Statistical analysis and risk prediction of lung cancer exploiting logistic regression

Weiqi He

School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China

8208201415@csu.edu.cn

Abstract. Cancer is a disease with a high mortality rate. Early prediction of cancer is an important means to completely cure the disease. Therefore, there is an ever-increasing demand for technologies to detect early cancer nodules. As an easily misdiagnosed disease, the mortality rate of lung cancer has reached the highest level in recent years. Early diagnosis of lung cancer can save many lives. The study used logistic regression to predict cancer risk by collecting and analyzing patient data. Logistic regression, referred to as logistic regression analysis, is a general linear regression analysis model that relates to supervised learning in machine learning. First, n sets of data are given to the model as a training set for training, and then the model could be leveraged to classify other data (test sets), and finally the classification results could be achieved. P indicators make up each batch of data. This study used a dataset from Kaggle, which contains information from one thousand patients with lung cancer, and it has 23 eigenvalues. This dataset classifies the predicted values as high, medium, and low. The regression results show that the precision value of the low prediction value is 84%, the precision value of the middle prediction value is 88%, and the precision value of the low prediction value is 93%. The result is that as the number of samples with different predicted values increases, the precision values increase.

Keywords: lung cancer, machine learning, logistic regression, cancer prediction.

1. Introduction

The most prevalent malignant tumor worldwide is lung cancer, which poses a major threat to people's health. In most cases, the manifestations of lung cancer in the patient's body are manifested through early symptoms. The cancer's histology, stage (how far it has spread), and the patient's physical fitness determine how well the treatment works. Radiation, chemotherapy, and surgery are all used as lung cancer treatments. Overall, the proportion that a patient could survive for more than five years is merely 19.7% [1]. According to estimates, globally, in the past 2020, 1.79 million patients died of cancer worldwide, and 2.21 million people were diagnosed with cancer [2].

The second most frequent malignancy, lung cancer has the greatest fatality rates. The emergence of the COVID-19 in early 2020 has affected the statistical counts of lung cancer cases. However, data rates can be considered similar to those previously reported.

In the US, both men and women with lung cancer on average were diagnosed at the age of 70. Estimates show that 37% of occurrences impact adults over 75, whereas 53% of incidents hit people between the ages of 55 and 74 [3]. For US citizens more than 59 years old, lung cancer is the most fatal disease [4]. Exposed to environment with intense air pollution for a long time could boost the lung cancer probability, especially in those who already face a high hereditary risk [5]. Over 80% of lung cancer cases are caused by tobacco use, which is also the greatest cause of mortality that may be prevented globally [6]. These factors of risk and non-clinical symptoms are some typical cancer illness indicators. Environmental factors are significant contributors to human cancer. The risk factors mentioned above are counted and leveraged as eigenvalues, which become an important part of cancer prediction.

Pre-diagnosis can discover or limit the potential for lung cancer disease screening. In the pre-diagnosis stage, symptoms, and risk variables (such as age, tobacco use, air pollution, obesity, and insulin resistance, etc.) exhibited statistically significant effects. Lung cancer prediction has attracted discussions among many researchers.

Machine learning is facing a rapid advancement currently. By fusing the intelligence from computer science and statistics, it demonstrates a promising effect in the field of data science [7]. Through the application of optimization, mathematical, and statistical techniques, machine learning enables computers to learn from a select group of samples and eventually identify challenging patterns in huge, complicated data sets [8]. Machine learning has more recently been used for cancer prognosis and prediction.

In this experiment, lung cancer risk is estimated leveraging logistic regression (LR). It is a multivariate model that could be applied for learning the relationship between input and output. This model is widely applied for health science research for its superior interpretable design and outstanding classification capacities. There are more sophisticated versions of logistic regression known as multiclass or multinomial logistic regression that can handle more than two categories of predictors [9]. This research aims to use logistic regression to study and train the lung cancer data set, and finally find a lung cancer prediction model with high accuracy.

2. Method

2.1. Dataset

The research is based on a dataset from Kaggle [10]. The dataset focused on several risky causes and several symptoms which may indicate lung cancer. It contains 1000 groups of data. Each group contains 23 features and 1 label. Before applying machine learning models, data cleaning, checking for null values, replacing “level” with integers were achieved. Training set and test set are split up from the dataset in an 8:2 ratio. The model will be evaluated on the test set using cross-validation.

