R Notebook

#BIS 581 ##Paul Dunn ###Missing Value imputation

The code below uses random forests (covered later, don’t worry about them for now) to fill in some missing values (na) in a data set. The iris dataset is built-in and has data about flowers in it. Feel free to look at the dataset and/or open the help file for it.

data(iris)  
#let's get our own copy  
iris.na <- iris  
set.seed(111)  
## artificially drop some data values.  
for (i in 1:4) iris.na[sample(150, 20), i] <- NA  
  
#how many missing values do we have?  
summary(iris.na)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.200 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.500 1st Qu.:0.300   
## Median :5.750 Median :3.000 Median :4.400 Median :1.300   
## Mean :5.832 Mean :3.062 Mean :3.758 Mean :1.198   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## NA's :20 NA's :20 NA's :20 NA's :20   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##   
##

View(iris)  
View(iris.na)

This shows us how many missing values we have, spread across the data set. Just for comparison’s sake, let’s look at an original value, the row where that value is missing and then the value our function (below) fills in

iris[6,4]

## [1] 0.4

iris.na[6,4]

## [1] NA

won’t be perfect, but let’s fill in the values with impute, randomforest

library(randomForest)

## randomForest 4.7-1.2

## Type rfNews() to see new features/changes/bug fixes.

set.seed(222)  
iris.imputed <- rfImpute(Species ~ ., iris.na)

## ntree OOB 1 2 3  
## 300: 5.33% 4.00% 6.00% 6.00%  
## ntree OOB 1 2 3  
## 300: 5.33% 4.00% 6.00% 6.00%  
## ntree OOB 1 2 3  
## 300: 5.33% 2.00% 6.00% 8.00%  
## ntree OOB 1 2 3  
## 300: 5.33% 4.00% 6.00% 6.00%  
## ntree OOB 1 2 3  
## 300: 4.67% 2.00% 6.00% 6.00%

View(iris.imputed)

OK, now let’s compare values again

iris[3,2]

## [1] 3.2

iris.na[3,2]

## [1] 3.2

iris.imputed[3,2]

## [1] 4.7

print(1-iris.imputed[3,2]/iris[3,2])

## [1] -0.46875

Not perfect by any means, but if we need all of the rows, it works. Take a look at the last line printed above. this is just one minus the ratio of the imputed value to the ‘real’ value.