Classification of malignancy in breast cancer using Logistic Regression

Soham Bhattachrjee

Paper Code: DSE-B2
Roll no: 193012-21-0392, Sem: VI
College Roll: 771
Reg no: 012-1111-0944-19
Asutosh College

July 12, 2022

Abstract

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error with increased accuracy. In this paper we predict the malignancy of the tumour based on the 30 characteristics of individual cells of breast cancer obtained from a minimally invasive fine needle aspirate (FNA), to discriminate benign from malignant lumps of a breast mass using Logistic Regression. We will use the outcome Benign or Malignant to predict if a new patient has a probability of developing malignancy or not, basing on the FNA data. Furthermore, our predictor will be an exciting occasion of exposing some basic concepts of Logistic Regression and implementing a code around the biomedical problem: which features are most essential in predicting malignant outcomes. In our analysis, we use the open-source tools RStudio and packages such as Tidyverse, ggplot2, caTools. After performing Logistic Regression with only the essential features included, we were able to predict the malignancy of a tumour in a patient with approximately 93% accuracy.

Keywords: Breast cancer, malignancy, classification, Logistic Regression

Contents

1 Introduction

Breast Cancer is the disease which affects women the most worldwide. From being fourth in the list of most common cancers in India during the 1990s, it has now become the first. About 1 in 28 Indian women is likely to develop breast cancer during her lifetime. The chances of developing breast cancer is more (1 in 22) for urban women than their rural counterpart (1 in 60). In India, the incidence has increased significantly, almost by 50%, between 1965 and 1985. The estimated number of incident cases in India in 2016 was 118000, 98.1% of which were females. In 2022, an estimated 287,850 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 51,400 new cases of non-invasive (in situ) breast cancer. About 2,710 new cases of invasive breast cancer are expected to be diagnosed in men in 2022. A man's lifetime risk of breast cancer is about 1 in 833. About 43,250 women in the U.S. are expected to die in 2022 from breast cancer. Women are seriously threatened by breast cancer with high morbidity and mortality. Every year, death rate increases drastically due to breast cancer. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump.

Here we predict the malignancy of the tumour based on the 30 characteristics of individual cells of breast cancer obtained from a minimally invasive fine needle aspirate (FNA), to discriminate benign from malignant lumps of a breast mass using Logistic Regression.

As prerequisites we shall be introducing the concepts of a Classification problem, Logistic Regression. And how one can use Logistic Regression to solve a Classification Problem.

1.1 Classification problems

Classification problems are the set of problems which deal with identifying the categories or the *classes* to which data points or observations belong to. In this problem we will be having observations with certain features which we are interested in and we will know which classes some of them belong to. For the rest we will be predicting the classes based on the information we already have.

Classification can be of two types: *binary classification* when we have just two target classes, or *multiclass classification* when we have more than two target classes. One of the most common algorithms to approach any

classification problem is logistic regression.

2 Problem Statement

This section will state our objective and the dataset that we will be using for this purpose.

2.1 Objective

Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumour. A tumour does not mean cancer - tumours can be benign (not cancerous) or malignant (cancerous). This analysis aims to observe which features are most helpful in predicting malignant or benign tumour and to see general trends that may aid us in understanding Breast Cancer better. Based on the 30 characteristics mentioned in the following dataset the goal is to classify whether the breast tumour is benign or malignant.

2.2 About the dataset

We will use the "Breast Cancer Wisconsin (Diagnostic)" (WBCD) dataset, provided by the University of Wisconsin, and hosted by the UCI Machine Learning Repository. In the dataset 30 characteristics of individual cells of breast cancer are studied. The dataset can be found in this GitHub link.

The first column of the dataset corresponds to the patient ID, while the last column represents the diagnosis (the outcome can be "Benign" or "Malignant" based on the type of diagnosis reported). The resulting dataset consists of 569 patients: 212 (37.2%) have an outcome of Malignancy, and 357 (62.7%) are Benign.

In detail, the dataset consists of ten real-valued features computed for each cell nucleus:

1. Radius (mean of distances from centre to points on the perimeter)
2. Texture (standard deviation of Gray-scale values)

3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (the ratio of the volume and the surface area)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension (The higher the number, the more abnormal the tissue is.)

The ten real-valued features correspond to the Mean, Standard error and the Worst or largest (mean of the three largest values of samples obtained from each individual). Column 32 contains the Benign or Malignant outcome.

3 Methodology

Our research methodology primarily consists of the following steps:

1. Cleaning the data with respect to variable names and dealing with missing values, abnormal data and so on.
2. Computing summary statistics, and performing exploratory analysis of the data.
3. Discovering relationships between data in terms of graphical visualisations and correlations between variables.
4. Performing logistic regression and computing the accuracy of the model.

3.1 Importing and cleaning the data

We start with loading the data:

```
library(tidyverse)
cancerdata <- read_csv("data.csv")
```

Then we print the dimensions of the data:

```
nrow(cancerdata) #Imported 569 rows

## [1] 569

ncol(cancerdata) #Imported 32 columns

## [1] 32
```

The structure of the data is as follows:

```
str(cancerdata)

## 'data.frame':    569 obs. of  32 variables:
##  $ id                      : int  842302 842517 84300903 84348301
   84358402 843786 844359 84458202 844981 84501001 ...
##  $ diagnosis               : Factor w/ 2 levels "B","M": 2 2 2 2 2 2
   2 2 2 2 ...
##  $ radius_mean             : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean            : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean          : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean               : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean         : num  0.1184 0.0847 0.1096 0.1425 0.1003
   ...
##  $ compactness_mean        : num  0.2776 0.0786 0.1599 0.2839 0.1328
   ...
##  $ concavity_mean          : num  0.3001 0.0869 0.1974 0.2414 0.198
   ...
##  $ concave.points_mean     : num  0.1471 0.0702 0.1279 0.1052 0.1043
   ...
##  $ symmetry_mean           : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588
   ...
##  $ radius_se               : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se              : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se            : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                 : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se           : num  0.0064 0.00522 0.00615 0.00911
   0.01149 ...
##  $ compactness_se          : num  0.049 0.0131 0.0401 0.0746 0.0246
   ...
##  $ concavity_se            : num  0.0537 0.0186 0.0383 0.0566 0.0569
   ...
##  $ concave.points_se       : num  0.0159 0.0134 0.0206 0.0187 0.0188
   ...
##  $ symmetry_se             : num  0.03 0.0139 0.0225 0.0596 0.0176
   ...
##  $ fractal_dimension_se    : num  0.00619 0.00353 0.00457 0.00921
   0.00511 ...
```

```
##  $ radius_worst             : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst            : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst          : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst               : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst         : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst        : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst          : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst     : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst           : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768
...
```