2.2. Logistic regression

It is primarily employed in classification jobs for forecasting the category of a given sample. Algorithms for classification use it. Regression is utilized because it takes the result of a linear regression function as input and applies a sigmoid function to estimate the likelihood for the given class.

2.2.1. Algorithm principle. First a binary classification problem is tackled using logistic regression. Divide the results into two categories, one labeled 1 and the other labeled 0. At this point, a function is needed that can map the input data x(i) to between 0 and 1. If the obtained result is less than 0.5, it is judged to belong to 0, otherwise it belongs to 1. Moreover, this function needs to be set as a parameter to be determined. In this way, different data sets can be used to train to obtain corresponding accurate parameters. The sigmoid function could be denoted as:

$$h(x^i) = \frac{1}{1+e^{-(w^T x^i + b)}} \tag{1}$$

2.2.2. Maximum likelihood estimation (MLE). It is a typical technique for parameter estimation. Evaluate model parameters on a given dataset. Its main idea is that the probability of an event occurring is greatest when the event occurs. Assuming there is data i, its mapping $y_i \in (0,1)$ could be regarded as a probability. The corresponding relationship is analyzed in detail. When y_i is 1, $h(x_i)$ can be regarded as a probability, which can be understood as the possibility that x_i belongs to 1. When y_i is 0, $1\text{-}h(x_i)$ can be regarded as a probability, which can be regarded as the possibility that x_i belongs to 0. Therefore, in this case, the maximum likelihood function constructed at this time is:

$$\prod_{i=1}^{k} h(x_i) \prod_{i=k+1}^{n} (1-h(x_i)) \tag{2}$$

2.2.3. Regularization. Overfitting means overfitting the training data, which increases the complexity of the model and makes it less prosperous (the ability to predict unknown data). In Figure 1, the left one is underfitting, the middle one is a suitable fit, and the right one is overfitting.

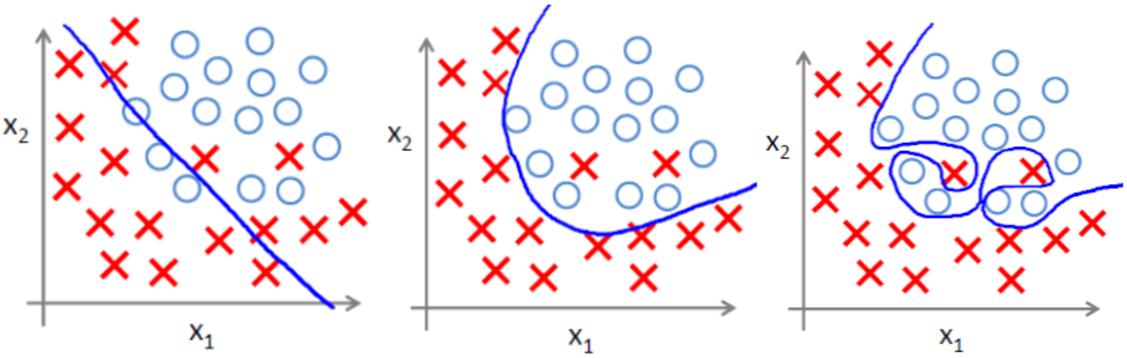

Figure 1. Three situations of fitting [11].

Regularization is a strategy that minimizes risk by adding some restrictions to the loss function, such as regularization terms or penalty terms. Overfitting is prevented by using regularization. In general, a regularization term is a function that describes the complexity of the model. This function is a monotonically increasing function. As the complexity of the model increases, the function value increases.

In regression problems, a regularization term could take different forms, where L1 and L2 norm could be used. When taking the squared loss, the loss function of the model is changed to:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \tag{3}$$

2.3. Evaluation indicators

There are four concepts. True positive (TP) means correctly categorized positive samples. False positive (FP) is the misclassified positive samples. True negative (TN) denotes the correctly classified negative samples. False negative (FN) means the wrongly classified negative samples.

2.3.1. Accuracy. It is the proportion correctly categorized samples. The correct prediction is TN+TP, and the wrong prediction is FP+FN. The equation can be redefined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

2.3.2. Precision. It is the correct positive predictions over all positive predictions. All results with a predicted value of positive may be TP or FP. It can be defined as:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

2.3.3. Recall. It is the proportion of all true positive examples that are predicted to be positive. This evaluation indicator reflects the capacity of a classifier identify all positive classes.

