Logistic Regression in Machine Learning

Student Id: 23097024 **Git hub link: Link**

1. Introduction

Logistic Regression (LR) is a widely known algorithm for machine learning employed to solve binary classification tasks. Logistic regression is a statistical model that gives a probability of a binary result based on one or more input variables. Though it has 'regression' in its name, logistic regression is indeed a model of classification, not a regression model. Logistic regression utilizes the logistic function (sigmoid function) in an attempt to mimic the probability that an event will occur with values between 0 and 1.
LR's readability and simplicity make it a common usage in many classification issues in a broad array of applications such as medical diagnosis, advertising, finance, and social sciences. Its ability to give probabilities along with class prediction gives it an edge in most real-world applications where decision boundaries are important.
This tutorial will include the main concepts of logistic regression, its mathematical basis, how to use it, ways to measure its performance, and its applications in real life, so that a complete picture of the algorithm is obtained.

2. How Logistic Regression Works

Logistic Regression models the relationship between a binary dependent variable and one or more independent variables by applying a logistic function. Unlike linear regression, which predicts continuous values, Logistic Regression is designed to predict probabilities, which always lie between 0 and 1. This makes it ideal for binary classification tasks, where the goal is to predict one of two possible outcomes (e.g., yes/no, success/failure).

The core idea behind Logistic Regression is that it uses the logistic function (also called the sigmoid function) to transform the linear output of the model into a probability value. The logistic function is defined as:

$$\boldsymbol{\sigma(x) = 1 + e - x1}$$

Where:

- **$\boldsymbol{\sigma}$ $(\boldsymbol{x})$ is the predicted probability of the positive class (i.e., class 1).**
- **$\boldsymbol{x}$ is the linear combination of input features, calculated $\boldsymbol{x = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn}$, where $\boldsymbol{X1, X2, \ldots, Xn}$ are the independent variables and $\boldsymbol{\beta 0, \beta 1, \ldots, \beta n}$ are the model coefficients or weights.**

The logistic function maps any real-valued number into a value between 0 and 1. This output represents the predicted probability that the input features belong to the positive class (class 1).

The predicted probability y is then used to classify the data into one of the two categories. A threshold, typically 0.5, is applied to decide which class the instance belongs to. If the probability is greater than or equal to 0.5, the instance is classified as class 1 (positive), and if it's less than 0.5, it's classified as class 0 (negative).

3. Key Concepts and Formulas

3.1 Logistic Function

The logistic function , also known as the sigmoid function , is central to logistic regression as it transforms the linear combination of the input features into a probability. The logistic function is as follows:

$$\boldsymbol{\sigma(x) = \frac{1}{1 + e - x}}$$

Where:

- $\boldsymbol{\sigma(x)}$ is the predicted probability that the event happens (i.e., the probability of class 1).

- $\boldsymbol{e^{-x}}$ is the exponential function of the linear combination of the input features.

The logistic function maps values between 0 and 1, making it well-suited to model probabilities. The function guarantees the output of the model to be the probability of an event taking place, transforming real-valued predictions onto a probability scale.

3.2 Log-Loss (Binary Cross-Entropy Loss)

Logistic regression minimizes a loss function known as log-loss or binary cross-entropy , which measures how close the predicted probabilities are to the actual labels. The formula for log-loss is:

$$\boldsymbol{Log - Loss = -\frac{1}{n}\sum_{i=1}^{n}[yi\ log(yi\hat{}) + (1 - yi)\ log\ (1 - yi\hat{})]}$$

Where:

- $\boldsymbol{yi}$ is the true label (0 or 1) of the I - th sample.

- $\boldsymbol{yi\hat{}}$ is the predicted probability for the $i - th$ sample.

Log-loss penalizes incorrect predictions, and the goal is to minimize this loss during model training. It calculates how dissimilar the predicted probability is from the true label, and it guides optimization algorithms like gradient descent.

3.3 Decision Boundary

The decision boundary is the threshold for making predictions. In logistic regression, the decision boundary is typically at a predicted probability of 0.5. The decision rule is:

- **If** $\boldsymbol{y\hat{} \geq 0.5}$, predict the positive class (1).

- **If** $\boldsymbol{y\hat{} < 0.5}$, predict the negative class (0).