The column `id` is unnecessary to our purpose and so we drop it.

```
cancerdata$id <- NULL
```

We preview the dataset:

```
head(cancerdata)

##   diagnosis radius_mean texture_mean perimeter_mean area_mean
   smoothness_mean
## 1         M       17.99        10.38          122.80    1001.0
           0.11840
## 2         M       20.57        17.77          132.90    1326.0
           0.08474
## 3         M       19.69        21.25          130.00    1203.0
           0.10960
## 4         M       11.42        20.38           77.58     386.1
           0.14250
## 5         M       20.29        14.34          135.10    1297.0
           0.10030
## 6         M       12.45        15.70           82.57     477.1
           0.12780
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.27760         0.3001              0.14710         0.2419
## 2          0.07864         0.0869              0.07017         0.1812
## 3          0.15990         0.1974              0.12790         0.2069
## 4          0.28390         0.2414              0.10520         0.2597
## 5          0.13280         0.1980              0.10430         0.1809
## 6          0.17000         0.1578              0.08089         0.2087
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.07871    1.0950     0.9053        8.589  153.40
## 2                0.05667    0.5435     0.7339        3.398   74.08
## 3                0.05999    0.7456     0.7869        4.585   94.03
## 4                0.09744    0.4956     1.1560        3.445   27.23
## 5                0.05883    0.7572     0.7813        5.438   94.44
## 6                0.07613    0.3345     0.8902        2.217   27.19
##   smoothness_se compactness_se concavity_se concave.points_se
   symmetry_se
## 1      0.006399        0.04904      0.05373           0.01587
   0.03003
## 2      0.005225        0.01308      0.01860           0.01340
   0.01389
## 3      0.006150        0.04006      0.03832           0.02058
   0.02250
## 4      0.009110        0.07458      0.05661           0.01867
   0.05963
```

```
## 5       0.011490         0.02461       0.05688            0.01885
    0.01756
## 6       0.007510         0.03345       0.03672            0.01137
    0.02165
##   fractal_dimension_se radius_worst texture_worst perimeter_worst
    area_worst
## 1              0.006193        25.38         17.33          184.60
         2019.0
## 2              0.003532        24.99         23.41          158.80
         1956.0
## 3              0.004571        23.57         25.53          152.50
         1709.0
## 4              0.009208        14.91         26.50           98.87
          567.7
## 5              0.005115        22.54         16.67          152.20
         1575.0
## 6              0.005082        15.47         23.75          103.40
          741.6
##   smoothness_worst compactness_worst concavity_worst concave.points_
    worst
## 1           0.1622            0.6656          0.7119
    0.2654
## 2           0.1238            0.1866          0.2416
    0.1860
## 3           0.1444            0.4245          0.4504
    0.2430
## 4           0.2098            0.8663          0.6869
    0.2575
## 5           0.1374            0.2050          0.4000
    0.1625
## 6           0.1791            0.5249          0.5355
    0.1741
##   symmetry_worst fractal_dimension_worst
## 1         0.4601                 0.11890
## 2         0.2750                 0.08902
## 3         0.3613                 0.08758
## 4         0.6638                 0.17300
## 5         0.2364                 0.07678
## 6         0.3985                 0.12440
```

3.2 Descriptive statistics

The summary statistics of the dataset is given as:

```
summary(cancerdata)

##  diagnosis  radius_mean      texture_mean    perimeter_mean      area_
    mean
##  B:357      Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.
    : 143.5
##  M:212      1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu
    .: 420.3
##             Median :13.370   Median :18.84   Median : 86.24   Median
    : 551.1
```

```
##            Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean
    : 654.9
##            3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu
    .: 782.7
##            Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.
    :2501.0
## smoothness_mean   compactness_mean   concavity_mean     concave.
    points_mean
## Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.
    :0.00000
## 1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu
    .:0.02031
## Median :0.09587   Median :0.09263   Median :0.06154   Median
    :0.03350
## Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean
    :0.04892
## 3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu
    .:0.07400
## Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.
    :0.20120
## symmetry_mean     fractal_dimension_mean   radius_se        texture_
    se
## Min.   :0.1060   Min.   :0.04996        Min.   :0.1115   Min.
    :0.3602
## 1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324   1st Qu
    .:0.8339
## Median :0.1792   Median :0.06154        Median :0.3242   Median
    :1.1080
## Mean   :0.1812   Mean   :0.06280        Mean   :0.4052   Mean
    :1.2169
## 3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu
    .:1.4740
## Max.   :0.3040   Max.   :0.09744        Max.   :2.8730   Max.
    :4.8850
##   perimeter_se       area_se         smoothness_se       compactness_
    se
## Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.
    :0.002252
## 1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169   1st Qu
    .:0.013080
## Median : 2.287   Median : 24.530   Median :0.006380   Median
    :0.020450
## Mean   : 2.866   Mean   : 40.337   Mean   :0.007041   Mean
    :0.025478
## 3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu
    .:0.032450
## Max.   :21.980   Max.   :542.200   Max.   :0.031130   Max.
    :0.135400
##   concavity_se      concave.points_se   symmetry_se        fractal_
    dimension_se
## Min.   :0.00000   Min.   :0.000000   Min.   :0.007882   Min.
    :0.0008948
## 1st Qu.:0.01509   1st Qu.:0.007638   1st Qu.:0.015160   1st Qu
    .:0.0022480
## Median :0.02589   Median :0.010930   Median :0.018730   Median
    :0.0031870
## Mean   :0.03189   Mean   :0.011796   Mean   :0.020542   Mean
    :0.0037949
```

```
##  3rd Qu.:0.04205   3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu
    .:0.0045580
##  Max.   :0.39600   Max.   :0.052790   Max.   :0.078950   Max.
    :0.0298400
##    radius_worst   texture_worst   perimeter_worst     area_worst
##  Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2
##  1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3
##  Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
##  Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6
##  3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0
##  Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0
##  smoothness_worst  compactness_worst  concavity_worst  concave.points
    _worst
##  Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.
    :0.00000
##  1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu
    .:0.06493
##  Median :0.13130   Median :0.21190   Median :0.2267   Median
    :0.09993
##  Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean
    :0.11461
##  3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu
    .:0.16140
##  Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.
    :0.29100
##  symmetry_worst   fractal_dimension_worst
##  Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2822   Median :0.08004
##  Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :0.6638   Max.   :0.20750
```