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

2.3.4. F1-score. It is another measurement of accuracy, which is the weighted harmonic mean of precision and recall. F1 weights both indicators. When there is a contradiction between Precision and Recall, F1-score can be used for comprehensive consideration.

$$F_1 = \frac{2\mathrm{TP}}{2TP + FP + FN} \tag{7}$$

2.3.5. ROC curve. The full name of ROC is Receiver Operating Characteristic. Each point on the curve reflects the same sensitivity. They all respond to the same signal stimulus. The difference lies in the different evaluation criteria. The vertical and horizontal axes of the curve are, respectively, the true positive rate and the false positive rate. It compares each coordinate point to the ROC curve after measuring coordinate points at various thresholds.

2.3.6. Area under curve (AUC). It is the area under the ROC curve, which value ranges between 0 and 1. AUC could be leveraged for intuitively validating the performance of the classifier. Works better with larger values.

3. Result

3.1. Statistics analysis

Figure 2 shows the number of samples with low (label 0), medium (label 1) and high (label 2) prediction results among 1000 samples.

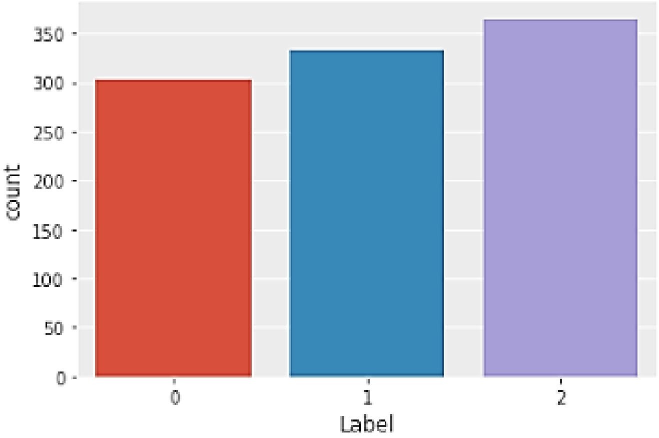

Figure 2. Label distribution.

3.2. Eigenvalues for testing

Table 1 shows the results obtained by using the logistic regression model for lung cancer prediction. In this experiment, 80% of the 1000 pieces of data which has 13 eigenvalues are used for training, while 20% are used for testing. Among the 200 samples, the number of samples with label 0, label 1, and label 2 are 54, 63, and 83, respectively.

Table 1. Results of predicting the dataset using logistic regression(c=10).

	precision	recall	f1-score
Low	0.84	0.80	0.82
Medium	0.88	0.84	0.86
High	0.93	1	0.97
average	0.89	0.88	0.88

In the detection system of lung cancer patients, it is required to detect as many cancer patients as possible, because we hope that they can be treated in time. Then at this time recall should be leveraged, that is, the recall rate, to collect all the cancer patients as much as possible. Cancer samples were identified. The recall rate from label 0 to label 2 gradually increases as the test set's sample count rises. Figure 3 is the ROC curve of the experiment, with an AUC value of 0.96.

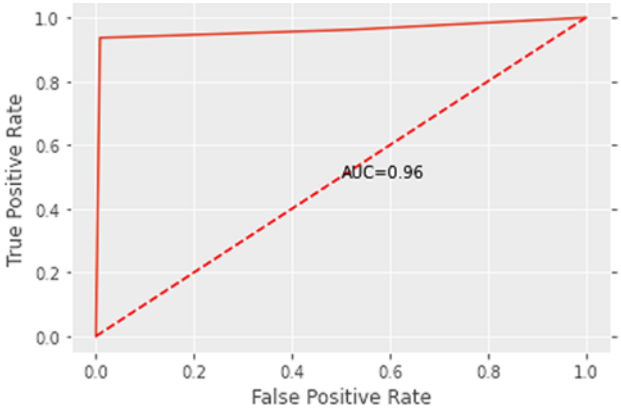

Figure 3. ROC curve of logistic regression.

The experimental results are affected by the amount of samples, the amount of eigenvalues, and model parameters. Next, the precision and recall will be further improved.

3.3. Grid search-based parameter optimization

Grid search is a way to tune parameters. Grid search works well with three or four (or fewer) hyperparameters. The computational complexity of this method grows exponentially when the number of hyperparameters increases. Grid search runs the model with each set of hyperparameters and chooses the set of hyperparameters that produces the validation set's minimum error.

