Breast Cancer Data

Jas Kainth

22/09/2020

```
raw_data <- read_csv("breast_cancer.csv")
  # Recode the class; 2 = benign (i.e. not cancer/negative) so turn this to 0
  # 4 = malignant so turn this to 1
data <- raw_data %>%
  mutate(class_logistic = case_when(Class == 2 ~ 0,
                                    TRUE ~ 1)) %>%
  # Get rid of the first column since it doesn't provide any information
  select(-`Sample code number`)

# Make sure we didn't make a mistake during the class_logistic mutation
data %>%
  count(Class)
data %>%
  count(class_logistic)
# That looks good

# Rename the columns
data <- data %>%
  rename("clump_thickness" = "Clump Thickness",
         "unif_of_cell_size" = "Uniformity of Cell Size",
         "unif_of_cell_shape" = "Uniformity of Cell Shape",
         "marginal_adhesion" = "Marginal Adhesion",
         "single_epithelial_cell_size" = "Single Epithelial Cell Size",
         "bare_nuclei" = "Bare Nuclei",
         "bland_chromatin" = "Bland Chromatin",
         "normal_nucleoli" = "Normal Nucleoli",
         "mitoses" = "Mitoses",
         "class" = "Class")
# Is there any missing data?
sapply(data, function(x) sum(is.na(x)))

# Nice, there is no missing data
# Also, looking at the values it seems that all their values are from 1 - 10 (other
# than the dependent variable), which means they have already been scaled
# Check if there are any values for any column which we wouldn't expect
# (i.e. outside of the range of values)
data %>%
  pivot_longer(cols = clump_thickness:mitoses, names_to = "variable",
               values_to = "values") %>%
  distinct(values)
# The only values in our dataset range from 1 - 10 which is what we want
```

```r
# We might want to use the covariates as factors rather than numerical but for now
# we will leave them as numerical and later, in the statistical analysis part we can
# explore both options
write_csv(data, "full_data.csv")
```

Data Summary

The data provided was relatively clean. We were given 699 instances (unique data points) along with 10 attributes (possible predictors) and the dependent variable. The dependent variable was Class (i.e. benign or malignant), which was either a "2" or a "4" representing benign or malignant, respectively. This was re-coded, for convenience, to "0" or "1" representing benign or malignant, respectively. The attributes were given on a scale of 1 to 10. No information was provided on the original units or values. We had information for Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class of Cancer. There were no missing data nor data outside of the given range. We were also provided with a Sample Code Number (ID) which was a random number for each instance. The dependent variable was re-coded under a new column and then was written as a new .csv file for convenience to read into other files.