In the results displayed, there are 569 records, each with 31 columns. Diagnosis is a categorical variable. Class distribution of the categorical variable: 357 benign, 212 malignant.

3.3 Univariate plots

One of the main goals of this paper is to observe which features are most helpful in predicting malignancy of a tumour in Breast Cancer. So visualising the data is going to let us do just that. We will analyze the features and try to understand which features have larger predictive value and which does not bring considerable predictive value if we want to create a model that allows us to guess if a tumor is benign or malignant.

```
library(lessR)
diagnosistable <- cancerdata$diagnosis

diagnosisdf <- data.frame(diagnosistable)

PieChart(diagnosistable, data = diagnosisdf, hole=0,
```

```
7             main="Distribution of Malignant and Benign tumours")
```

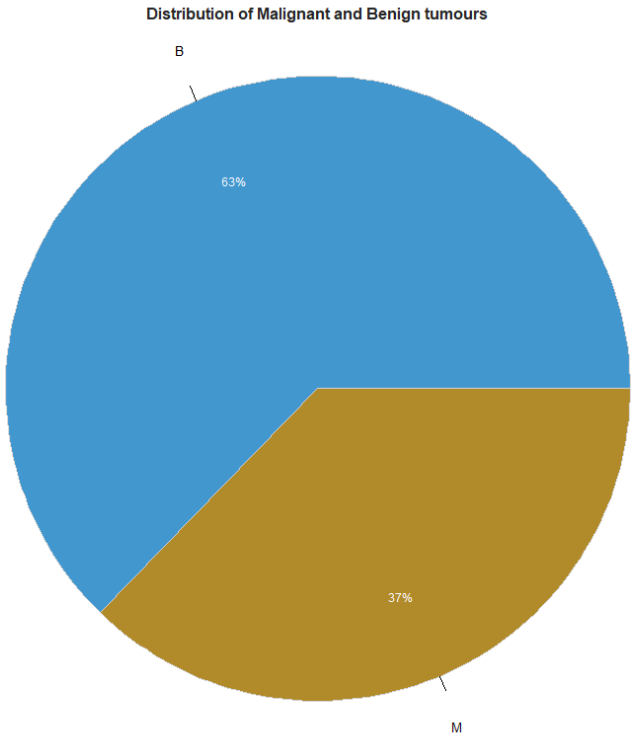

Figure 1: Distribution of benign and malignant tumours

```
## >>> Suggestions
## PieChart(diagnosistable, hole=0) # traditional pie chart
## PieChart(diagnosistable, values="%") # display %'s on the chart
## PieChart(diagnosistable) # bar chart
## Plot(diagnosistable) # bubble plot
## Plot(diagnosistable, values="count") # lollipop plot
##
## ── diagnosistable ──
##
##                    B      M     Total
## Frequencies:     357    212       569
## Proportions:   0.627  0.373     1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 36.951, df = 1, p-value = 0.000
```

M = Malignant (indicates presence of cancer cells); B = Benign (indicates absence of cancer cells)

Here we can see that 357 observations which account for 62.7% of all observations indicate the absence of cancer cells and 212 which account for 37.3% of all observations shows the presence of cancerous cell.

The percentage for malignant tumours is unusually large; this dataset

does not represent in this case a typical medical analysis distribution. Typically, there will be a considerable large number of cases that represents negative against a small number of cases that represents positives (malignant) tumor.

Following this, we observe that we have data corresponding to the 10 parameters that we listed previously. Hence it is in our interest to visualise their values with the help of a histogram across the two categories of benign and malignant.

```
library(reshape)
cancerdata_mean <- cancerdata[ ,c("diagnosis", "radius_mean", "texture_
    mean","perimeter_mean", "area_mean", "smoothness_mean", "compactness
    _mean", "concavity_mean", "concave.points_mean", "symmetry_mean", "
    fractal_dimension_mean" )]

meanlong <- melt(data=cancerdata_mean, id.vars="diagnosis")

ggplot(data=meanlong,mapping = aes(x=value)) +
     geom_histogram(bins=10, aes(fill=diagnosis),colour="Black",alpha
         =0.4) +
     facet_wrap(.~variable, scales="free_x") +
     labs(title="Histogram of the following parameters")+
     theme(plot.title = element_text(hjust = 0.5))
```

