Wednesday: Logistic Regression

Agenda

5 min: Overview

50 min: Suggested Resources

1 hr 5 min: Exercise

Specific Learning Outcomes

- I can recognize when to choose logistic regression.
- I can distinguish the use of odds, odds ratios, and transformations in logistic regression.
- I can correctly interpret the results of logistic regression.
- I can select the best logistic model that describes the relationship under question.
- I can demonstrate how logistic regression can be extended for nominal and ordinal outcomes.
- I can explain how fitting a logistic regression differs from other regression models.

Overview

Logistic regression is another family of commonly used regression algorithms. In a lot of ways, linear regression and logistic regression are similar. But, the biggest difference lies in what they are used for. Linear regression algorithms are used to predict/forecast values while logistic regression algorithms are used for classification tasks i.e. 1 / 0, Yes / No, True / False given a set of independent variables. Some practical examples of classification tasks include classifying whether an email is a spam or not, classifying whether a tumour is malignant or benign, classifying whether a website is fraudulent or not, etc.

Logistic regression is applicable, if:

- We want to model the probabilities of a response variable as a function of some explanatory variables, e.g. "success" of admission as a function of gender.
- We want to perform descriptive discriminate analyses such as describing the differences between individuals in separate groups as a function of explanatory variables, e.g. student admitted and rejected as a function of gender.

- We want to predict probabilities that individuals fall into two categories of the binary response as a function of some explanatory variables, e.g. what is the probability that a student is admitted given she is a female.
- We want to classify individuals into two categories based on explanatory variables, e.g. classify new students into "admitted" or "rejected" group depending on their gender.

Logistic regression can be classified in the following ways:

1. **Binomial Logistic Regression:** target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
2. **Multinomial Logistic Regression:** target variable can have 3 or more possible types which are not ordered(i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.
3. **Ordinal Logistic Regression:** it deals with target variables with ordered categories. For example, a test score can be categorized as:“very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.

In order to improve the accuracy of our logistic regression model we can perform any of the following techniques:

1. **Explore more classifiers** - Logistic Regression learns a linear decision surface that separates our classes. It could be possible that our 2 classes may not be linearly separable. In such a case we might need to look at other classifiers such as Support Vector Machines which are able to learn more complex decision boundaries. We can also start looking at Tree-Based classifiers such as Decision Trees which can learn rules from our data. We can think of them as a series of If-Else rules which the algorithm automatically learns from the data.
2. **Optimize other scores** - We can optimize on other metrics also such as Log Loss and F1-Score. The F1-Score could be useful, in case of class imbalance. This is a good guide that talks more about scoring.
3. **Class Imbalance** - We can look for class imbalance in our data. Since we are working with admit/reject data, then the number of rejects would be significantly higher than the admits. Most classifiers in SkLearn including LogisticRegression have a class_weight parameter. Setting that to balanced might also work well in case of a class imbalance.
4. **Feature Scaling and/or Normalization** - We can check the scales of our gre and gpa features. They differ on 2 orders of magnitude. Therefore, our gre feature will end up dominating the others in a classifier like Logistic Regression. We can normalize all our features to the same scale before putting them in a machine learning model.
5. **Hyperparameter Tuning - Grid Search / Random Search** - We can improve our accuracy by performing a Grid Search to tune the hyperparameters of our model. As we will get to learn, for example in case of LogisticRegression, the parameter C is a hyperparameter. Also, we should avoid using the test data during grid search. Instead perform cross-validation. Use our test data only to report the final numbers for our final model. We note that GridSearch should be done for all models that we try because then only we will be able to tell what is the best we can get from each model. Scikit-Learn provides the GridSearchCV class for this.
6. **Error Analysis** - For each of our models, go back and look at the cases where they are failing. We might end up finding that some of our models work well on one part of the parameter space while others work better on other parts. If this is the case, then Ensemble Techniques such as VotingClassifier

techniques often give the best results. Models that win Kaggle competitions are many times ensemble models.

7. **More Features** - If all of the above fail, we can start looking for more features.

Let's begin by going through the following resources, then work on the provided exercise below;

Suggested Resources

- Introduction to Logistic Regression. **[Link]** **(https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#introduction)**
- Logistic Regression for Machine Learning. **[Link]** **(https://machinelearningmastery.com/logistic-regression-for-machine-learning/)**
- Explaining Logistic Regression Results to Non-Statistical Audiences. **[Link]** **(https://www.theanalysisfactor.com/explaining-logistic-regression/)**
- Building and Applying Logistic Regression Models. **[Link]** **(https://personal.utdallas.edu/~pkc022000/6390/SP06/NOTES/Logistic_Regression_4.pdf)** **(https://towardsdatascience.com/a-comprehensive-study-of-linear-vs-logistic-regression-to-refresh-the-basics-7e526c1d3ebe)**

Exercise

- Python Programming: Logistic Regression [**Link** **(https://colab.research.google.com/drive/1fERHbnlUiDd_z48aO8kv7u5MNQyl1uCx?usp=sharing)**]

"Given enough time any business can discover the truth, however, Data Scientists offer a faster solution." Damian Mingle.