

Binary Logistic Regression

Aaron Avram

June 4 2025

Introduction

In this write up I will go through my derivation of the objective function and the optimization algorithm for my Multinomial Logistic Regression model.

Construction

For my notation let $X \in \mathbb{R}^{n \times p}$ be the matrix with each of the n training inputs as rows. Where we define p to be the dimension of each input vector (or equivalently the number of features). Next let $y \in \{1, \dots, K\}^n$ equal the vector of training labels, where there are K output classes. I will apply the shortcut used in the Binary Model Construction where each row of X is given an additional final entry with a 1, to account for the bias term. So X is now a $n \times (p + 1)$ dimensional matrix. The model is parametrized by $K - 1$ $p + 1$ dimensional vectors $\theta^{(1)}, \dots, \theta^{(K-1)}$, which I will collectively refer to by the flattened combined vector θ . Where the probabilities are given by:

$$\Pr(\kappa|x; \theta^{(\kappa)}) = \begin{cases} \frac{\exp((\theta^{(\kappa)})^T x)}{1 + \sum_{1 \leq i < K} \exp((\theta^{(i)})^T x)} & \kappa < K \\ \frac{1}{1 + \sum_{1 \leq i < K} \exp((\theta^{(i)})^T x)} & \kappa = K \end{cases}$$

The log likelihood function is then given by:

$$\mathcal{L}(\theta) = \sum_{1 \leq r \leq n} \log \Pr(y_r | x_r; \theta^{y_r})$$

Using a one hot encoding with the kronecker delta function we can write this as:

$$= \sum_{1 \leq r \leq n} \sum_{1 \leq s \leq K} \delta_{y_r s} \log \Pr(y_r | x_r; \theta^{(s)})$$

Consider the Likelihood function with respect to one input vector x_r and one parameter $\theta^{(j)}$:

$$\mathcal{L}_{rj}(\theta) = \log \Pr(y_r | x_r)$$

However we can condense this further, as if $y_r \neq j$ the numerator of the probability function is irrelevant to $\theta^{(j)}$, and if we decompose the fraction via the logarithm rules we can omit the numerator

term and replace it with a kronecker delta. I.e Suppose $\Pr = N/D$ then $\log(\Pr) = \log(N) - \log(D)$ and so if N is independent of $\theta(j)$ we can omit it. Thus:

$$\begin{aligned}\mathcal{L}_{rj}(\theta) &= \delta_{y_rj} \log(\exp((\theta^{(j)})^T x_r)) - \log(1 + \sum_{1 \leq i < K} \exp((\theta^{(i)})^T x)) \\ &= \delta_{y_rj} (\theta^{(j)})^T x_r - \log(1 + \sum_{1 \leq i < K} \exp((\theta^{(i)})^T x))\end{aligned}$$

Taking the partial derivative with respect to $\theta^{(j)}$ we find

$$\begin{aligned}\frac{\partial \mathcal{L}_{rj}}{\partial \theta^{(j)}} &= \delta_{y_rj} x_r - \frac{x_r \exp((\theta^{(j)})^T x_r)}{1 + \sum_{1 \leq i < K} \exp((\theta^{(i)})^T x)} \\ &= x_r (\delta_{y_rj} - \Pr(j|x_r))\end{aligned}$$

Now let us take the second partial derivative of this expression with respect to $\theta^{(i)}$

$$\frac{\partial}{\partial \theta^{(i)}} \frac{\partial \mathcal{L}_{rj}}{\partial \theta^{(j)}} = \begin{cases} -x_r x_r^T \Pr(j|x_r) (1 - \Pr(j|x_r)) & i = j \\ x_r x_r^T \Pr(j|x_r) \Pr(i|x_r) & i \neq j \end{cases}$$

We can make this more compact utilizing the kronecker delta function:

$$\frac{\partial}{\partial \theta^{(i)}} \frac{\partial \mathcal{L}_{rj}}{\partial \theta^{(j)}} = x_r x_r^T \Pr(i|x_r) (\delta_{ij} - \Pr(j|x_r))$$

So the total second derivative of \mathcal{L} is:

$$\begin{aligned}\frac{\partial}{\partial \theta^{(i)}} \frac{\partial \mathcal{L}}{\partial \theta^{(j)}} &= \sum_{1 \leq r \leq n} x_r x_r^T \Pr(i|x_r) (\delta_{ij} - \Pr(j|x_r)) \\ &= \sum_{1 \leq r \leq n} x_r x_r^T \Pr(i|x_r) (\delta_{ij} - \Pr(j|x_r))\end{aligned}$$