C++_Implementation

4375.004
Chris Talley

Running Code

➜ ./logreg.out Titanic_LOG.csv ","
Reading Dataset
Splitting Dataset
Creating LR model from Dataset
Optimized Coefficients:
W: 0.259625
B: 0.109785
Train Accuracy: 60.9785
Test Accuracy: 52.1531
Elapsed Time in milliseconds: 4903 ms

Written Assignment

This is the result of my Logistic Regression implementation in C++. The model resulted in somewhat optimized coefficients. It took a little under five seconds and was lacking in accuracy. This can be explained by using inly one predictor.

Generative classifiers, such as naive bayes, are created using joint probability distribution. This means that the features and taret variable are used together to determine the classification. This is done in Naive Bayes by turning the joint probability into a conditional P(Y|X). As a result, generative models learn the distribution of data, and can be used to predict new values.

On the other hand, discriminative classifiers work differently. For example, in logistic regression, probability is used to draw a boundary condition between classes. If a data point belongs to class 1, then LR uses P(Y=1|X). This focuses the solution to determine all boundary thresholds within the data being classified. As a result, the data distribution is not learned, so predictions are not as capable.

According to Randal LeVeque of the University of Washington, computational science is facing a credibility crisis. Most computational results presented at conferences and in papers are impossible to verify. Part of the problem is the diverse set of tools. Researchers in disparate fields have their own set of community tools and systems. Due

to this, sharing original data sets along with scholarly papers proves to be a challenge. One way to reduce this barrier is by creating better tools.

Tools that keep track of data and allow for more open collaboration will enable researchers to better cross-reference. This is supporte in a paper by researchers at Princeton. They claim that Independent datasets created by third parties can provide a more standardized environment. This process of working on common bases of data can be beneficial in fostering a more communicative research system for the purposes of validating data science.