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: 2022-09-20

Contents

1 Setting up R

For the data analysis and further processes multiple libraries are needed, they are loaded in here. A few other options/settings are also configured here.

```
# Set code chunks visibility in pdf output to true
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.align = "center")

# Create vector with all packages that will be used
packages <- c("pander", "readr", "ggplot2", "gridExtra", "scales", "ggpubr", "reshape2")
# Load each package in the vector with lapply
invisible(lapply(packages, library, character.only = TRUE))
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# Drop the packages variable from memory since it will not be used again
remove(packages)
```

2 Data Set

The data set that will be used for this project is the Wisconsin Breast Cancer (Diagnostic) Data Set, publicly available from the UCI Machine Learning Repository.
Diagnosing if breast masses/lumps are benign or malignant, and thus if they are breast cancer, was done by full biopsies, which are invasive surgical procedures. So the researchers wanted a faster and less invasive way to diagnose breast cancer. A promising technique was to take fine needle aspirate fluid samples from the breast mass and then look at the cells under a microscope, the diagnosis would then be based on the characteristics and contextual features of the cells that could visually be seen. This gave mixed results because it was highly subjective and depended a lot on the skill of the physician.
To make this process faster, improve the correctness and objectivity of the diagnosis process, image processing and machine learning techniques were used.

2.1 Origin of the data

The data was gathered by first collecting the fine needle aspirates, which are expressed on a glass slide and stained. A color video camera mounted on top of a microscope, where the images were projected into the camera with a 63x objective and 2.5x ocular. The image is then captured by a color frame grabber board as a 512x480, 8-bit-per-pixel Targa file.
The digitized image is then analyzed in the program Xcyt (made by Nick Street), where the boundaries of the nuclei are marked by the user and the features then further computed and extracted. The features are numerically modeled such that larger values will typically indicate a higher likelihood of malignancy.

2.2 Data structure

The full data set consists of 569 cases and has the class label 'Diagnosis', with all cases labeled as either benign or malignant. From the digitized image 10 features are extracted for each nucleus, namely the following:

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter (total distance between points of marked boundary)
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

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)]
# Disable printing 'table continues' lines between split sections of the table
panderOptions("table.continues", "")
# Pretty print the summary of the data frame
pander::pander(codebook[,1:3], style = "rmarkdown")
```

Column Name	Full Name	Type
id	ID	dbl
diagnosis	Diagnosis	fct
radius_mean	Mean Radius	dbl
texture_mean	Mean Texture	dbl
perimeter_mean	Mean Perimeter	dbl
area_mean	Mean Area	dbl
smoothness_mean	Mean Smoothness	dbl
compactness_mean	Mean Compactness	dbl
concavity_mean	Mean Concavity	dbl
concave_pts_mean	Mean Concave Points	dbl
symmetry_mean	Mean Symmetry	dbl
fractal_dim_mean	Mean Fractal Dimension	dbl
radius_se	Radius Standard Error	dbl
texture_se	Texture Standard Error	dbl
perimeter_se	Perimeter Standard Error	dbl
area_se	Area Standard Error	dbl
smoothness_se	Smoothness Standard Error	dbl
compactness_se	Compactness Standard Error	dbl
concavity_se	Concavity Standard Error	dbl
concave_pts_se	Concave Points Standard Error	dbl
symmetry_se	Symmetry Standard Error	dbl
fractal_dim_se	Fractal Dimension Standard Error	dbl
radius_worst	Worst Radius	dbl
texture_worst	Worst Texture	dbl
perimeter_worst	Worst Perimeter	dbl
area_worst	Worst Area	dbl
smoothness_worst	Worst Smoothness	dbl
compactness_worst	Worst Compactness	dbl
concavity_worst	Worst Concavity	dbl
concave_pts_worst	Worst Concave Points	dbl
symmetry_worst	Worst Symmetry	dbl
fractal_dim_worst	Worst Fractal Dimension	dbl

2.3 Loading in the data

```
# Load in data from file with codebook column names
data <- read_csv("data/wdbc.data", col_names = codebook[[1]], show_col_types = FALSE)
# Set diagnosis column to factor
data$diagnosis <- factor(data$diagnosis, labels = c("Benign", "Malignant"))

# Print data to check if it is loaded in correctly
pander::pander(head(data[,c(1:6, 12:16, 22:26)]))
```

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
842302	Malignant	17.99	10.38	122.8	1001
842517	Malignant	20.57	17.77	132.9	1326
84300903	Malignant	19.69	21.25	130	1203
84348301	Malignant	11.42	20.38	77.58	386.1

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
84358402	Malignant	20.29	14.34	135.1	1297
843786	Malignant	12.45	15.7	82.57	477.1

fractal_dim_mean	radius_se	texture_se	perimeter_se	area_se
0.07871	1.095	0.9053	8.589	153.4
0.05667	0.5435	0.7339	3.398	74.08
0.05999	0.7456	0.7869	4.585	94.03
0.09744	0.4956	1.156	3.445	27.23
0.05883	0.7572	0.7813	5.438	94.44
0.07613	0.3345	0.8902	2.217	27.19

fractal_dim_se	radius_worst	texture_worst	perimeter_worst	area_worst
0.006193	25.38	17.33	184.6	2019
0.003532	24.99	23.41	158.8	1956
0.004571	23.57	25.53	152.5	1709
0.009208	14.91	26.5	98.87	567.7
0.005115	22.54	16.67	152.2	1575
0.005082	15.47	23.75	103.4	741.6

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

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

```r
# Print diagnosis counts
table(data$diagnosis)
```

```
##
##    Benign Malignant
##       357       212
```

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, https://minds.wisconsin.edu/bitstream/handle/1793/59692/TR1131.pdf (accessed Sep 16, 2022).