Predictung Lung Cancer with Machine Learning

September 23, 2022

1 Problem Formulation

In this paper, I will try to predict, if a patient with specific attributes has lung cancer or not.

1.1 Datapoints

The dataset consists of 309 entries in an execel sheet, with patients between 21 and 87 years. Each datapoint describes a patients with various attributes. The dataset is licsensed under the *CCO: Public Domain* and was found on *Kaggle*. For transparency reasons, this dataset is used in another project which can be found on *here*, but since this this work was done independantly it was by no means copied.

The dataset is complete and has no missing features. Each datapoint is depicted with 16 different attributes.

'Print table head'

The label for this prediction will be wether or not a patient has lungcancer (yes/no), which implys that this problem is a classification problem. I chose to change the dataset for the labels from 'yes' and 'no' to -1 and 1, for easier prediction and because I will be using the SVC hinge loss funciton in the second part.

All attributes except for the gender and age are represented with 1 or 2 which correspond to no/yes respectively. The Gender is represented with M or F which corresponds to male and female, respectively. The age is the actual age as a number.

The histogramm, shows that the data is mostly balanced except for few outliers like *Fatigue* and *Lung cancer*. Since I will be using logistic regression these outliers will not impact our results significantly.

Additionally, I was able to find 33 duplicates which I got rid of. As one may have already realized, the dataset is not all too big, as there are just 276 entries.

1.2 Feature selection

For the features I selected all attributes apart from the lungcancer column, as it is used for the label. For the selected features I chose to change the datapoint from 1 and 2 to 0 and 1 for convention and for the sake of understandability. Important to note here, is that except for the age, all features are representable in binary, meaning a normalization is not needed. After looking at the correlation heatmap of the feature, I could not identify any bad datasets. All datapoints seem to be in range of *-0.75* and *0.75* which is a common indicator for correlation. All labels, gender, age, smoking, yellow fingers, (. . .) intuitively seem to be important as features of wether a patient has lung cancer or not. After some minor testing with Variance Inlfation Factor (VIF) I could validate that all 15 features seem to help the accuracy of the prediction.

1.3 Machine learning model

Our goal is to classify wether a patient has lungcancer. This can be described as a binary classification problem.

For the first machine learning model I chose a logistic regresion, as it seemed to be a reasonable choice for a simple binary classification problem. Logistic regression uses a linear hypothesis space and works by setting a demlimiter between the datapoints.

On a 2d space one would put a line between the given datapoints. In 3d space the points would be seperated by a plane and in higher dimension a hyperplane would be used to describe the separation.

As for the loss function, I chose the logistic loss function as it is a very common and established function for logistic regresion.

1.4 Model validation

As already mention, the dataset is not very big, which is why I chose to use the k-fold cross validation method. This method is suitable for datasets with fewer entries. The algorithm works by randomly splitting up the data sets in equal parts and using one set as the test set and the rest as training set. I decided to split the data in 5 parts (k=5) to achieve a training to test ratio of 80% to 20% which is common for achieving good results.

2 Code Appendics