C is known as a hyperparameter. Hyperparameters are used to guide the model on how to choose appropriate parameters. Parameters instruct the model how to treat the eigenvalues. Extreme parameters can lead to overfitting, so regularization can be used to constrain it.

Table 2 demonstrates the different results of modifying the value of C parameter using grid search. According to Table 2, it can be known that when C parameter is set to 10 or 100, the mean test score achieves the maximum value of 0.8538.

Table 2. Parameter optimization results using grid search.

C	mean_test_score
0.001	0.7288
0.01	0.8075
0.1	0.8313
1	0.8425
10	0.8538
100	0.8538

Animesh, et al. predictied lung cancer survivability using logistic regression algorithms, and its precision was as high as 90.5% [12]. Fang, et al. established a logistic regression model to predict lung cancer, and its precision value was as high as 97.1% [13]. These show that the precision value can be further improved. The precision value can be improved by increasing eigenvalue. In this experiment, only 13 of the 23 eigenvalues are selected in the data table. When the eigenvalues were increased to 17, the precision and recall of the model increased to over 90%. Table 3 shows the experimental data when the number of eigenvalues increases to 17.

Table 3. Results of increasing the eigenvalue to 17(c=10).

	precision	recall	f1-score
Low	0.98	0.92	0.95
Medium	0.93	0.99	0.96
High	1.00	1.00	1.00
average	0.97	0.97	0.97

When the number of eigenvalues increases to 19 and above, the precision and recall can reach 100%.

4. Conclusion

In this study, a lung cancer dataset was used to train and test a logistic regression model. When selecting 13 eigenvalues in the dataset, the experimental accuracy is 85.1%. Further increasing the eigenvalues can improve the experimental accuracy. Selecting 15 eigenvalues, the accuracy rate reaches 90.4%. Selecting 17 eigenvalues, the accuracy rate reaches 94.8%. When 19 or more eigenvalues are selected, the accuracy is close to 100%. If a higher accuracy rate is required with fewer eigenvalues, it could be achieved by increasing the number of samples. On the contrary, if it is required to get a higher accuracy rate with fewer data sets, it could be achieved by increasing the number of eigenvalues. For this experiment, the relatively optimal experimental results can also be obtained by selecting the optimal parameter C. The suitable strategy could be chosen by considering which way to optimize the most resource-saving in real life.

References

[1] Guan, Y., Ren, M., Guo, D., & He, Y. (2020). Research progress on lung cancer screening. Chinese Journal of Lung Cancer, 23(11), 954-960.

[2] Chhikara, B. S., & Parang, K. (2023). Global Cancer Statistics 2022: the trends projection analysis. Chemical Biology Letters, 10(1), 451-451.

[3] Torre, L. A., Siegel, R. L., & Jemal, A. (2016). Lung cancer statistics. Lung cancer and personalized medicine: current knowledge and therapies, 1-19.

[4] Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., et, al. (2020). Colorectal cancer statistics, 2020. CA: a cancer journal for clinicians, 70(3), 145-164.

[5] Huang, Y., Zhu, M., Ji, M., Fan, J., Xie, J., Wei, X., et, al. (2021). Air pollution, genetic factors, and the risk of lung cancer: a prospective study in the UK Biobank. American journal of respiratory and critical care medicine, 204(7), 817-825.

[6] Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. Contemporary Oncology/Współczesna Onkologia, 25(1), 45-52.

[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 59-78.

[9] Boateng, E., & Oduro, F. (2018). Predicting microfinance credit default: a study of Nsoatreman rural bank, Ghana. Journal of Advances in Mathematics and Computer Science, 26(1), 1-9.

[10] Kaggle. (2022), The Devastator Lung Cancer Prediction. Available from: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link\. cited 2023 May 1

[11] Rhys, W. (2017), Underfitting, Overfitting and Regularization. Available from: https://www.cnblogs.com/rhyswang/p/6991564.html. cited 2023 May 1

[12] Hazra, A., Bera, N., & Mandal, A. (2017). Predicting lung cancer survivability using SVM and Logistic Regression Algorithms. International Journal of Computer Applications, 174(2), 19-24.

[13] Zihan, F, Huanming, Z, Jiaming, Z. (2018). Cancer Logistic Regression Prediction Based on Cell Factors. Journal of Changsha University, 32(2), 42-44.