EDA - Assignments 6 and 7

Aaron Niskin

September 8, 2016

1. Read the data

```
library("RWeka", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.3")
dats <- read.arff("Chronic_Kidney_Disease/chronic_kidney_disease_full.arff")
```

So we check out our data, and everything seems to look fine, but just in case, let's check out the file (I've been told that there may be an extra comma in there).

```
grep ',,' Chronic_Kidney_Disease/chronic_kidney_disease_full.arff
```

```
## 75,70,1.020,0,0,normal,normal,notpresent,notpresent,107,48,0.8,144,3.5,13.6,46,10300,4.8,no,,no,no,g
```

and sure enough, it's there. At this point we want to go into our favorite VIM editor, and check this out (after all, as we can see from the data, null values are supposed to be denoted by single question marks).

```
kidneyDats <- read.arff("Chronic_Kidney_Disease/chronic_kidney_disease_full_edit.arff")
```

2. Assess the completeness of the data

```
sum(complete.cases(kidneyDats))
```

```
## [1] 158
```

```
dim(kidneyDats)
```

```
## [1] 400  25
```

```
because I feel as though this isn't being too descriptive, I'll use a function I wrote for assignment 3
```

```
hasInvalidDataNames <- function(dats, nullValues) {
  hasInvalidData <- function(col_name) {
    occurrences <- sapply(dats[,col_name], function(x) {
      is.null(x) || is.na(x) || x %in% nullValues
      })
    length(which(occurrences))
  }
  sapply(names(dats), hasInvalidData)
}
invalidRows_tmp <- hasInvalidDataNames(dats, c("", "None", "?", "Unknown/Invalid"))
invalidRows     <- invalidRows_tmp[invalidRows_tmp > 0]
invalidRows
```

```
##   age    bp    sg    al    su   rbc    pc   pcc    ba   bgr    bu    sc
##     9    12    47    46    49   152    65     4     4    44    19    17
##   sod   pot  hemo   pcv  wbcc  rbcc   htn    dm   cad appet    pe   ane
##    87    88    52    71   106   131     2     2     2     1     1     1
```

So, if you're tryingto model pcc against age, for instance, this should be sufficient. If, on the other

3. Discretize the potasium feature using 5 bins containing equal numbers of observations

```
discrete_pot <- cut(dats$pot,
                    breaks=quantile(dats$pot, seq(0, 1, 0.2), na.rm = TRUE),
                    include.lowest = TRUE)
```

4. Discretize the potasium feature using 5 bins of equal length

```
disc_pot_2 <- cut(dats$pot, 5, include.lowest = TRUE)
```