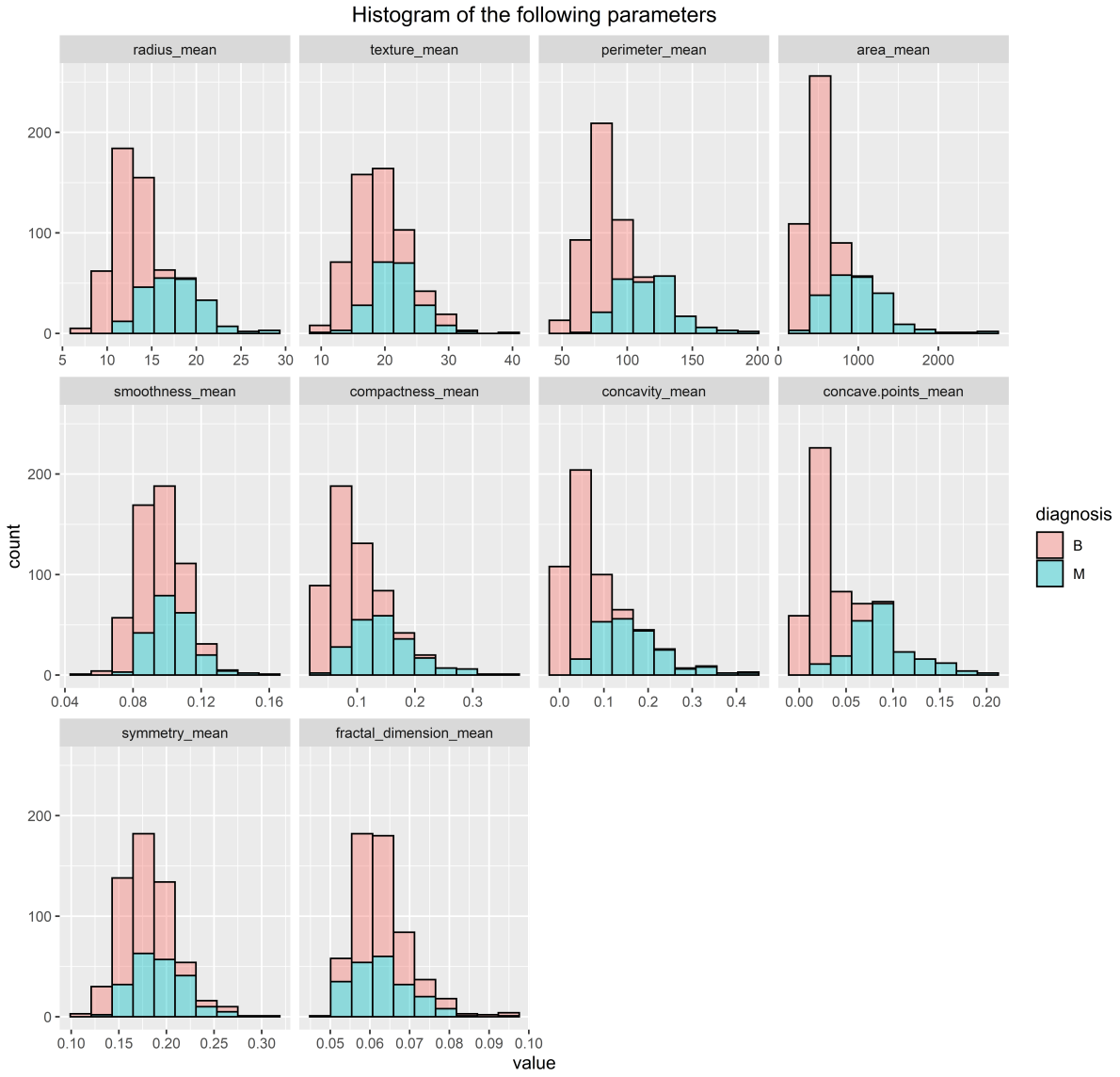

Figure 2: Histogram of the mean of the 10 concerned parameters separated by diagnosis

From the histogram of radius mean, it is observed that higher radius mean of sample cell nuclei corresponds to higher probability of malignancy of the tumour. As for the malignant cell samples, we can see that the radius mean follows normal distribution with mean 17.35 units. The total data is positively skewed which shows lower number of people have high tumour radius mean in general.

The histogram of the texture mean shows that benign and malignant tumour cells superimpose in terms of texture mean. The entire data on texture mean is almost symmetric in nature, with a slight positive skewness.

The distribution of perimeter mean of tumour cells is similar to that of radius mean, where high perimeter mean corresponds to higher probability of malignancy. The distribution of malignant tumour cells is positively skewed which means lower number of people have very high perimeter mean.The entire distribution of the perimeter mean data is also highly skewed in nature.

Area mean of tumour cells also follows a positively skewed distribution. Higher area mean of tumour cells corresponds to higher probability of malignancy of the tumour. However, the malignant cells follow a normal distribution with mean at 1062.5 units.

Smoothness mean of both benign and malignant tumour cells superimpose on each other. They follow a normal distribution with mean at 0.1 units.

From the histogram, we can see benign and malignant tumour cells superimpose on each other in the range 0.00 to 0.27 units. Very high compactness mean corresponds to high probability of malignancy.

Concavity mean follows a positively skewed distribution, where lower number of cell nuclei have higher concavity mean. Moreover higher values of concavity mean also corresponds to higher probability of malignancy of tumour cells.

Concave points mean also follow positively skewed distribution. Higher concave points mean corresponds to higher probability of malignancy of tumour. The malignant cells themselves follow normal distribution with mean at 0.1 units.

Symmetry mean of both benign and malignant tumour cells superimpose. They follow a normal distribution with mean at 0.18 units.

