DS 310 Final Project
Author: Kyle Dennis

Overview

In this project, my goal is to fit a logistic regression model from scratch, using Python. Instead of using existing libraries such as scikit-learn for model fitting, I will build my algorithms from the ground up. This work's purpose is to help understand the mathematics behind logistic regression model fitting.

The main algorithm I will implement is the Newton-Raphson algorithm. This method, at the fundamental level, iteratively estimates the roots of a function. In this application, it will be used to calculate logistic regression parameters that best fit a dataset. My chosen dataset contains data on credit card transactions and tracks whether they were fraudulent or not. Once the model is fitted, it can be used to classify whether a given credit card transaction is fraudulent or not.

Mathematical Background

A likelihood function measures how well a set of parameters describes a dataset(i.e. the probability of the dataset given the parameters). A common goal of machine learning optimization is to maximize the likelihood function of a dataset. In other words, the goal is to find parameter values that best describe or predict a dataset. Here, the Newton-Raphson method will be used to find logistic regression model parameter values that maximize the log-likelihood function of the dataset(we take the log to avoid underflow). To apply the Newton-

Raphson method, we complete the following steps: We approximate the log-likelihood as a quadratic function using the Taylor Series. We then take the first derivative of this function with respect to the parameters(the parameters are represented as a vector) and set it equal to zero(a vector of zeros). We are left with a linear system we can solve for, where the solution is the step size of our parameter guesses(Heath). The Newton-Raphson method repeats this process until convergence(until we maximize the log-likelihood function).

To obtain the linear system, we need to calculate the gradient vector and the Hessian matrix of the log-likelihood function. The gradient vector contains the first-ordered partial derivatives with respect to each parameter. This represents the ith first derivative in the truncated Taylor series expansion. The Hessian matrix contains the second-ordered partial derivatives with respect to the parameters. This represents the second derivative in the truncated Taylor series expansion(Heath). I will implement two main algorithms for this- one to calculate the gradient and one to calculate the hessian for each iteration of the Newton-Raphson algorithm.

Once the best weights(parameters) are found, we can use them for classification. To do so, we can model the logit(the log-odds ratio) as a linear combination of parameters with predictor variables. We then plug in this linear combination into the sigmoid function to obtain the probability of success(Agresti). Then, we make a threshold for classification. Finally, we have a classification model. This model can classify fraudulent transactions, when provided input(when provided a transaction). For the purposes of this project however, I will focus on model fitting rather than classification examples.

Logistic regression at its core predicts the probability of a binary outcome. Of course, it has assumptions about the dataset we will be using. One important assumption is that the observations in the dataset are independent of each other. Another is that the observations are Bernoulli trials, which means they are random experiments with only two outcomes. In addition, it assumes the relationship between the logit transformation and the predictor variables is linear.

Dataset

The dataset I will use contains credit card transactions made in September 2013 by European cardholders. The data has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection. The sample size of this dataset is 284,807 transactions.

For model fitting, I will be using the independent variables “V1” and “V2” to predict the dependent variable “Class.” The V1 and V2 variables are principal components resulting from a principal components analysis. The original variables and their contexts are anonymous to maintain confidentiality,(Machine Learning Group 2013). The class variable is binary: 1 represents a fraudulent transaction and 0 represents a non-fraudulent transaction.

Reflection

This project was very productive for me. It gave me the chance to understand the core mathematical principles that underly logistic regression. Also, I learned the cost of working with

big data. If I were to use more predictor variables and observations during model fitting, the time cost would grow exponentially. The problem of big data in the context of machine learning is very interesting to me, and I will continue to explore it in the future.

Citations

Dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Other:

- Heath, MT. 2005. *Scientific Computing: An introductory survey* (2nd ed.). McGraw-Hill.
- Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, Incorporated.