The decision boundary separates the two classes based on the predicted probabilities. You can move the threshold to manage the precision-recall trade-off in classification problems**.**

4. Implementation of Code

This example trains a Logistic Regression model on binary classification (Malignant vs. Benign tumours) using the Breast Cancer dataset. We first standardize the dataset using Standard Scaler() for better model convergence. We split the data into training (70%) and testing (30%) sets using train test split() with class balance preserved using stratify=y.

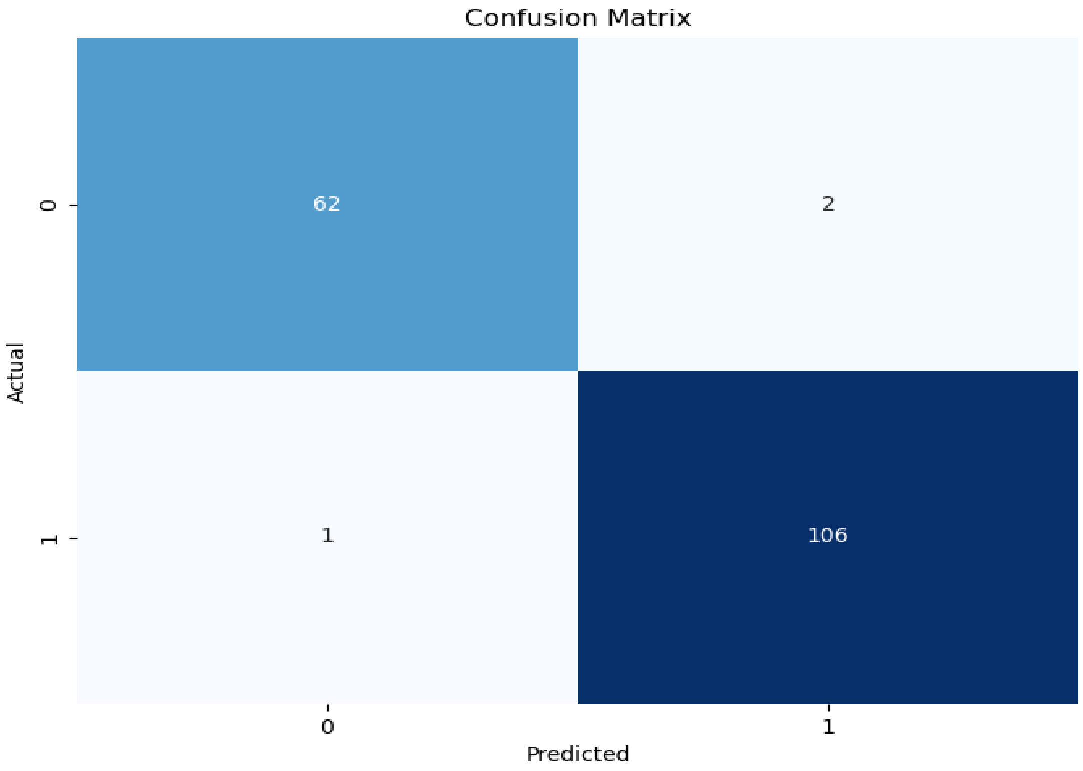
A Logistic Regression model is fitted with solver='saga', max iter=5000 to prevent convergence issues. The model is used to predict on the test set. Its accuracy, confusion matrix, classification report are computed, and 5-fold cross-validation is performed using Stratified K Fold() to see its performance.

A confusion matrix is plotted using Seaborn's heatmap() showing correct and false classifications. An ROC curve is also plotted to check the capability of the model in distinguishing between classes, and the AUC score is used to measure its performance.

The model is highly accurate, demonstrating that Logistic Regression is an appropriate method for medical diagnosis issues. Using cross val score(), we check its consistency across different data splits. This approach ensures an exhaustive evaluation of the model's classification ability.