Fractal dimension also approximately follows a symmetric distribution. The malignant cells however follows positively skewed distribution. However in general, the benign and malignant cell observations superimpose.

```
1   cancerdata_se <- cancerdata[ ,c("diagnosis", "radius_se", "texture_se",
        "perimeter_se", "area_se", "smoothness_se", "compactness_se", "
        concavity_se", "concave.points_se", "symmetry_se", "fractal_
        dimension_se" )]
2
3   selong <- melt(data=cancerdata_se, id.vars="diagnosis")
4
5   ggplot(data=selong,mapping = aes(x=value)) +
6       geom_histogram(bins=10, aes(fill=diagnosis),colour="Black",alpha
            =0.4) +
```

```
        facet_wrap(.~variable, scales="free_x") +
        labs(title = "Histogram of the standard error of the parameters") +
        theme(plot.title = element_text(hjust = 0.5))
```

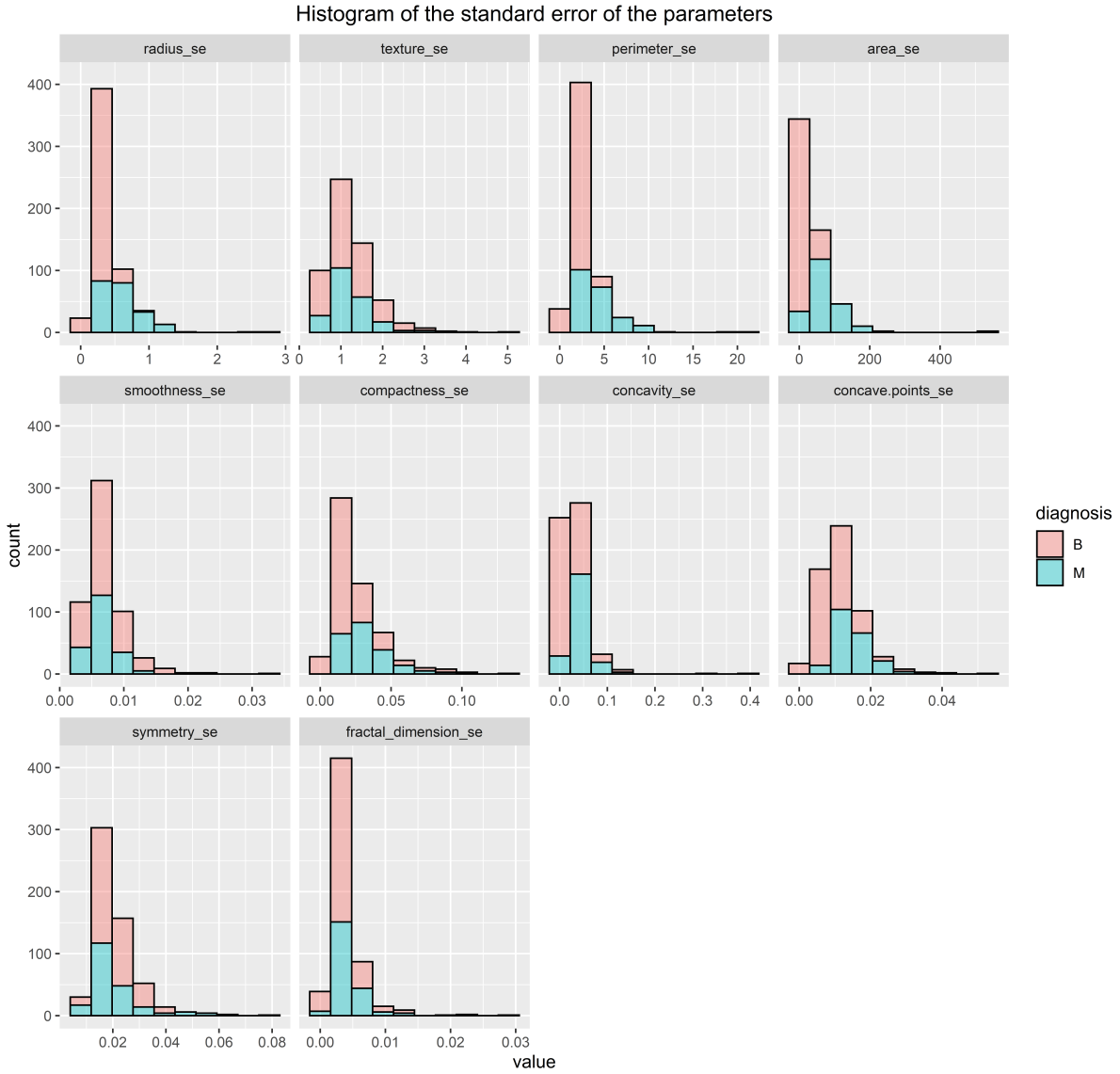

Figure 3: Histogram of the standard error of the 10 concerned parameters separated by diagnosis

```
    cancerdata_worst <- cancerdata[ ,c("diagnosis", "radius_worst", "
        texture_worst","perimeter_worst", "area_worst", "smoothness_worst",
        "compactness_worst", "concavity_worst", "concave.points_worst", "
        symmetry_worst", "fractal_dimension_worst" )]

    worstlong <- melt(data=cancerdata_worst, id.vars="diagnosis")

    ggplot(data=worstlong,mapping = aes(x=value)) +
        geom_histogram(bins=10, aes(fill=diagnosis),colour="Black",alpha
            =0.4) +
```

```
7         facet_wrap(.~variable, scales="free_x") +
8         labs(title = "Histogram of the worst value of the parameters") +
9         theme(plot.title = element_text(hjust = 0.5))
```

