Predicting Lung Cancer with Machine Learning

October 7, 2022

Sede	Max cites	H-Index
Chile	257.72	21.39
Leeds	165.77	19.68
Sao Paolo	71.00	11.50
Stockholm	134.19	14.33
Morelia	257.56	17.67

1 Introduction

Lung cancer is the most common form of cancer in the world. Over 200.000 new cases are being diagnosed every year making up to 13% of of all cancer diagnosis. Almost half of the patients with lungcancer die after a year after the diagnosis. As for all forms of cancer, an early diagnosis increases the survival rate drasticaly. Earlier stages of lung cancer are easier to treat and can decrease the death rate by 14% to up to 20%[1]. Thus, applying machine learning, appears to be reasonable use case for the diagnosis of lung cancer, since it can achieve high effectivnes.

In this paper, I will try to determine whether a patient with certain characteristics has lung cancer or not.

1.1 Report structure

The report will continue with the section 2, *"Problem Statement"*, where the data set is presented and all the data points are clarified. Section 3, *"Methods"*, will explain how and why the feature selection was done and how the data sets were constructed. Furthermore, the two machine learning methods, *Logistic Regression* and *Support Vector Machine (SVC)*, will be discussed. In section 4, *"Results"*, the results are evaluated and section 5, *"Conclusion"*, will draw a final conclusion.

2 Problem Formulation

2.1 Data set

The data set consists of 309 entries in an Excel spreadsheet, with patients between 21 and 87 years old. Each data point describes a patient with different characteristics. The data set is licensed under the *CCO: Public Domain* and was found on *Kaggle* [2]. For transparency, this dataset is used in another project found *here* [3], but as this work was done independently, it has not been copied in any way.

The data set is complete and has no missing features. Each data point is mapped with 16 different attributes (gender, age, smoking, yellow fingers, chronic disease, *see appendics for more...*). Additionally, I was also able to find 33 duplicates, which I have eliminated. As you may have already noticed, the dataset is not too large, now containing only 276 entries. How this problem was dealt with will be further explained in the "Data Splitting" section.

2.2 Label and feature clarification

The label for this prediction is whether a patient has lung cancer or not (yes/no), which means that this problem is a classification problem. I decided to change the dataset for the labels from *yes* and *no* to *-1* and *1* to make the prediction easier and because I will use the SVC hinge loss function in the second part. All attributes except gender and age are represented with *1* or *2*, which corresponds to *no* and *yes* respectively. The gender is represented with *M* or *F*, which means male or female respectively. The age is the actual age as an integer. The histogram shows that the data is largely balanced, except for a few outliers such as *fatigue* and *lung cancer*. Since I will be using Logistic Regression, these outliers will not significantly affect our results [4][5].

3 Methods

3.1 Feature selection and engineering

For the features, I selected all attributes except for the lung cancer column, as it is used for the label. For the selected features, I have changed the data points from *1* and *2* to *0* and *1* for convention purposes and ease of understanding. It is important to note here that with the exception of age, all characteristics can be represented in binary, meaning, a normalisation is not required. After looking at the correlation heat map of the characteristic, I could not identify any bad data sets. All data points appear to be in the range of -0.75 and 0.75, which is a common indicator of correlation. Also, all the identifiers, gender, age, smoking, yellow fingers, (...) seem intuitively important in determining whether a patient has lung cancer or not. After some manual tests with *Variance Inflation Factor (VIF)* I could confirm that all 15 characteristics seem to improve the accuracy of the prediction.

3.2 Data Splitting

At first the data is split into two parts, the validation/training set and the test set. I chose the ratio to be 80 to 20 which was discussed in the lecture. This helps us to split the whole data into two part, so that the test set is not used in neither the training set nor the validation set. This was achieved by using the $train_{testsplit}$ function of the *sklearn* library. The remaining part (80%) of the data set will be used to train and validate the machine learning models.

