Research Log Project Machine Learning

Differentiating between benign and malignant samples of breast masses

Student: Vincent Talen

Student number: 389015

Class: BFV3

Study: Bio-Informatics

Institute: Institute for Life Science & Technology

Teachers: Dave Langers (LADR) and Bart Barnard (BABA)

Date: 2022-09-19

Contents

1 Data Set

1.1 Structure of data set

To easily annotate figures and to better understand and distinguish the data a codebook was manually created. This codebook will now be imported and the explanations of the multiple columns shown.

```
# Import codebook
codebook <- read_delim("data/codebook.txt", delim = "|", show_col_types = FALSE)
codebook[,c(1,4)]
```

```
## # A tibble: 32 x 2
##    ‘Column Name‘    Description
##    <chr>            <chr>
##  1 id               "Identification number for patient/sample"
##  2 diagnosis        "The class label of the breast mass with \"B\" = \"Benign\"~
##  3 radius_mean      "The mean of the radius’; A radius is the mean of distances~
##  4 texture_mean     "The mean of the textures; Texture is the standard deviatio~
##  5 perimeter_mean   "The mean of the perimeters; "
##  6 area_mean        "The mean of the areas; "
##  7 smoothness_mean  "The mean of the smoothnesses; Smoothness is the local vari~
##  8 compactness_mean "The mean of the compactnesses; Compactness is calculated w~
##  9 concavity_mean   "The mean of the concavities; Concavity is the severity of ~
## 10 concave_pts_mean "The mean of the concave points; The concave points are the~
## # ... with 22 more rows
```

1.2 Loading in the data

```
# Load in data from file
data <- read_csv("data/wdbc.data", col_names = codebook[[1]], show_col_types = FALSE)
head(data)
```

```
## # A tibble: 6 x 32
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
##      <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
## 1   842302 M                18.0         10.4          123.       1001
## 2   842517 M                20.6         17.8          133.       1326
## 3 84300903 M                19.7         21.2          130        1203
## 4 84348301 M                11.4         20.4           77.6       386.
## 5 84358402 M                20.3         14.3          135.       1297
## 6   843786 M                12.4         15.7           82.6       477.
## # ... with 26 more variables: smoothness_mean <dbl>, compactness_mean <dbl>,
## #   concavity_mean <dbl>, concave_pts_mean <dbl>, symmetry_mean <dbl>,
## #   fractal_dim_mean <dbl>, radius_se <dbl>, texture_se <dbl>,
## #   perimeter_se <dbl>, area_se <dbl>, smoothness_se <dbl>,
## #   compactness_se <dbl>, concavity_se <dbl>, concave_pts_se <dbl>,
## #   symmetry_se <dbl>, fractal_dim_se <dbl>, radius_worst <dbl>,
## #   texture_worst <dbl>, perimeter_worst <dbl>, area_worst <dbl>, ...
```

```r
# Print the amount of samples and columns
cat("Amount of samples:", dim(data)[1], "\tColumns in dataframe", dim(data)[2])
```

```
## Amount of samples: 569   Columns in dataframe 32
```

References

[1] W.N. Street, W.H. Wolberg and O.L. Mangasarian. (1993), *Nuclear feature extraction for breast tumor diagnosis.*, 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, http://rexa.info/paper/b98475235164960529ad2ff9fda3816e9335cf8a (accessed Sep 16, 2022).