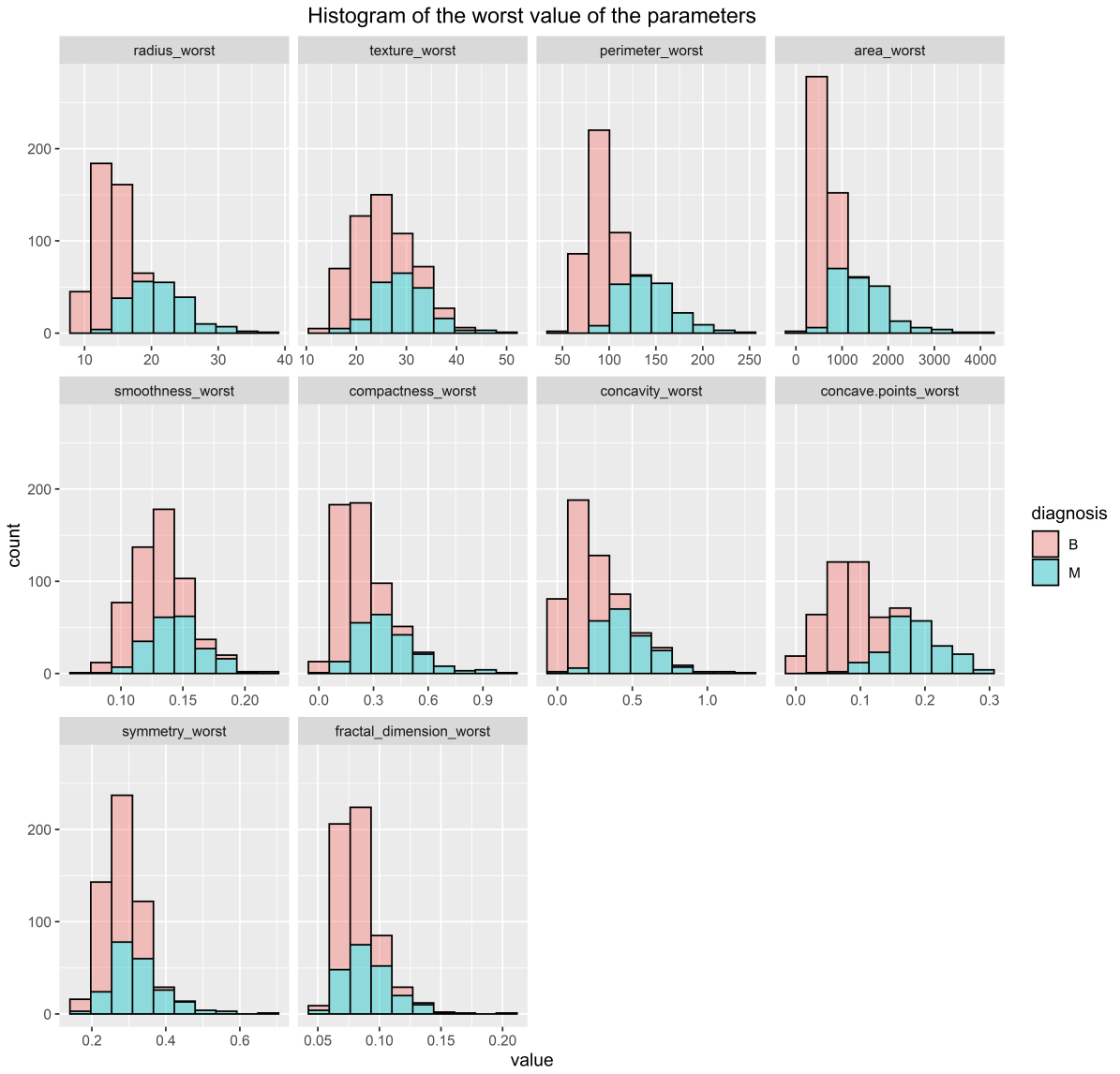

Figure 4: Histogram of the worst values of the 10 concerned parameters separated by diagnosis

We can observe that most of the features that we see are normally distributed.

Comparison of radius distribution by malignancy shows that there is no perfect separation between any of the features. Although we do have fairly good separations for `concave.points_worst`, `concavity_worst`, `perimeter_worst`, `area_mean`, `perimeter_mean`. We do have as well tight superposition for some of the values, like `symmetry_se`, `smoothness_se`.

3.4 Graphical visualisation of relationships between multiple variables

We are also interested in how the 30 predictor variables relate to each other. To see bivariate relationships among these 30 predictor variables, we will look at their correlation coefficients.

```
1  library("corrplot")
2
3  corMatMy <- cor(cancerdata[,2:31])
4  corrplot(corMatMy, method = "square", order = "hclust", tl.cex = 0.7,
       tl.col="black")
```

