hw10

Mark Pearl

3/26/2020

" ## Question 14.1 The breast cancer data set breast-cancer-wisconsin.data.txt from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ (description at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) has missing values. 1. Use the mean/mode imputation method to impute values for the missing data. 2. Use regression to impute values for the missing data. 3. Use regression with perturbation to impute values for the missing data. 4. (Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using (1) the data sets from questions 1,2,3; (2) the data that remains after data points with missing values are removed; and (3) the data set when a binary variable is introduced to indicate missing values.

```
cancer_data <- read.csv('C:/Users/mjpearl/Desktop/omsa/ISYE-6501-OAN/hw10/breast-cancer-wisconsin.data.
summary(cancer_data)
```

```
##        id           clump_thickness  unif_cellsize    unif_cellshape
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
##  3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##    marg_adhes     single_epitheilial  bare_nuclei  bland_chromatin
##  Min.   : 1.000   Min.   : 1.000     1      :402   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 2.000     10     :132   1st Qu.: 2.000
##  Median : 1.000   Median : 2.000     2      : 30   Median : 3.000
##  Mean   : 2.807   Mean   : 3.216     5      : 30   Mean   : 3.438
##  3rd Qu.: 4.000   3rd Qu.: 4.000     3      : 28   3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000     8      : 21   Max.   :10.000
##                                      (Other): 56
##  normal_nucleoli     mitoses           class
##  Min.   : 1.000   Min.   : 1.000   Min.   :2.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
##  Median : 1.000   Median : 1.000   Median :2.00
##  Mean   : 2.867   Mean   : 1.589   Mean   :2.69
##  3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
##  Max.   :10.000   Max.   :10.000   Max.   :4.00
##
```

When we take a look at the summary plot we can see that the bare_nuclei feature contains missing values that we will need to impute, Let's determine the sum of those numbers of missing records before we complete imputation.

1) Using mean/mode imputation method

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
sum(cancer_data$bare_nuclei=="?")
```

```
## [1] 16
```

Now we can see that the ? values have been replaced correctly and we'll be able to use the mean / mode imputation method on the bare_nuclei column

```
cancer_data$bare_nuclei <- na_if(cancer_data$bare_nuclei,"?")
sum(cancer_data$bare_nuclei=="?")
```

```
## [1] NA

## Warning: package 'imputeMissings' was built under R version 3.6.3

##
## Attaching package: 'imputeMissings'

## The following object is masked from 'package:dplyr':
##
##     compute

## [1] 0
```

2. Regression Imputation

Now we're going to use regression for the imputation to provide a prediction for the missing value We'll first find the indexes for all of the observations that have an NA value.

```
missing <-which(is.na(cancer_data$bare_nuclei), arr.ind = TRUE)

########################## Regression Imputation ####################

# Not to include the response variable in regression imputation
data_modified <- cancer_data[-missing, 2:10]
data_modified$bare_nuclei <- as.integer(data_modified$bare_nuclei)

names(data_modified)
```

```
## [1] "clump_thickness"    "unif_cellsize"      "unif_cellshape"
## [4] "marg_adhes"         "single_epitheilial" "bare_nuclei"
## [7] "bland_chromatin"    "normal_nucleoli"    "mitoses"
```

```r
#---- Linear regression Imputation --------

model <- lm(bare_nuclei~clump_thickness+unif_cellsize+unif_cellshape+marg_adhes+single_epitheilial+blanc
summary(model)
```

```
##
## Call:
## lm(formula = bare_nuclei ~ clump_thickness + unif_cellsize +
##     unif_cellshape + marg_adhes + single_epitheilial + bland_chromatin +
##     normal_nucleoli + mitoses, data = data_modified)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1137 -0.7185 -0.4731 -0.2994  7.3848
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.862817   0.162497  11.464  < 2e-16 ***
## clump_thickness     0.068118   0.034746   1.960  0.05035 .
## unif_cellsize       0.087939   0.063482   1.385  0.16643
## unif_cellshape      0.110046   0.061190   1.798  0.07255 .
## marg_adhes         -0.076950   0.038270  -2.011  0.04475 *
## single_epitheilial  0.043216   0.052123   0.829  0.40733
## bland_chromatin     0.044536   0.049211   0.905  0.36579
## normal_nucleoli     0.119422   0.037076   3.221  0.00134 **
## mitoses             0.001405   0.049448   0.028  0.97733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.896 on 674 degrees of freedom
## Multiple R-squared:  0.2326, Adjusted R-squared:  0.2235
## F-statistic: 25.54 on 8 and 674 DF,  p-value: < 2.2e-16
```

```r
# predict V7
V7_hat <- predict(model, newdata = cancer_data[missing,])
```

15. Optimization is used frequently in the energy to determine the proper resource allocation for power dams on an existing project. Models are run to determine what is the proper constraints to put in place to determine when is the ideal time to do maintenance on assets to avoid shutting down during peak times. Data required is weather data and pricing data.