Softmax algorithm

Enzhi Li

September 6, 2018

I Introduction

Softmax is a natural generalization of logistic regression algorithm which is generally used for binary classification. In contrast to logistic regression, softmax is designed for multiple classification. Given a data set of $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{0, 1, ..., N-1\}$, our goal is to train a model that can predict the label of an unknown feature vector $\boldsymbol{x}$. Softmax gives us one such model. In this algorithm, we assume that for a certain data point $(\boldsymbol{x}, y)$, the probability of the observing label y is

$$\begin{aligned} p(y=0) &= \frac{1}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}}, \\ p(y=k) &= \frac{e^{\theta_k^T x}}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}}, k = 1, 2, ..., N-1. \end{aligned} \tag{1}$$

Here, $x = (\boldsymbol{x}, 1)^T$ is an extended column vector that represents the feature vector, and $\theta = (\boldsymbol{w}, b)^T$ is a column vector that represents the model parameters. We can rewrite the above probability as

$$\begin{aligned} p &= \left(\frac{1}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}}\right)^{\delta_{y,0}} \prod_{k=1}^{N-1} \left(\frac{e^{\theta_k^T x}}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}}\right)^{\delta_{y,k}} \\ &= \left(\frac{1}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}}\right)^{\sum_{j=0}^{N-1} \delta_{y,j}} \prod_{k=1}^{N-1} e^{\theta_k^T x \delta_{y,k}} \\ &= \frac{1}{1+\sum_{i=1}^{N-1} e^{\theta_i^T x}} \prod_{k=1}^{N-1} e^{\theta_k^T x \delta_{y,k}} \end{aligned} \tag{2}$$

Using the maximum likelihood estimation (MLE), we can obtain the total probability of observing the data set $\{(\boldsymbol{x}_l, y_l), l = 1, 2, ..., m\}$ as

$$\begin{aligned} P &= \prod_{l=1}^{m} p_l \\ &= \prod_{l=1}^{m} \left(\frac{1}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l}} \right)^{\delta_{y_l,0}} \prod_{j=1}^{N-1} \left(\frac{e^{\theta_j^T x_l}}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l}} \right)^{\delta_{y_l,j}} \\ &= \prod_{l=1}^{m} \frac{1}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l}} \prod_{j=1}^{N-1} e^{\theta_j^T x_l \delta_{y_l,j}} \end{aligned} \tag{3}$$

With the MLE, we are to find the parameters $\theta_i, i = 1, 2, ..., N-1$ that can maximize the total probability P. This problem can be reformulated as minimizing the loss function which is defined as

$$\begin{aligned} L &= -\log P \\ &= -\sum_{l=1}^{m} \left(\delta_{y_l,0} \log \frac{1}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l}} + \sum_{j=1}^{N-1} \delta_{y_l,j} \log \frac{e^{\theta_j^T x_l}}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l}} \right) \\ &= \sum_{l=1}^{m} \left(\sum_{j=0}^{N-1} \delta_{y_l,j} \log \left(1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l} \right) - \sum_{j=1}^{N-1} \delta_{y_l,j} \theta_j^T x_l \right) \\ &= \sum_{l=1}^{m} \left(\log \left(1 + \sum_{i=1}^{N-1} e^{\theta_i^T x_l} \right) - \sum_{j=1}^{N-1} \delta_{y_l,j} \theta_j^T x_l \right) \end{aligned} \tag{4}$$

We need to calculate the gradient of this function to find the optimal parameters that can minimize this loss function. In the next section, we are going to calculate its gradient.

II Gradient and Hessian matrix

The loss function defined in the previous section is a summation of independent functions. The summand function takes the form

$$\begin{aligned} f(\theta; x, y) &= \sum_{j=0}^{N-1} \delta_{y,j} \log \left(1 + \sum_{i=1}^{N-1} e^{\theta_i^T x} \right) - \sum_{j=1}^{N-1} \delta_{y,j} \theta_j^T x \\ &= \log \left(1 + \sum_{i=1}^{N-1} e^{\theta_i^T x} \right) - \sum_{j=1}^{N-1} \delta_{y,j} \theta_j^T x \end{aligned} \tag{5}$$

This is a function with $\theta = (\theta_1, \theta_2, ..., \theta_{N-1})$ as variables and x, y as parameters. The gradient of this function with respect to $\theta_k, k = 1, 2, ..., N-1$ is

$$\nabla_{\theta_k} f(\theta) = \left(\frac{e^{\theta_k^T x}}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x}} - \delta_{y,k} \right) x \tag{6}$$

Hessian matrix of this function is

$$\nabla^2_{\theta_{k'}\theta_k} f(\theta) = e^{\theta_k^T x} \frac{1}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x}} \left(\delta_{kk'} - \frac{e^{\theta_{k'}^T x}}{1 + \sum_{i=1}^{N-1} e^{\theta_i^T x}} \right) x x^T \tag{7}$$