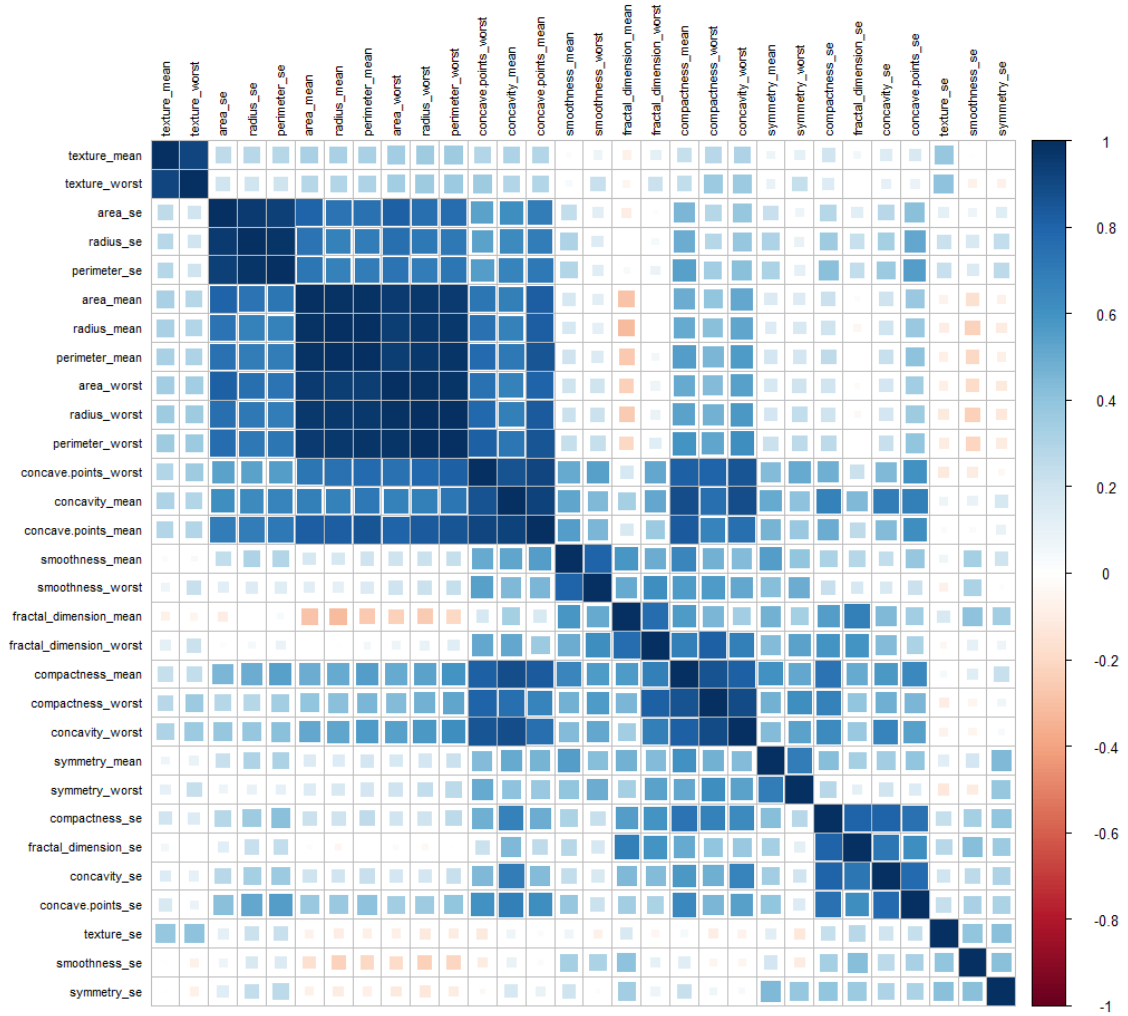

Figure 5: Correlation matrix of the vairables

There are quite a few variables that are correlated. Often we have features that are highly correlated and those provide redundant information. By eliminating highly correlated features we can avoid a predictive

bias for the information contained in these features. This also shows us, that when we want to make statements about the biological/ medical importance of specific features, we need to keep in mind that just because they are suitable to predicting an outcome they are not necessarily causal - they could simply be correlated with causal factors. We will now remove all features with a correlation higher than 0.9, keeping the feature with the lower mean.

```
library(caret)
highlyCor <- findCorrelation(corMatMy, cutoff=0.9,verbose=T,names=T)

## Compare row 7  and column  8 with corr  0.921
##   Means:  0.571 vs 0.389 so flagging column 7
## Compare row 8  and column  28 with corr  0.91
##   Means:  0.542 vs 0.377 so flagging column 8
## Compare row 23  and column  21 with corr  0.994
##   Means:  0.48 vs 0.367 so flagging column 23
## Compare row 21  and column  3 with corr  0.969
##   Means:  0.446 vs 0.359 so flagging column 21
## Compare row 3  and column  24 with corr  0.942
##   Means:  0.414 vs 0.353 so flagging column 3
## Compare row 24  and column  1 with corr  0.941
##   Means:  0.39 vs 0.349 so flagging column 24
## Compare row 1  and column  4 with corr  0.987
##   Means:  0.35 vs 0.347 so flagging column 1
## Compare row 13  and column  11 with corr  0.973
##   Means:  0.372 vs 0.346 so flagging column 13
## Compare row 11  and column  14 with corr  0.952
##   Means:  0.323 vs 0.347 so flagging column 14
## Compare row 22  and column  2 with corr  0.912
##   Means:  0.224 vs 0.357 so flagging column 2
## All correlations <= 0.9
```

The following 10 columns have been flagged for removal:

```
highlyCor

##  [1] "concavity_mean"       "concave.points_mean" "perimeter_worst"
##  [4] "radius_worst"         "perimeter_mean"      "area_worst"
##  [7] "radius_mean"          "perimeter_se"        "area_se"
## [10] "texture_mean"

cdata <- cancerdata
cancerdata[,highlyCor] <- NULL
colnames(cancerdata)

##  [1] "diagnosis"              "area_mean"
##  [3] "smoothness_mean"        "compactness_mean"
##  [5] "symmetry_mean"          "fractal_dimension_mean"
##  [7] "radius_se"              "texture_se"
##  [9] "smoothness_se"          "compactness_se"
## [11] "concavity_se"           "concave.points_se"
## [13] "symmetry_se"            "fractal_dimension_se"
## [15] "texture_worst"          "smoothness_worst"
## [17] "compactness_worst"      "concavity_worst"
## [19] "concave.points_worst"   "symmetry_worst"
```

```
## [21] "fractal_dimension_worst"

ncol(cancerdata)

## [1] 21
```

We have removed the redundant predictor variables from our dataset, and we are now left with only the 21 essential features that will help us in predicting the malignancy of a Breast Cancer tumuor.

3.5 What is Logistic regression?

A logistic regression is a special case of regression analysis which is used whenever the dependent variable is categorical.

Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables, and not the actual value of the variable. This type of statistical model is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In order to do this we use the sigmoid function (a special case of the logistic function),

$$\text{sig}(t) = \frac{1}{1+e^{-t}} \tag{3.1}$$

This is a positive valued function bounded between 0 and 1, as seen in Figure 1 below.