5. Discretize the potasium feature using a topdown approach from the discretize package

```
tmp = cbind(dats$pot[!is.na(dats$pot)], dats$pot[!is.na(dats$pot)])
tmp <- with(dats, discretization::disc.Topdown(tmp, method = 1))
tmp_doit <- function(a) {
  return(paste(unname(tmp$cutp[[1]])[a], unname(tmp$cutp[[1]])[a + 1], sep="-"))
}
tmp2 <- lapply(tmp$Disc.data[[1]], function(a) {return(c(tmp_doit=a))})
tmp2
```

```
## [[1]]
## tmp_doit
##        1
##
## [[2]]
## tmp_doit
##        2
##
## [[3]]
## tmp_doit
##        3
##
## [[4]]
## tmp_doit
##        2
##
## [[5]]
## tmp_doit
##        3
##
## [[6]]
## tmp_doit
##        4
##
```

```
## [[7]]
## tmp_doit
##        2
##
## [[8]]
## tmp_doit
##        5
##
## [[9]]
## tmp_doit
##        3
##
## [[10]]
## tmp_doit
##        3
##
## [[11]]
## tmp_doit
##        2
##
## [[12]]
## tmp_doit
##        3
##
## [[13]]
## tmp_doit
##        4
##
## [[14]]
## tmp_doit
##        2
##
## [[15]]
## tmp_doit
##        3
##
## [[16]]
## tmp_doit
##        4
##
## [[17]]
## tmp_doit
##        2
##
## [[18]]
## tmp_doit
##        3
##
## [[19]]
## tmp_doit
##        3
##
## [[20]]
## tmp_doit
```

```
##        3
##
## [[21]]
## tmp_doit
##        3
##
## [[22]]
## tmp_doit
##        2
##
## [[23]]
## tmp_doit
##        3
##
## [[24]]
## tmp_doit
##        4
##
## [[25]]
## tmp_doit
##        3
##
## [[26]]
## tmp_doit
##        3
##
## [[27]]
## tmp_doit
##        2
##
## [[28]]
## tmp_doit
##        3
##
## [[29]]
## tmp_doit
##        3
##
## [[30]]
## tmp_doit
##        3
##
## [[31]]
## tmp_doit
##        3
##
## [[32]]
## tmp_doit
##        3
##
## [[33]]
## tmp_doit
##        3
##
```

```
## [[34]]
## tmp_doit
##        3
##
## [[35]]
## tmp_doit
##        3
##
## [[36]]
## tmp_doit
##        3
##
## [[37]]
## tmp_doit
##        4
##
## [[38]]
## tmp_doit
##        5
##
## [[39]]
## tmp_doit
##        7
##
## [[40]]
## tmp_doit
##        3
##
## [[41]]
## tmp_doit
##        2
##
## [[42]]
## tmp_doit
##        3
##
## [[43]]
## tmp_doit
##        3
##
## [[44]]
## tmp_doit
##        3
##
## [[45]]
## tmp_doit
##        3
##
## [[46]]
## tmp_doit
##        3
##
## [[47]]
## tmp_doit
```

```
##        2
##
## [[48]]
## tmp_doit
##        4
##
## [[49]]
## tmp_doit
##        4
##
## [[50]]
## tmp_doit
##        2
##
## [[51]]
## tmp_doit
##        4
##
## [[52]]
## tmp_doit
##        3
##
## [[53]]
## tmp_doit
##        2
##
## [[54]]
## tmp_doit
##        2
##
## [[55]]
## tmp_doit
##        6
##
## [[56]]
## tmp_doit
##        3
##
## [[57]]
## tmp_doit
##        2
##
## [[58]]
## tmp_doit
##        2
##
## [[59]]
## tmp_doit
##        3
##
## [[60]]
## tmp_doit
##        1
##
```

```
## [[61]]
## tmp_doit
##        3
##
## [[62]]
## tmp_doit
##        2
##
## [[63]]
## tmp_doit
##        2
##
## [[64]]
## tmp_doit
##        3
##
## [[65]]
## tmp_doit
##        2
##
## [[66]]
## tmp_doit
##        2
##
## [[67]]
## tmp_doit
##        4
##
## [[68]]
## tmp_doit
##        3
##
## [[69]]
## tmp_doit
##        3
##
## [[70]]
## tmp_doit
##        3
##
## [[71]]
## tmp_doit
##        3
##
## [[72]]
## tmp_doit
##        3
##
## [[73]]
## tmp_doit
##        2
##
## [[74]]
## tmp_doit
```

