MATH2021-1 High-dimensional statistics

Faculty of Applied Science

Generalized linear models and classification

Teachers :
Gentiane Haesbroeck

Students :
Romain LAMBERMONT, s190931
Arthur LOUIS, s191230

December 21, 2022

Contents

List of Figures

List of Tables

1 Introduction

For this project, we have decided to change our dataset as the one we used for the first project had no binary indicator that was really interesting. The chosen dataset is the Brest Cancer Diagnostic in Wisconsin[1]. This dataset is a collection of data that has been compiled for the purpose of studying breast cancer. It contains various characteristics of breast cancer tumors, such as radius, texture, perimeter, and smoothness, as well as the diagnosis (malignant or benign). This last variable is the target variable we'll use for this project. It's value is contained in the first column `malignant` and is a binary indicator :

- 1 : The breast cancer is malignant
- 0 : The breast cancer is beningn

The other variables contained in the dataset are describing the tumor and are the following :

- `radius_mean` : mean of distances from center to points on the perimeter
- `texture_mean` : standard deviation of gray-scale values
- `perimeter_mean` : mean size of the core tumor
- `area_mean` : mean area of the core tumor
- `smoothness_mean` : mean of local variation in radius lengths
- `compactness_mean` : mean of $\frac{\text{perimeter}^2}{\text{area}-1}$
- `concavity_mean` : mean of severity of concave portions of the contour
- `concave points_mean` : mean for number of concave portions of the contour
- `symmetry_mean` : mean of simmilarity between left and right part of the tumor
- `fractal_dimensions_mean` : mean for "coastline approximation" - 1, represents the self-similarity in the tumor's structure

We have randomly sampled 500 lines from the original dataset, which has 569 observations with 357 benign and 212 malignant tumors (62.74% benign). After sampling, we have 315 benign and 185 malignant tumors in our set (63%) keeping our ratio benign/malignant fairly similar.

Our goal in this project is going to classify the observations of the quantitative variables between the benign and malignant tumors.

2 Preliminaries for the supervised classification

To determine whether some information about the classification might be available in the explanatory variables, a variety of graphical and statistical summaries can be used :

[1] This data can be downloaded here : https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

2.1 Statistical summary

A first way to approach the problem is by showing a statistical summary of our dataset splitted in the two tumor groups.

	benign	malignant
radius_mean	12.123	17.556
texture_mean	17.931	21.557
perimeter_mean	77.941	116.026
area_mean	461.304	989.846
smoothness_mean	0.093	0.103
compactness_mean	0.081	0.145
concavity_mean	0.047	0.162
concave.points_mean	0.026	0.088
symmetry_mean	0.175	0.192
fractal_dimension_mean	0.063	0.062

	benign	malignant
radius_mean	3.333	10.444
texture_mean	14.731	13.264
perimeter_mean	147.290	487.167
area_mean	18839.566	140044.214
smoothness_mean	0.000	0.000
compactness_mean	0.001	0.003
concavity_mean	0.002	0.006
concave.points_mean	0.000	0.001
symmetry_mean	0.001	0.001
fractal_dimension_mean	0.000	0.000

By looking at these tables, we can get a first insight that the variables have some kind of correlation between the classification of the tumor and the explanatory variables as larger values seems to result in malignant tumors. We can jump into more graphical analysis to show in more detail the distribution of the explanatory variables relative to their classification.

2.2 Box and density plots

By using the box and density plots, we can easily visualize whether the two groups have different distributions for each explanatory variable by showing in a more visual way the previously computed summary values.

As can clearly be seen in the 2 figures here under, there's a logical relationship between the tumor's size and it's classification between benign and malignant. Indeed, in the boxplots and density plots, we clearly see that the mean of each explanatory variable is distributed along smaller values for benign tumors than for malignant tumors. This relationship seems logical to us as the bigger the tumor, the more dangerous it gets. These facts, are all true except for 1 variable, `fractal_dimension_mean`. All these observations results in giving us some confidence that we'll be able to classify the observations between the two groups using the majority of the explanatory variables.

