MLE and Logistic Regression

(https://github.com/learn-co-curriculum/dsc-mle-logistic-regression) (https://github.com/learn-co-curriculum/dsc-mle-logistic-regression/issues/new/choose)

Introduction

In this lesson, you'll further investigate the connections between maximum likelihood estimation and logistic regression. This is a common perspective for logistic regression and will be the underlying intuition for upcoming lessons where you'll code the algorithm from the ground up using NumPy.

Objectives

You will be able to:

- Determine how MLE is tied into logistic regression

MLE formulation

As discussed, maximum likelihood estimation finds the underlying parameters of an assumed distribution to maximize the likelihood of the observations. Logistic regression expands upon this by investigating the conditional probabilities associated with the various features, treating them as independent probabilities and calculating the respective total probability.

For example, when predicting an individual's risk for heart disease, you might consider various factors such as their family history, weight, diet, exercise routines, blood pressure, and cholesterol. When looked at individually, each of these has an associated conditional probability that the individual has heart disease based on each of these factors. Mathematically, you can write each of these probabilities for each factor X as:

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

This is the standard linear regression model ($\beta_0 + \beta_1 x_i$) you have seen previously, modified to have a range of 0 to 1. The range is modified and constrained by applying the sigmoid function since you're predicting probabilities.

Then, combining these conditional probabilities from multiple features, you maximize the likelihood function of each of those independent conditional probabilities, giving you:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i} = \prod_{i=1}^{N} \frac{\exp y_i(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Notes on mathematical symbols

Recall that the $\prod$ sign stands for a product of each of these individual probabilities. (Similar to how $\sum$ stands for the sum of a series.) Since this is a monotonically increasing function, its maximum will be the same as the logarithm of the function, which is typically used in practice in order to decompose this product of probabilities into a sum of log probabilities for easier calculation of the derivative. In future sections, you'll investigate the derivative of this function and then use that in order to code up our own function for logistic regression.

Algorithm bias and ethical concerns

It should also be noted that while this is mathematically sound and a powerful tool, the model will simply reflect the data that is fed in. For example, logistic regression and other algorithms are used to inform a wide range of decisions including whether to provide someone with a loan, the degree of criminal sentencing, or whether to hire an individual for a job. In all of these scenarios, it is again important to remember that the algorithm is simply reflective of the underlying data itself. If an algorithm is trained on a dataset where African Americans have had disproportionate criminal prosecution, the algorithm will continue to perpetuate these racial injustices. Similarly, algorithms trained on data that reflect a gender pay-gap will also continue to promote this bias unless adequately accounted for through careful preprocessing and normalization. With this, substantial thought and analysis regarding problem set up and the resulting model is incredibly important. While future lessons and labs in this section return to underlying mathematical theory and how to implement logistic regression on your own, it is worthwhile to investigate some of the current problems regarding some of these algorithms, and how naive implementations can perpetuate unjust biases.

Additional resources

Below are a handful of resources providing further information regarding some of the topics discussed here. Be sure to check out some of the news articles describing how poor safeguards and problem formulation surrounding algorithms such as logistic regression can lead to unjust biases:

Algorithm bias and ethical concerns

- **Machine Bias (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)**
- **Amazon’s Gender-Biased Algorithm Is Not Alone (https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone)**

- The software that runs our lives can be bigoted and unfair. But we can fix it (https://www.bostonglobe.com/business/2017/12/21/the-software-that-runs-our-lives-can-bigoted-and-unfair-but-can-fix/RK4xG4gYxcVNVTIubeC1JI/story.html)
- Why artificial intelligence is far too human (https://www.bostonglobe.com/ideas/2017/07/07/why-artificial-intelligence-far-too-human/jvG77QR5xPbpwBL2ApAFAN/story.html)
- Can Computers Be Racist? The Human-Like Bias Of Algorithms (https://www.npr.org/2016/03/14/470427605/can-computers-be-racist-the-human-like-bias-of-algorithms)

Additional mathematical resources

If you want to really go down the math rabbit-hole, check out section 4.4 on Logistic Regression from the Elements of Statistical Learning which can be found here: https://web.stanford.edu/~hastie/ElemStatLearn// (https://web.stanford.edu/%7Ehastie/ElemStatLearn//) .

Summary

In this lesson, you further analyzed logistic regression from the perspective of maximum likelihood estimation. Additionally, there was a brief pause to consider the setup and interpretation of algorithms such as logistic regression. In particular, remember that issues regarding racial and gender bias that can be perpetuated by these algorithms. Always try to ensure your models are ethically sound. In the proceeding labs and lessons, you will continue to formalize your knowledge of logistic regression, implementing gradient descent and then a full logistic regression algorithm using Python packages in order to give you a deeper understanding of how logistic regression works.

How do you feel about this lesson?

Have specific feedback?

Tell us here! (https://github.com/learn-co-curriculum/dsc-mle-logistic-regression/issues/new/choose)