```
##        4
##
## [[75]]
## tmp_doit
##        2
##
## [[76]]
## tmp_doit
##        4
##
## [[77]]
## tmp_doit
##        2
##
## [[78]]
## tmp_doit
##        2
##
## [[79]]
## tmp_doit
##        3
##
## [[80]]
## tmp_doit
##        3
##
## [[81]]
## tmp_doit
##        3
##
## [[82]]
## tmp_doit
##        8
##
## [[83]]
## tmp_doit
##        3
##
## [[84]]
## tmp_doit
##        3
##
## [[85]]
## tmp_doit
##        3
##
## [[86]]
## tmp_doit
##        3
##
## [[87]]
## tmp_doit
##        4
##
```

```
## [[88]]
## tmp_doit
##        3
##
## [[89]]
## tmp_doit
##        5
##
## [[90]]
## tmp_doit
##        4
##
## [[91]]
## tmp_doit
##        2
##
## [[92]]
## tmp_doit
##        3
##
## [[93]]
## tmp_doit
##        4
##
## [[94]]
## tmp_doit
##        3
##
## [[95]]
## tmp_doit
##        4
##
## [[96]]
## tmp_doit
##        3
##
## [[97]]
## tmp_doit
##        4
##
## [[98]]
## tmp_doit
##        3
##
## [[99]]
## tmp_doit
##        3
##
## [[100]]
## tmp_doit
##        2
##
## [[101]]
## tmp_doit
```

```
##        4
##
## [[102]]
## tmp_doit
##        2
##
## [[103]]
## tmp_doit
##        2
##
## [[104]]
## tmp_doit
##        3
##
## [[105]]
## tmp_doit
##        3
##
## [[106]]
## tmp_doit
##        3
##
## [[107]]
## tmp_doit
##        2
##
## [[108]]
## tmp_doit
##        3
##
## [[109]]
## tmp_doit
##        3
##
## [[110]]
## tmp_doit
##        3
##
## [[111]]
## tmp_doit
##        3
##
## [[112]]
## tmp_doit
##        3
##
## [[113]]
## tmp_doit
##        2
##
## [[114]]
## tmp_doit
##        4
##
```

```
## [[115]]
## tmp_doit
##        1
##
## [[116]]
## tmp_doit
##        1
##
## [[117]]
## tmp_doit
##        3
##
## [[118]]
## tmp_doit
##        2
##
## [[119]]
## tmp_doit
##        2
##
## [[120]]
## tmp_doit
##        3
##
## [[121]]
## tmp_doit
##        2
##
## [[122]]
## tmp_doit
##        3
##
## [[123]]
## tmp_doit
##        3
##
## [[124]]
## tmp_doit
##        3
##
## [[125]]
## tmp_doit
##        3
##
## [[126]]
## tmp_doit
##        2
##
## [[127]]
## tmp_doit
##        2
##
## [[128]]
## tmp_doit
```

```
##        4
##
## [[129]]
## tmp_doit
##        3
##
## [[130]]
## tmp_doit
##        1
##
## [[131]]
## tmp_doit
##        5
##
## [[132]]
## tmp_doit
##        2
##
## [[133]]
## tmp_doit
##        2
##
## [[134]]
## tmp_doit
##        4
##
## [[135]]
## tmp_doit
##        3
##
## [[136]]
## tmp_doit
##        4
##
## [[137]]
## tmp_doit
##        3
##
## [[138]]
## tmp_doit
##        3
##
## [[139]]
## tmp_doit
##        4
##
## [[140]]
## tmp_doit
##        4
##
## [[141]]
## tmp_doit
##        5
##
```

```
## [[142]]
## tmp_doit
##        3
##
## [[143]]
## tmp_doit
##        4
##
## [[144]]
## tmp_doit
##        4
##
## [[145]]
## tmp_doit
##        3
##
## [[146]]
## tmp_doit
##        3
##
## [[147]]
## tmp_doit
##        4
##
## [[148]]
## tmp_doit
##        3
##
## [[149]]
## tmp_doit
##        2
##
## [[150]]
## tmp_doit
##        2
##
## [[151]]
## tmp_doit
##        3
##
## [[152]]
## tmp_doit
##        1
##
## [[153]]
## tmp_doit
##        3
##
## [[154]]
## tmp_doit
##        2
##
## [[155]]
## tmp_doit
```