Figure 1: Boxplot of explanatory variables separated by groups

Figure 2: Density plot of explanatory variables separated by groups

2.3 Correlation matrix and scatterplot

Another method to assess if the explanatory variables can be used to classify the data between the 2 classes is a correlation matrix :

(a) Malignant tumors

(b) Benign tumors

Figure 3: Correlation matrices

An even more visual way to assess the correlation between the class of the observation and the explanatory variables is a scatter where points are color coded respectively to their class :

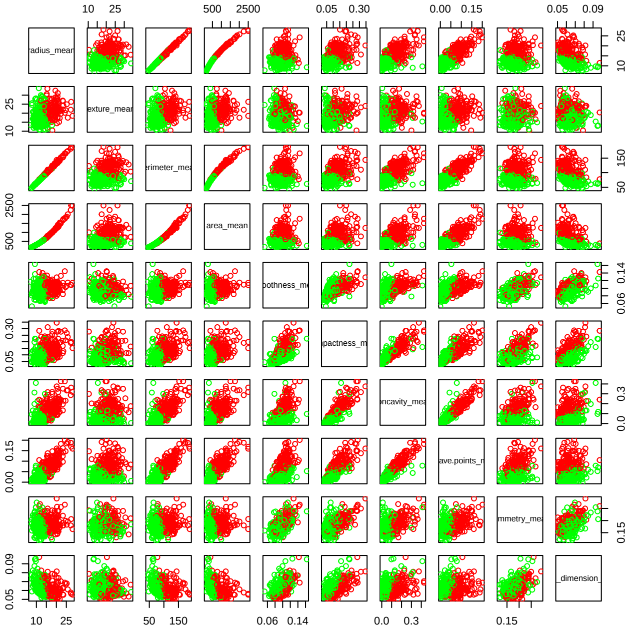

Figure 4: Caption

By looking at the figures here above, we again clearly see a correlation between the class of an observation and the explanatory variables values. Indeed, the values corresponding to benign tumors are concentrated to the left, representing smaller values. We can even see a clear linear correlation for the variables relative to the tumor's size.

2.4 Adequacy conclusion

By exploring the data in these ways, we can gain a better understanding of whether the explanatory variables contain useful information for classifying observations into the two groups defined by the binary indicator and knowing the data has been collected on several period of times with a lot of observations, we can then clearly conclude that we can use our classification techniques on this dataset.

3 Preparation of the classification rules

3.1 Logistic regression

The `glm()` function in R uses the coefficients of the model to make predictions about the response variable. The `glm()` function fits a generalized linear model, which is a type of statistical model that allows you to model the relationship between a response variable and one or more predictor variables.

The `glm()` function estimates the coefficients of the model by maximizing the likelihood of the data given the model. The likelihood is a measure of how well the model fits the data. The `glm()` function estimates the coefficients such that the likelihood is maximized.

Once the coefficients of the model have been estimated, the `glm()` function can use them to make predictions about the response variable. Specifically, the `glm()` function uses the coefficients to calculate the predicted value of the response variable for a given set of predictor variables.

It is important to select the relevant column to have a good model. In our case, we decided to keep all the variables because looking at the figure **??**, all variables separate the binary indicator in 2 distinct areas.

GLM	Intercept	radius_mean	texture_mean	perimeter_mean
	-18.1368587	1.3499735	0.4830149	-0.6073519

GLM	area_mean	smoothness_mean	compactness_mean	concavity_mean
	0.0441134	70.3754716	-11.2074699	16.9460664

GLM	concave.points_mean	symmetry_mean	fractal_dimension_mean
	76.4346011	18.9287531	48.8811318

Table 1: GLM coefficients

This gives us the following model :

$$\begin{aligned}\text{pred} &= g^{-1}\mathbf{X}\beta \\ &= g^{-1}(-18.1368587 + 1.3499735 \times \text{radius\_mean} + 0.4830149 \times \text{texture\_mean} \\ &+ ... + 48.8811318 \times \text{fractal\_dimension\_mean})\end{aligned}$$

3.2 LDA

Linear Discriminant Analysis (LDA) is a classification algorithm that is used to predict the class label of a sample based on its features. It is a supervised learning algorithm that assumes that the features follow a Gaussian distribution within each class.

