Reflection for Assessment 1, Week 7

Jiaqi Zhao

Introduction

In our project, we consider a binary classification problem. The project topic was determined to be human health and aiming is to use Body Signals to predict the presence or absence of an individual's drinking behavior. After searching and analysis for an appropriate database (Smoking and Drinking Dataset with body signal) on kaggle, the 4 team members selected different classification models for prediction. Finally, use performance matrix to compare 4 models and discuss result.

Consider to suitability for binary outcomes, I chose the logistic regression model after performing EDA on the data.

Literature Search and Analysis

In exploring the dataset, I conducted an EDA. Histograms were drawn to analyze the distribution of each variables like sex, age, height, weight, and various health parameters and Heatmap can visually observe the correlation between labels. This understanding formed the foundation of our model.My work included refining the dataset, feature selection, and implementing the logistic regression model.

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. In binary classification tasks, logistic regression models the probability that a given input belongs to a particular class. [Hands-on Machine Learning (http://oreilly.com/catalog/errata.csp?isbn=9781492032649)]

In logistic regression, the linear combination of input features X and their corresponding weights θ (parameters) is calculated as:

$$z = \theta_0 + \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \ldots + \theta_n \cdot X_n$$

Where: - z is the raw score, representing the weighted sum of input features plus a bias term θ_0. - $X_1, X_2, \ldots, X_n$ are the input features. - $\theta_1, \theta_2, \ldots, \theta_n$ are the corresponding weights.

The raw score z is passed through the sigmoid function(Logistic function) to squash it between 0 and 1, representing a probability. The probability that the input X belongs to the positive class is given by the sigmoid function applied to the raw score:

$$P(Y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Where: - $P(Y = 1|X)$ is the probability of the instance belonging to the positive class given the input features X.

challenges and Solutions:

Dataset

How to find the suitable data set is a challenges for our team.In addition to considering quality, quantity, accuracy,Sufficient Size and balanced data distribution, a suitable dataset should also consider connecting reality: including real data ensures that the model trained on the dataset is applicable to the actual situation.

After a group discussion and online search, We found huge size dataset about human body signals on KAGGLE collected from National Health Insurance Service in Korea. It contains 20 numerical labels and 3 categorical labels(sex, Smoking state and drinking state).This is a qualified dataset for the criteria proposed above. The dataset is related to the field of health and medical research. It includes basic health indicators that are important for medical research and highly relevant for predicting whether a person has a healthy lifestyle (not smoking, not drinking).

Model Selection

It's hard to choose a suitable and accurate model. After comparing some classification models, I choose to use logistic regression model for prediction because it is easy to understand and can effectively deal with large data sets. It is faster to train and provides fast predictions, which is highly efficient for working with large-scale data sets. Theoretically, the logistic regression model assumes a linear relationship between the characteristics and the log odds of the response variable. Logistic regression can perform well in cases where the relationship is approximately linear.

Comparative Analysis

Compare the models of the group members to determine which model fits best.

The Performance Metrics used by our team to compare the model prediction results are four different classification models: logistic regression, KNN,xgboost and decision tree. Performance metrics are crucial for measuring the accuracy of machine learning models, as they provide quantifiable and interpretable measures of the model's performance on a particular task. It can allow different models or algorithms to be compared and aid in decision-making processes.

The understanding of Performance Metrics's mathematical principles and conclusion analysis can be found in my individual section. And discussion and analysis of model comparisons are presented in the report.

Improvent and Innovations

According to the precision of the measurement model, it is found that the precision of logistic regression is not high.Therefore,I sum up the reasons for the low accuracy of the analysis and the shortcomings of the analysis model.

1.The rationale for logistic regression is the assumption that there is a linear relationship between the log odds of the independent and dependent variables. But if the true relationship between the data is nonlinear, logistic regression may not accurately capture the patterns in the data.

2.Irrelevant Features or outlier in dataset. Logistic regression is sensitive to irrelevant or redundant features. Including irrelevant features in the model can lead to overfitting and reducev accuracy. Therefore, in EDA data, it may be helpful to use heatmap to observe features between features and delete weakly correlated features.

Then, I using L1 Regularization in Logistic Regression to improve the accuracy of the model, which works according to the results of the code run and I mentioned results for code: accuracy increase from 71.46% to 71.75% and recall score increase from 0.697 to 0.703. precision increase from 72.02% to 72.17% and F_1 score also increase from 0.708283 to 0.712129.

From the results, the accuracy of the model has been improved. Because L1 Regularization adds a penalty term to the loss function, encouraging the model to use fewer features by shrinking some of the feature coefficients to zero. In standard logistic regression, the loss function for binary classification is given by:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right)$$

With L1 regularization, the loss function becomes:

$$\text{Regularized Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right) + \lambda \sum_{j=1}^{p} |\theta_j|$$

[classification notes (https://www.ole.bris.ac.uk/bbcswebdav/pid-7461256-dt-content-rid-32001453_2/xid-32001453_2)]

For group work,everyone was very motivated. In the process of discussing topics, finding data and comparative analysis, team members communicate and learn from each other.At the same time, I also learned the application of other three classification models and I can use more comprehensive data packages to conduct efficient and accurate comparative analysis of model accuracy. And with the help of team members, I became more proficient in the use of GitHub.

In the future, I will focus on the analysis of data and model results. In terms of methods to improve model accuracy, in addition to using regularization and other methods, more attention should be paid to the relationship between features of data sets.