```
##        3
##
## [[156]]
## tmp_doit
##        4
##
## [[157]]
## tmp_doit
##        3
##
## [[158]]
## tmp_doit
##        3
##
## [[159]]
## tmp_doit
##        2
##
## [[160]]
## tmp_doit
##        2
##
## [[161]]
## tmp_doit
##        1
##
## [[162]]
## tmp_doit
##        3
##
## [[163]]
## tmp_doit
##        5
##
## [[164]]
## tmp_doit
##        4
##
## [[165]]
## tmp_doit
##        3
##
## [[166]]
## tmp_doit
##        3
##
## [[167]]
## tmp_doit
##        5
##
## [[168]]
## tmp_doit
##        4
##
```

```
## [[169]]
## tmp_doit
##        3
##
## [[170]]
## tmp_doit
##        3
##
## [[171]]
## tmp_doit
##        3
##
## [[172]]
## tmp_doit
##        2
##
## [[173]]
## tmp_doit
##        2
##
## [[174]]
## tmp_doit
##        4
##
## [[175]]
## tmp_doit
##        2
##
## [[176]]
## tmp_doit
##        2
##
## [[177]]
## tmp_doit
##        4
##
## [[178]]
## tmp_doit
##        2
##
## [[179]]
## tmp_doit
##        2
##
## [[180]]
## tmp_doit
##        2
##
## [[181]]
## tmp_doit
##        4
##
## [[182]]
## tmp_doit
```

```
##        3
##
## [[183]]
## tmp_doit
##        3
##
## [[184]]
## tmp_doit
##        3
##
## [[185]]
## tmp_doit
##        2
##
## [[186]]
## tmp_doit
##        4
##
## [[187]]
## tmp_doit
##        3
##
## [[188]]
## tmp_doit
##        3
##
## [[189]]
## tmp_doit
##        3
##
## [[190]]
## tmp_doit
##        4
##
## [[191]]
## tmp_doit
##        3
##
## [[192]]
## tmp_doit
##        3
##
## [[193]]
## tmp_doit
##        3
##
## [[194]]
## tmp_doit
##        3
##
## [[195]]
## tmp_doit
##        4
##
```

```
## [[196]]
## tmp_doit
##        3
##
## [[197]]
## tmp_doit
##        3
##
## [[198]]
## tmp_doit
##        3
##
## [[199]]
## tmp_doit
##        4
##
## [[200]]
## tmp_doit
##        2
##
## [[201]]
## tmp_doit
##        3
##
## [[202]]
## tmp_doit
##        2
##
## [[203]]
## tmp_doit
##        4
##
## [[204]]
## tmp_doit
##        2
##
## [[205]]
## tmp_doit
##        3
##
## [[206]]
## tmp_doit
##        3
##
## [[207]]
## tmp_doit
##        2
##
## [[208]]
## tmp_doit
##        3
##
## [[209]]
## tmp_doit
```

```
##        4
##
## [[210]]
## tmp_doit
##        3
##
## [[211]]
## tmp_doit
##        4
##
## [[212]]
## tmp_doit
##        2
##
## [[213]]
## tmp_doit
##        4
##
## [[214]]
## tmp_doit
##        3
##
## [[215]]
## tmp_doit
##        3
##
## [[216]]
## tmp_doit
##        2
##
## [[217]]
## tmp_doit
##        3
##
## [[218]]
## tmp_doit
##        3
##
## [[219]]
## tmp_doit
##        4
##
## [[220]]
## tmp_doit
##        2
##
## [[221]]
## tmp_doit
##        4
##
## [[222]]
## tmp_doit
##        3
##
```

```
## [[223]]
## tmp_doit
##        2
##
## [[224]]
## tmp_doit
##        3
##
## [[225]]
## tmp_doit
##        4
##
## [[226]]
## tmp_doit
##        3
##
## [[227]]
## tmp_doit
##        3
##
## [[228]]
## tmp_doit
##        3
##
## [[229]]
## tmp_doit
##        3
##
## [[230]]
## tmp_doit
##        4
##
## [[231]]
## tmp_doit
##        3
##
## [[232]]
## tmp_doit
##        3
##
## [[233]]
## tmp_doit
##        2
##
## [[234]]
## tmp_doit
##        3
##
## [[235]]
## tmp_doit
##        3
##
## [[236]]
## tmp_doit
```