In our case, when the linear regression of y on x is given by $\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k$, we enter this y in place of t in the sigmoid function, and then we determine the coefficients $b_1, b_2, \ldots b_k$.

$$f(x) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(b_0+b_1\cdot x_1+b_2\cdot x_2+\cdots+b_k\cdot x_k)}} \tag{3.2}$$

In logistic regression, a logit transformation is applied on the odds which is the probability of success divided by the probability of failure. This is also known as the log odds and this logistic function is represented by the formula

$$\begin{aligned} \text{Logit}(f(x)) &= \frac{1}{1+e^{-f(x)}} \\ \ln\left(\frac{f(x)}{1-f(x)}\right) &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k \end{aligned} \tag{3.3}$$

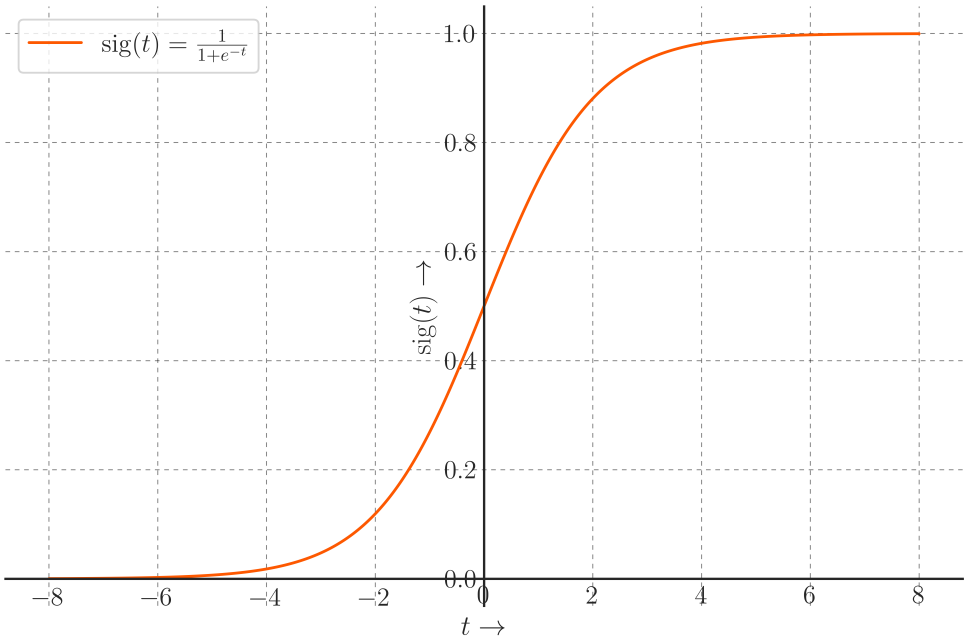

Figure 6: Sigmoid function

In this logistic regression equation, Logit($f(x)$) is the dependent or response variable and x is the independent variable. The coefficients in this model are commonly estimated via maximum likelihood estimation (MLE). This method tests different values of b_i through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than 0.5 will predict 0 while a probability greater than 0.5 will predict 1. After the model has been computed we used the confusion matrix to check for the accuracy of the model.

4 Performing logistic regression

In order to predict the diagnosis given the data, we need to fit a logistic regression model to the data. In order to do that we need to split the data set into two data sets: a training set and a test set.

The purpose of the training set is to train the model on this data. Then we test the model on the ”unknown data” which is in the test set.

By generating predictions on the test set and comparing with the actual values, we can generate a confusion matrix and calculate the accuracy of the model.

```
    library(caTools)
    set.seed(123)
    split <- sample.split(cancerdata$diagnosis, SplitRatio = 0.7)
    training_set <- subset(cancerdata, split==TRUE)
    test_set <- subset(cancerdata, split==FALSE)
```

After we split the dataset, we standardise the data to 0 mean and 1 standard deviation. To do this, we iterate over the columns of the training set, then compute the mean and standard deviation of each column, using which we standardise that column of the training set and the corresponding column of the test set. The reason for doing this is that we do not want to standardise the test set with the statistics of the test set as that may lead to overfitting.

```
    for(i in seq(2,21))
    {
        col_mean <- mean(training_set[,i])
        col_se <- sd(training_set[,i])

        training_set[,i] <- (training_set[,i] - col_mean)/col_se
        test_set[,i] <- (test_set[,i] - col_mean)/col_se
    }
```

Next we fit a generalised linear model, where we specify that we want to predict `diagnosis` with the help of the other variables. The parameter `binomial` signifies that it is a logistic regression problem.

```
    #Fitting Logistic Regression to Training Set
    classifier <- glm(formula = diagnosis~., family=binomial,
                      data=training_set)

    #Predicting the test set result
    prob_pred <- predict(classifier, type = "response", newdata = test_set
        [-1])
    y_pred <- ifelse(prob_pred>0.5,"M","B")
    y_true <- test_set$diagnosis
```

Next we generate the confusion matrix. It is a simple table which is used to describe the performance of a classifier by providing the numbers of correctly and incorrectly classified observations across the categories. For a binary classification problem the table looks as below:

Where,

- TN: True negative
- FP: False positive

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

Table 1:

- FN: False negative
- TP: True positive

Following this, accuracy is computed as:

$$\text{accuracy} = \frac{TP + TN}{TN + FP + FN + TP} \tag{4.1}$$

In code:

```
cm <- table(y_true, y_pred)
cm

##    B    M
## B 101  6
## M   5  59

accuracy <- 160/171
accuracy

## [1] 0.9356725
```

Thus we get $\approx 93\%$ accuracy.

5 Concluding remarks

After cleaning and analysing the data we applied a logistic regression model in order to predict, whether, we know the required parameter values, the cancer diagnosis is malignant or benign. In that regard, we have achieved a satisfactorily high 93% accuracy.

Acknowledgements

I would sincerely like to thank **Oindrila Bose, Asutosh College, Kolkata** for her valuable inputs during the completion of this paper. Without her valuable inputs and insights, the paper would not have made it to completion.

I would also like to thank my parents **Subrata Bhattacharjee** and **Rubi Bhattacharjee** for their unfettered support throughout the duration of the project. It was their efforts which enabled me to complete my project without having to think of anything else.

References

1. Logistic Regression for malignancy prediction in cancer - Luca Zammatoro, Towards Data Science, Dec 23, 2019

2. Hosmer, D., Sturdivant, R. and Lemeshow, S., 2013. Applied logistic regression. New York , Toronto: Wiley.

3. Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO. 7.

4. Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

Declaration: *I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.*

Soham Bhattacharjee