A Study of Sigmoid-based Multi-class Logistic Regression

Saniya Abushakimova
MATH 540 Statistical Learning, Fall 2021
Nazarbayev University, Nur-Sultan, Kazakhstan

Abstract

The aim of this research project is to investigate how sigmoid-based multi-class logistic regression would perform compared to softmax-based one. The results showed that both models produce the same accuracy rate, however, the softmax-based model was computationally faster.

Introduction

Multi-class logistic regression is an extension of logistic regression that is used to predict the probabilities of more than two classes. Logistic regression by default is limited only to binary classification problems, and one of the possible ways to generalize it to multi-class problems is to use the one-vs-rest technique. It involves splitting the multi-class classification problem into multiple binary classification problems, however, this approach might be computationally expensive. A more efficient way to solve multi-class classification problems would be to use an extension of logistic regression that involves changing the activation function from sigmoid to softmax. More precisely, let us consider any input $\mathbf{x}_i \in R^d$, then softmax normalization for K classes would be:

$$\boxed{\text{softmax}(\mathbf{w}_k^T \mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i}}, \quad k \in 1, .., K} \tag{1}$$

If we feed (1) into the cross-entropy loss, we will get:

$$\begin{aligned}\ell(\mathbf{w}_k^T \mathbf{x}_i) &= -\sum_{k=1}^K \mathbf{e}_y \ln \left(\frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i}} \right) \\ &= \ln \left(\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i} \right) - \mathbf{e}_y (\mathbf{w}_k^T \mathbf{x}_i),\end{aligned}$$

where $\mathbf{e}_y$ is the one-hot representation of y, i.e. it is a vector with all entries being zero, except y-th component, which is one.

The gradient of the loss with respect to $\mathbf{w}_k$ is then:

$$\begin{aligned}\nabla_{\mathbf{w}_k} \ell &= \nabla_{\mathbf{w}_k} \ln \left(\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i} \right) - \nabla_{\mathbf{w}_k} \mathbf{e}_y (\mathbf{w}_k^T \mathbf{x}_i) \\ &= \frac{1}{\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i}} \left(\nabla_{\mathbf{w}_k} \sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i} \right) - \mathbf{e}_y \mathbf{x}_i \\ &= \frac{e^{\mathbf{w}_k^T \mathbf{x}_i} \mathbf{x}_i}{\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i}} - \mathbf{e}_y \mathbf{x}_i = \left(\frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{j=1}^K e^{\mathbf{w}_j^T \mathbf{x}_i}} - \mathbf{e}_y \right) \mathbf{x}_i \\ &= (\text{softmax}(\mathbf{w}_k^T \mathbf{x}_i) - \mathbf{e}_y) \mathbf{x}_i\end{aligned}$$

From the obtained gradient we can conclude that gradient-based optimization will result in $\text{softmax}(\mathbf{w}_k^T \mathbf{x}_i) \rightarrow \mathbf{e}_y$. However, can we use element-wise sigmoid instead of softmax in multi-label classification. The following research project is going to address this question by conducting both theoretical analysis and empirical evaluation.

Theoretical Analysis

The sigmoid function is given by:

$$\boxed{\sigma(\mathbf{w}_k^T \mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{1 + e^{\mathbf{w}_k^T \mathbf{x}_i}}, \quad k \in 1, .., K} \tag{2}$$

In order to investigate whether sigmoid-based multi-class logistic regression will be able to converge, one must show that gradient-based optimization will force $\sigma(\mathbf{w}_k^T \mathbf{x}_i) \rightarrow \mathbf{e}_y$. Let us first derive a new cross-entropy loss.

$$\begin{aligned}\ell(\mathbf{w}_k^T \mathbf{x}_i) &= -\sum_{k=1}^K \mathbf{e}_y \ln \left(\frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{1 + e^{\mathbf{w}_k^T \mathbf{x}_i}} \right) \\ &= \sum_{k=1}^K \mathbf{e}_y \ln \left(1 + e^{\mathbf{w}_k^T \mathbf{x}_i} \right) - \sum_{k=1}^K \mathbf{e}_y \ln (e^{\mathbf{w}_k^T \mathbf{x}_i}) \\ &= \ln \left(1 + e^{\mathbf{w}_k^T \mathbf{x}_i} \right) \sum_{k=1}^K \mathbf{e}_y - \sum_{k=1}^K \mathbf{e}_y (\mathbf{w}_k^T \mathbf{x}_i) \\ &= \ln \left(1 + e^{\mathbf{w}_k^T \mathbf{x}_i} \right) - \mathbf{e}_y (\mathbf{w}_k^T \mathbf{x}_i)\end{aligned}$$