```
##        2
##
## [[237]]
## tmp_doit
##        3
##
## [[238]]
## tmp_doit
##        4
##
## [[239]]
## tmp_doit
##        3
##
## [[240]]
## tmp_doit
##        2
##
## [[241]]
## tmp_doit
##        3
##
## [[242]]
## tmp_doit
##        3
##
## [[243]]
## tmp_doit
##        3
##
## [[244]]
## tmp_doit
##        2
##
## [[245]]
## tmp_doit
##        2
##
## [[246]]
## tmp_doit
##        3
##
## [[247]]
## tmp_doit
##        4
##
## [[248]]
## tmp_doit
##        3
##
## [[249]]
## tmp_doit
##        2
##
```

```
## [[250]]
## tmp_doit
##        3
##
## [[251]]
## tmp_doit
##        3
##
## [[252]]
## tmp_doit
##        2
##
## [[253]]
## tmp_doit
##        2
##
## [[254]]
## tmp_doit
##        3
##
## [[255]]
## tmp_doit
##        4
##
## [[256]]
## tmp_doit
##        3
##
## [[257]]
## tmp_doit
##        3
##
## [[258]]
## tmp_doit
##        2
##
## [[259]]
## tmp_doit
##        3
##
## [[260]]
## tmp_doit
##        3
##
## [[261]]
## tmp_doit
##        4
##
## [[262]]
## tmp_doit
##        2
##
## [[263]]
## tmp_doit
```

```
##        2
##
## [[264]]
## tmp_doit
##        3
##
## [[265]]
## tmp_doit
##        2
##
## [[266]]
## tmp_doit
##        3
##
## [[267]]
## tmp_doit
##        4
##
## [[268]]
## tmp_doit
##        3
##
## [[269]]
## tmp_doit
##        4
##
## [[270]]
## tmp_doit
##        3
##
## [[271]]
## tmp_doit
##        2
##
## [[272]]
## tmp_doit
##        2
##
## [[273]]
## tmp_doit
##        2
##
## [[274]]
## tmp_doit
##        2
##
## [[275]]
## tmp_doit
##        2
##
## [[276]]
## tmp_doit
##        4
##
```

```
## [[277]]
## tmp_doit
##        2
##
## [[278]]
## tmp_doit
##        2
##
## [[279]]
## tmp_doit
##        2
##
## [[280]]
## tmp_doit
##        2
##
## [[281]]
## tmp_doit
##        4
##
## [[282]]
## tmp_doit
##        2
##
## [[283]]
## tmp_doit
##        2
##
## [[284]]
## tmp_doit
##        4
##
## [[285]]
## tmp_doit
##        3
##
## [[286]]
## tmp_doit
##        2
##
## [[287]]
## tmp_doit
##        2
##
## [[288]]
## tmp_doit
##        3
##
## [[289]]
## tmp_doit
##        2
##
## [[290]]
## tmp_doit
```

```
##        3
##
## [[291]]
## tmp_doit
##        3
##
## [[292]]
## tmp_doit
##        2
##
## [[293]]
## tmp_doit
##        2
##
## [[294]]
## tmp_doit
##        3
##
## [[295]]
## tmp_doit
##        4
##
## [[296]]
## tmp_doit
##        3
##
## [[297]]
## tmp_doit
##        3
##
## [[298]]
## tmp_doit
##        3
##
## [[299]]
## tmp_doit
##        2
##
## [[300]]
## tmp_doit
##        2
##
## [[301]]
## tmp_doit
##        2
##
## [[302]]
## tmp_doit
##        4
##
## [[303]]
## tmp_doit
##        2
##
```

```
## [[304]]
## tmp_doit
##        3
##
## [[305]]
## tmp_doit
##        3
##
## [[306]]
## tmp_doit
##        3
##
## [[307]]
## tmp_doit
##        4
##
## [[308]]
## tmp_doit
##        3
##
## [[309]]
## tmp_doit
##        2
##
## [[310]]
## tmp_doit
##        3
##
## [[311]]
## tmp_doit
##        3
##
## [[312]]
## tmp_doit
##        2
```

6. Discretize the potasium feature using a bottomup approach from the discretize package

---