;
* процедуру назначения номеров класса для символьных значений целевого атрибута, которая заменяет символьные значения классов их порядковыми номерами;
* процедуру замены пропущенных значений средний значением среди примеров этого класса.

2.5.Изучил графические возможности среды R, для выборки Iris построил график зависимости (plot) атрибута sepal\_length от sepal\_width, гистограмму атрибута petal\_length. График имеет легенду. заголовок, заголовки осей.

3. Описание 5-ти рассмотренных выборок:

3.1. IRIS включает 4 атрибута, 150 экземпляров, 3 класса, не имеет пропущенных значений и используется в основном для задачи классификации.

3.2. WINE включает 13 атрибутов, 178 экземпляров, 3 класса, не имеет пропущенных значений и используется в основном для задачи классификации.

3.3. ZOO включает 17 атрибутов, 101 экземпляр, 7 классов, не имеет пропущенных значений и используется в основном для задачи классификации.

3.4. HEART DEASISES включает 75 атрибутов однако нам необходимы только 14 из них, 303 экземпляра, имеет пропущенные значения.

3.5. AIRQUALITY включает 6 атрибутов, 131 экземпляр, 5 классов, имеет пропущенные значения и используется в основном для задач регрессии.

4. Код:

iris<-read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"), header=**FALSE**)

irisEtalon<-read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"), header=**FALSE**)

zoo<-read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/zoo/zoo.data"), header=**TRUE**)

wine<-read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"), header=**TRUE**)

ac <- airquality

heart <- read.csv("heart.csv")

In [310]:

names(iris)<-c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Species")

names(zoo)<-c('Name','Hair','Geathers','Eggs','Milk','Airborne','Aquatic',

'Predator','Toothed','Backbone','Breathes','Venomous','Fins','Legs','Tail','Domestic','Catsize','Type')

names(wine)<-c('Classes','Alcohol','Malic.Acid','Ash','Alcalinity.Of.Ash',

'Magnesium','Total.Phenols','Flavanoids','Nonflavanoid.Phenols','Proanthocyanics',

'Color.Intensity','Hue','0D280/0D315','Proline')

names(heart)[1] <- "age"

In [275]:

change\_col\_names <- function(dataSet, colNames){

*#This function*

names(dataSet) <-colNames

return(dataSet)

}

shuffle <- function(dataSet){

*#This function randomly shuffles input dataset*

shuffledDS <-dataSet[sample(nrow(dataSet)),]

return(shuffledDS)

}

replace\_helper <- function(col){

col[is.na(col)] <- median(col, **TRUE**)

return(col)

}

replace\_na <- function(dataSet, targetAttr){

*#This function gets dataset, and the name of target att., and replaces all na's with median*

for (class in 1:length(unique(dataSet[targetAttr][[1]])))

{

for(num in 1:(length(dataSet)-1)){

dataSet[which(dataSet[targetAttr]==class),][[num]] <-

replace\_helper(dataSet [which(dataSet[targetAttr]==class),][[num]])

}

}

return(dataSet)

}

replace\_class <- function(classCol){

if(is.null(levels(classCol)))

{

i=1

for(num in unique(classCol)){

classCol[classCol==num]<-i

i=i+ 1

}

return(classCol)

}

else{

levels(classCol) <- seq(length(levels(classCol)))

return(classCol)

}

}

train\_test\_sep <- function(df){

*#Function sepparates dataframe into test and train sets*

*#Returns list where 1st is train set, 2nd is test set*

set.seed(111)

sample\_size <- floor(0.7 \* nrow(df))

train\_ind <- sample(seq\_len(nrow(df)), size = sample\_size)

train\_set <- df[train\_ind,]

test\_set <-df[-train\_ind,]

head(test\_set)

head(train\_set)

ls<-list(train\_set, test\_set)

return(ls)

}

makeplot<-function(){

plot(iris$Sepal.Length, iris$Sepal.Width, main="Sepal", pch=21,bg=c('red','green','blue')[unclass(iris$Species)],

xlab='Width',ylab='Length')

options(repr.plot.width=5, repr.plot.height=5)

legend(2,7.9,legend=c('setoza','versicolor','virginica'),col=c('red','green','blue'),

lty=1,cex=0.6,title='Species',text.font=6)

}

5. График и гистограмма:

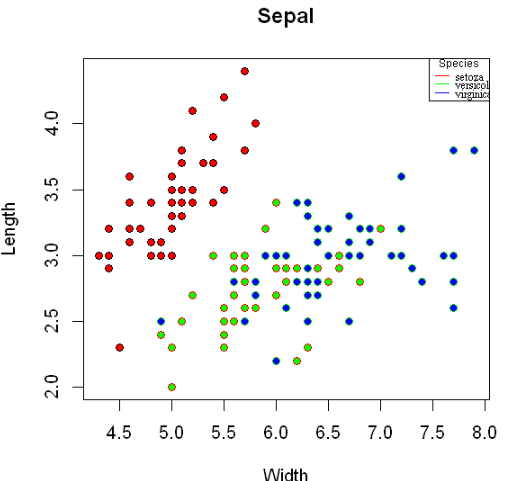


Рис. 1 - График зависимости Sepal Length от Sepal Width

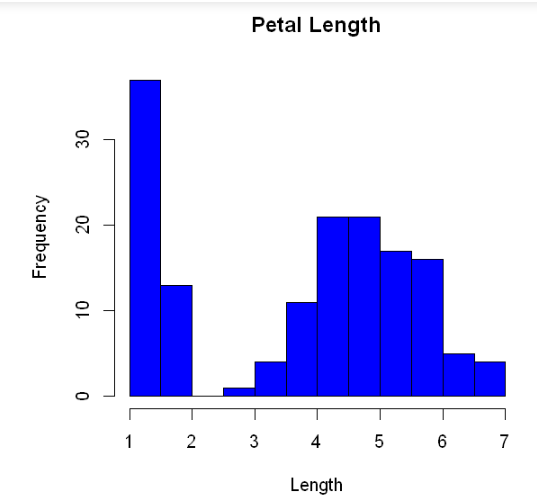


Рис. 2 - Гистограмма Petal Length.

6. Выводы:

При выполнении данной лабораторной работы было проведено ознакомление с языком и средой R. Ознакомление с репозиторием баз данных, с такими выборками, как : iris, wine, zoo, airquality, heart deasises. Были усвоены понятия обучающей и тестовой выборок. В итоге была реализована процедура предварительной обработки: разделение выборки на тестовую и обучающую, заменены пропущенные значения, перемешивание записей выборки.