**Add on behind the paragraph of the first method:**

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.

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.

A picture containing diagram

Description automatically generated

Chart, line chart

Description automatically generated

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.

The dataset for the SVC method is also split into 5 parts to achieve the same 80% training to 20% validation ratio, for a fairer comparison between the two methods. We also decided to use a linear kernel for the SVC given that the nature of our data.

**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.

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | **SVC** |
| **Training Error** | 2.597054628 | 0.201995064 |
| **Training Accuracy** | 0.924808721 | 0.935684903 |
| **Test Error** | 3.133209756 | 0.221038961 |
| **Test Accuracy** | 0.909285714 | 0.916623377 |

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.