Logistic Regression and Cross Entropy

Michael Hauser

The goal of this short article is to quickly review logistic regression, softmax and cross-entropy so that one has a working knowledge of these tools.

1 Problem Setting

Given a dataset $D := \{(x_n, y_n)\}_{n=1,2,\dots,N}$ of N-many samples of input-output pairs (x_n, y_n), where $x_n \in \mathbb{R}^d$ is the input to be classified, and $y_n \in [0,1]^K$ is the output label, the goal is to find a probability distribution that assigns to each x_n the correct probability label $y_n = [0, \dots, 0, 1, 0, \dots, 0]$, where the 1 is at the k^{th} index.

We can think of this as a regression problem, in the sense of finding a map $h_\Theta : \mathbb{R}^d \to [0,1]^K$ defined by $h_\Theta : x \mapsto h_\Theta(x)$ such that $\sum_{k=1}^K h_\Theta(x)_k = 1$, where Θ are the model parameters over some general type of function h. We can then *interpret* $h_\Theta(x)_k$ as being the probability that input x belongs to class k, i.e. $P_\Theta(y = k|x) := h_\Theta(x)_k$. With this notation, $h_\Theta(x)$ is the histogram of probabilities of x, while $h_\Theta(x)_k$ is the k^{th} region of the histogram, i.e. the probability that x is of class k (and of course we want the histogram of probabilities to sum to 1).

2 Cross Entropy

The Kullback-Leibler divergence is a way of measuring the distance between two probability distributions. It is not a proper metric in the sense of metric spaces, but it tells us the number of bits required to construct one distribution from another, so one can think of the two distributions as being a certain number of bits apart. Suppose for input-output pair (x_n, y_n) the learned distribution is $P_\Theta(y|x_n) := h_\Theta(x_n)$ while the true distribution is $Q(y|x_n) = y_n$ (which is usually a vector with a 1 at the k^{th} element and zeros everywhere else), then the Kullback-Leibler divergence is defined:

$$KL(Q||P_\Theta) := \sum_{k=1}^K Q(y = k|x_n) \log \frac{Q(y = k|x_n)}{P_\Theta(y = k|x_n)} \tag{1}$$

We can expand this into two terms, namely in terms of the entropy $H(Q) := -\sum_{k=1}^K Q(y = k|x_n) \log Q(y = k|x_n)$ as well as the cross entropy:

$$XE\left(Q, P_\Theta\right) := -\sum_{k=1}^{K} Q\left(y = k|x_n\right) \log P_\Theta\left(y = k|x_n\right) \tag{2}$$

In this form, minimizing $KL\left(Q||P_\Theta\right) = -H\left(Q\right) + XE\left(Q, P_\Theta\right)$ with respect to the parameters Θ through, say, some gradient descent search, only the cross entropy term is dependent on Θ and so that is the only term we need to consider in the optimiziation.

It is important to note that $XE\left(Q, P_\Theta\right)$ is *linear* in $\log P_\Theta\left(y = k|x_n\right)$, which is important to keep in mind for our discussion on the logistic regression.

3 Logistic Regression

Suppose we have some model regression function $h_\Theta : \mathbb{R}^d \to [0,1]^K$, defined by $h_\Theta : x \mapsto h_\Theta\left(x\right)$, interpreted as being the probability that input x belongs to class k, i.e. $P_\Theta\left(y = k|x\right) := h_\Theta\left(x\right)_k$.

Logistic regression then assumes that log probabilities are themselves linear, i.e. $\log P_\Theta\left(y_n = k|x_n\right) = W_k \cdot x_n + C$, where $W_k \in \mathbb{R}^K$ and $C \in \mathbb{R}$ is some normalization constant. To find C, we have $\sum_{k=1}^{K} P_\Theta\left(y_n = k|x_n\right) = e^C \sum_{k=1}^{K} e^{W_k \cdot x_n} = 1$, so we define the partition function as follows:

$$Z := \sum_{k=1}^{K} e^{W_k \cdot x_n} = e^{-C} \tag{3}$$

With this new (re-)definition of the normalization constant, we then define the log-probability:

$$\log P_\Theta\left(y_n = k|x_n\right) = W_k \cdot x_n - \log Z \tag{4}$$

This is written for specifically the probability of class k. To see the probability of a given x belonging to class $y = k$, we have the following:

$$P_\Theta\left(y = k|x\right) = \frac{1}{Z} e^{W_k \cdot x} \tag{5}$$

This is the softmax classifier. If instead we are interested in the entire histogram of probabilities we can write:

$$\log P_\Theta\left(y|x\right) = W \cdot x - \log Z \tag{6}$$

where $W \in \mathbb{R}^{K \times K}$ and the y-argument is left open denoting that it is over all classes.

We then have the final form of the cross entropy:

$$XE\left(Q, P_\Theta\right) := -\sum_{n=1}^{N} \sum_{k=1}^{K} Q\left(y = k|x_n\right) \left[W \cdot x_n - \log Z\right] \tag{7}$$