To classify a sample using LDA, the algorithm estimates the class-conditional probability density functions (PDFs) for each class based on the training data. The class-conditional PDFs are modeled as multivariate normal distributions with class-specific means and a common covariance matrix.

Given a new sample, LDA computes the posterior probabilities of the sample belonging to each class using Bayes' theorem. The posterior probability for each class is calculated as the product of the prior probability of the class and the likelihood of the sample belonging to the class, normalized by the sum of the likelihoods over all classes.

The class label with the highest posterior probability is then assigned to the sample. In other words, LDA assigns the sample to the class that has the highest probability of generating the sample's features, given the class labels.

Using all the quantitative variables, the LDA procedure resulted in the most discriminant direction described by the coefficients in the following table :

LD1	radius_mean	texture_mean	perimeter_mean	area_mean
	3.023569425	0.113449782	-0.363161854	-0.004811602

LD1	smoothness_mean	compactness_mean	concavity_mean	concave.points_mean
	1.120549699	-0.528340038	6.051544650	29.491683288

LD1	symmetry_mean	fractal_dimension_mean
	6.936328511	13.254548376

Table 2: LDA1 coefficients

This gives us the following model :

$$\text{pred} = 3.023 \times \text{radius\_mean} + 0.1134 \times \text{texture\_mean} + ... + 13.2445 \times \text{fractal\_dimension\_mean}$$

When we use all the variables, we have a discriminant power of -0.8025743. The power is negative, which means that the model is more accurate on the "malignant" cases.

LD1	radius_mean	texture_mean	perimeter_mean	area_mean
	3.042813457	0.113662563	-0.376842430	-0.004409713

LD1	smoothness_mean	compactness_mean	concavity_mean	concave.points_mean
	4.935756229	3.160583782	6.091609258	29.667446169

Table 3: LDA without useless columns

This gives us the following model :

$$\begin{aligned}\text{pred} = 3.043 \times \text{radius\_mean} + 0.1137 \times \text{texture\_mean}\\ + ... + 69.667 \times \text{concave.points\_mean}\end{aligned}$$

By removing the 2 columns "symmetry_mean" and "fractal_dimension_mean" we have a discriminant score of -0.7978517 which is better than in the previous case.
We chose these two column by looking at the figure **??** and looking for the variables the less correlated.

3.3 ROC

By computing the scores of each of the 2 chosen models thanks to the `pROC` package, we get these ROC curves :

Figure 5: ROC curves

The areas under the curves for GLM and LDA are respectively 0.99 and 0.98. We can than compute our threshold using the Youden index, for GLM and LDA respectively, these thresholds are 0.2952 and 0.7032.

4 Classification of the test data set

To compare the performance of the LDA model and the GLM model, we split the data into a training set (80% of the data) and a test set (20% of the data). We used the training set to fit both models, and then used the test set to evaluate their performance.

We evaluated the performance of the models using the area under the receiver operating characteristic (ROC) curve, which is a measure of the model's ability to distinguish between the two classes. We computed the ROC curves for both models using the `roc()` function from the `pROC` package, and then computed the AUC for each model using the `auc()` function from the `pROC` package.

Surprisingly, the results showed that the GLM model had a higher AUC on the test set (0.99) compared to the LDA model (0.98). While the AUC is an important measure of model performance, it is not the only factor to consider when choosing a model.

One potential reason for the higher AUC of the GLM model is that it is more flexible than the LDA model, as it allows for non-linear relationships between the predictor variables and the response variable. However, this flexibility also comes with the risk of overfitting, which can lead to poor generalization to new data.

On the other hand, the LDA model is a simpler and more interpretable model, as it assumes a linear relationship between the predictor variables and the response variable. While it may have a lower AUC on the test set, it is likely to have better generalization performance on new data.

Overall, our results suggest that the LDA model may be a better choice for this study compared to the GLM model, despite the higher AUC of the GLM model on the test set. The simpler and more interpretable nature of the LDA model may outweigh the slightly higher performance of the GLM model on the test set.

However, the 2 models have almost the same performance, LDA has 5 mistakes and GLM has 6 mistakes on the test set which contains 100 observations.