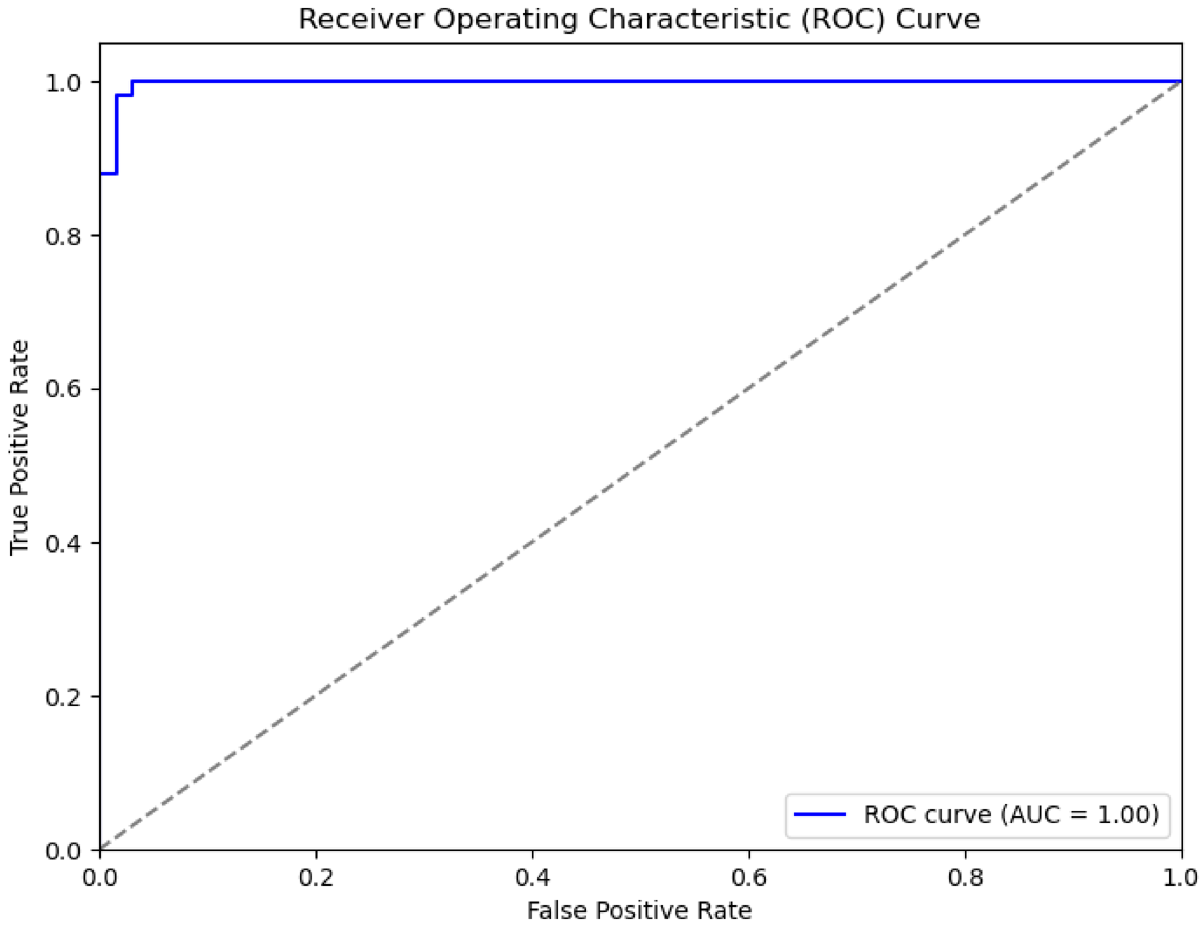
5. Model Evaluation and Performance

The logistic regression model for breast cancer classification performed very well, achieving 98 percent accuracy on the test set. This means that the model correctly classified almost all the instances in the dataset.



Confusion matrix analysis indicates that the model accurately predicted 106 malignant cases and 62 benign cases. It misclassified two benign cases as malignant, potentially resulting in unnecessary medical interventions. More seriously, it misclassified one malignant case as benign, which is a worse mistake in medical diagnosis.

The classification report also confirms the model's performance. Precision, or the ratio of predicted positive cases that were indeed positive, is 0.98 for both classes. Recall, or the ratio of actual cases identified correctly, is 0.99 for malignant cases. The F1-score, a trade-off between precision and recall, is near perfect at 0.98 to 0.99.



Stratified five-fold cross-validation confirmed that the model generalizes well across different data splits. ROC curve analysis yielded an AUC score of around 0.99, confirming that the model can distinguish well between malignant and benign cases.

Although the model is effective, the tuning of decision thresholds or the application of ensemble methods would further reduce false negatives, resulting in even higher reliability when applied to real-world scenarios.

6. Advantages & Cons, and Comparison with Other ML Algorithms

Advantages of Logistic Regression:

- Easy and Quick : Logistic regression is cheap computationally and simple to apply, and therefore appropriate for big data and when results are needed urgently. It works fine even when there are many features.