Then, the gradient of the new cross-entropy loss will be:

$$\begin{aligned}\nabla_{\mathbf{w}_k}\ell &= \nabla_{\mathbf{w}_k}\ln\left(1+e^{\mathbf{w}_k^T\mathbf{x}_i}\right) - \nabla_{\mathbf{w}_k}\mathbf{e}_y(\mathbf{w}_k^T\mathbf{x}_i)\\ &= \frac{1}{1+e^{\mathbf{w}_k^T\mathbf{x}_i}}\left(\nabla_{\mathbf{w}_k}(1+e^{\mathbf{w}_k^T\mathbf{x}_i})\right) - \mathbf{e}_y\mathbf{x}_i\\ &= \frac{e^{\mathbf{w}_k^T\mathbf{x}_i}\mathbf{x}_i}{1+e^{\mathbf{w}_k^T\mathbf{x}_i}} - \mathbf{e}_y\mathbf{x}_i = \left(\frac{e^{\mathbf{w}_k^T\mathbf{x}_i}}{1+e^{\mathbf{w}_k^T\mathbf{x}_i}} - \mathbf{e}_y\right)\mathbf{x}_i\\ &= (\sigma(\mathbf{w}_k^T\mathbf{x}_i) - \mathbf{e}_y)\mathbf{x}_i\end{aligned}$$

As it can be noted from the gradient, sigmoid function will approach $\mathbf{e}_y$, hence, our hypothesis was theoretically justified. Now, we can proceed to the empirical evaluation of this hypothesis.

Empirical Evaluation

Dataset

The dataset used in empirical evaluation describes hazelnuts. It contains 10 different attributes, namely length, width, thickness, surface area, mass, compactness, hardness, shell top radius, water content, carbohydrate content. They were used to predict hazelnut varieties, which could be either c-avellana, c-americana, or c-cornuta. The data across all hazelnut types is approximately evenly distributed. As a part of preprocessing, the data was also normalized and classes were one-hot encoded.

Results

Baseline multi-class logistic regression was trained on 141 instances of data with gradient descent under 75 epochs. The resulting accuracy rate was 0.95. Modified multi-class logistic regression with the sigmoid function was trained under the same conditions, and the resulting accuracy rate was also 0.95. Since the latter model converged and achieved the baseline result, we can conclude that our hypothesis was proven empirically. A complete solution can be accessed through the following github repository.

Discussion

The initial assumption was that it will take more epochs for the sigmoid-based multi-class logistic regression to start converging because the sum of individual probabilities across all K classes does not add up to one due to the specificity of the sigmoid function. Interestingly, during training, the sigmoid-based model converged faster than the softmax-based one (See Fig. 1). This is probably because the data was normalized before being passed to the model. Another assumption was that the softmax-based model would probably take more time to train because of the summation term in the denominator. However, the training process for the softmax-based model took 0.034 seconds, while for the sigmoid-based 1.08 seconds. One of the possible explanations for this phenomenon could be a small size of the dataset. Presumably, for a larger dataset (eg. 100K instances), the softmax-based model will perform slower due to the computational cost of the sum.

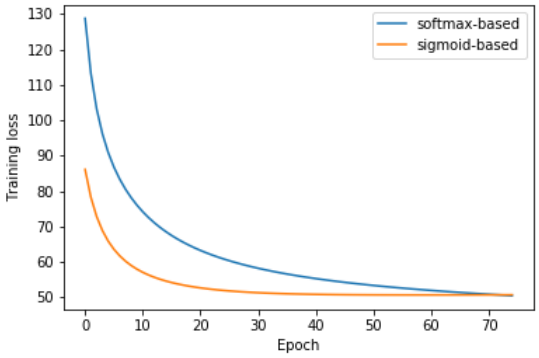

Figure 1: Training loss vs epoch

Conclusion

This research project aimed to investigate whether it is possible to replace softmax with sigmoid function in multi-class logistic regression. After careful theoretical and empirical analysis, it was concluded that for a relatively small dataset sigmoid-based model converges and achieves the baseline result. A further investigation of the sigmoid-based multi-class model on a larger dataset is recommended.