As mentioned earlier, the data set is not very large. I decided to use the k-fold cross-validation method. This method is suitable for data sets with fewer entries, to split it up into training and validation sets. The algorithm works by randomly dividing the datasets into equal parts and using one set as the test set and the rest as the training set. I decided to split the data into 5 parts (k=5) to achieve an 80 to 20 (training to validation) ratio,

which is very common. Our goal is to classify whether a patient has lung cancer. This can be described as a binary classification problem.

3.3   Logistic Regression

For the first machine learning model, I chose Logistic Regression because it seemed reasonable for a simple binary classification problem. Logistic Regression uses a linear hypothesis space and works by setting a limiter between the data points. In a 2D space, you would put a line between the given data points. In 3D space, the points would be separated by a plane, and in higher dimensions you would use a hyperplane to describe the separation.

3.3.1   Loss function for Logistic Regression

For the loss function, I chose the Logistic Loss function, as it is a very common and proven function for Logistic Regression and was easy to use as it is already implemented in the used library.

3.4   SVC

For the second machine learning model, we chose to use the Support Vector Machine (SVM) classifier class, Support Vector Classification (SVC), for this simple binary classification task. For SVC, it also uses a linear hypothesis space that maps $h(x) = w^T(x)$, identical to that of the Logistic Regression method used previously. The decision to opt for this method was so that we can compare the classification methods and evaluate which has better performance. We did not choose to make use of LinearSVC or SGDClassifier over the basic SVC due to the dataset being not too large.

3.4.1   Loss function for SVC

We also decided on using hinge loss to calculate the loss of our SVC method. The motivation for making use of hinge loss as the function for this method % hinge loss function ,% is mainly because it is widely known as the loss function tailored for SVM and also is easily accessible with sklearn.metrics. Furthermore, it also makes sense to use this loss function for this binary classification task. A visualization and representation of the hinge loss function is shown in the following figures.

From these figures, we can see that for observations that are of a margin distance of greater than or equal to 1, the hinge loss is valued at zero. While for observations of margin distance less than 1, the hinge loss value incurs a loss the increases linearly. To put simply, while the SVC bears the similarity with Logistic Regression in that it aims to separate both classes with a line, the difference lies in this hinge loss function, that aims to maximize the margin distance between each data point and the separating line.

4   Results

To evaluate and compare the two models, we have calculated the errors and the accuracy scores for each training and validation sets, obtaining the following results shown in the charts below.

u

As we can see from the charts and table above, both methods performed well in this binary classification to predict persons with lung cancer, with the training and validation accuracies for both Logistic Regression and SVC valued above 90%. It can also be observed that for both training and validation, SVC performed better than the Logistic Regression model, with a 93.5% training and 91.7% validation accuracy, compared to the 92.5% training and 90.9% validation accuracy of the latter. Additionally, from our results, we can clearly see how the absolute errors for SVC using hinge loss is significantly smaller than that of Logistic Regression. From these results, we therefore come to the decision that SVC is the better method for this prediction of lung cancer binary classification task.

5   Conclusion

Conclusion: briefly summarise the report and interpret the results; discuss if the obtained results seem to be optimal or if there is room for improvement speculate about future directions on how to further

6   References

- [1] Lung cancer fact sheet website: `https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet`

- [1] Data set from Kaggle: `https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer`

- [2] Other Kaggle project with same data set: `https://www.kaggle.com/code/gaganmaahi224/lung-cancer-5ml-models-full-analysis-plotly`

- [3] How to handle unbalanced sets tutorial : `https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html`

- [4] Unbalanced data in Logistic Regression: `https://stats.stackexchange.com/questions/6067/does-an-unbalanced-sample-matter-when-doing-logistic-regression`

- [5] Heatmap Tutorial *Medium* an seaborn library: `https://medium.com/@szabo.bibor/how-to-\\create-a-seaborn-correlation-heatmap-in-python-834c0686b88e`

7   Code Appendics

Sede	Max cites	H-index
Chile	257.72	21.39
Leeds	165.77	19.68
Sao Paolo	71.00	11.50
Stockholm	134.19	14.33
Morelia	257.56	17.67