- Interpretable : Logistic regression is arguably the most interpretable machine learning model. The model coefficients are directly proportional to each feature's effect on the probability of the outcome, and it is easy to visualize how individual features influence predictions. This is particularly valuable if model action is to be interpreted or explanations of decisions to be made in fields like medicine or finance.

- Probabilistic Output : Unlike simple binary classifiers, logistic regression provides probabilities for all classes, which gives more information regarding the model's belief in its prediction. Probabilistic output is helpful in decision-making when the cost of false positives and false negatives varies.

Disadvantages of Logistic Regression

- Assumes Linearity : Logistic regression is an assumption of linear relation among the input features and the log-odds of the outcome, which gives it less ability to map more complex, non-linear relations in the data.

- Sensitive to Outliers : The logistic regression is sensitive to outliers in the data. Abnormal values can bias the model coefficients and decrease predictive performance, especially when working with small datasets.

Comparison with Other Algorithms

- Logistic regression is simpler to interpret and understand compared to more advanced algorithms like Support Vector Machines (SVM) or Random Forests . It may perform worse at times on non-linear data.

- Algorithms like Decision Trees or Neural Networks may be applied to data with non-linear decision boundaries since they tend to perform better. These models possess the ability to identify more complex relationships between the target variable and features.

---

7. Applications

Logistic Regression finds many applications in different domains, mostly for binary classification issues, where out of two outcomes, one needs to be predicted. Some of the most common applications are:

- **Medical Diagnostics :** Logistic regression is often used in medicine for the prediction of whether a patient has a certain disease or condition. For example, it can be used to predict whether a patient has cancer, diabetes, or heart disease based on various diagnostic attributes like age, symptoms, results of tests, and medical history.

- **Spam Email Detection :** Logistic regression is used in email programs to classify an email as spam or not spam . An email content, sender's email, and so on, features can be trained using the model to predict whether the email is unsolicited or genuine.

- **Predicting Customer Churn :** Firms use logistic regression to predict whether a customer will cancel a subscription or discontinue service. This helps firms take action in advance to retain customers, based on utilization trends, customer call volume, and contract terms.

Such applications illustrate the application of logistic regression to decision-making in numerous sectors by providing accurate binary predictions.

---

8. How to Improve Accuracy

- **Feature Engineering :** the most effective model enhancement method is likely to be engineering better features. This can include creating new features that find useful patterns in the data or modifying existing ones, e.g., scaling, normalization, encoding categorical variables. Good features allow the model to capture more accurately the underlying structure of the data.

- **Regularization :** L1 (Lasso) and L2 (Ridge) regularization methods are used to prevent overfitting by including a penalty term for large coefficients. Lasso does feature selection by setting some of the coefficients to zero, while Ridge discourages large coefficients but does not set them to zero. Both methods help to improve the model's capacity to generalize to new observations.

- **Non-linear Features :** Logistic regression assumes a linear relationship between features and the log-odds of the target. You can fit non-linear relationships by applying feature transformations (e.g., polynomial features or interactions), which may improve model accuracy in certain cases.

- **Cross-Validation :** k-fold cross-validation is utilized to determine the best model and hyperparameters by testing the performance of the model on different subsets of data. It provides a more accurate estimate of model accuracy and avoids overfitting, ensuring that the model works well on new data.

9. Conclusion

Logistic Regression is a general and widely used machine learning algorithm, particularly for binary classification tasks. It is remarkable in its simplicity, interpretability, and efficacy, and is a favourite in real-world usage like medical diagnosis, spam filtering in emails, and predicting customer churn. As it offers class prediction probabilities, logistic regression offers valuable insights into model confidence and decision-making. However, its linear hypothesis between the features and the target variable can be limiting. Where data is in non-linear relationships, advanced algorithms like decision trees, support vector machines, or neural networks might be more effective than logistic regression. Overall, logistic regression remains a simple machine learning tool since it is easy to apply, simple to implement, and efficient. It is particularly strong in situations where a linear boundary between decisions is sufficient, and therefore it is an easy choice for most real-world classification problems.

10. References

1. **Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). Wiley-Interscience.**
2. **Cox, D. R. (1958). *The Regression Analysis of Binary Variables*. Journal of the Royal Statistical Society.**
3. **Scikit-learn Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html**
4. **Data Camp Logistic Regression Tutorial: https://www.datacamp.com/community/tutorials/logistic-regression-R**