Why Sums are Bad

Jason D. M. Rennie
jrennie@alum.mit.edu

November 4, 2004

Abstract

Sums aren't really bad, but they have a tendency to create non-convexities. For example, learning parameters for Logistic Regression is convex, but learning parameters for a mixture of Logistic Regression models is not. In fact, it is not a single sum of logarithms that causes problems, but two. Every model includes at least an implicit sum or integral hidden inside the normalization constant. When the unnormalized model also includes a sum, there is the potential for it to be the difference between two convex functions, which is not generally convex.

Consider the problem of binary classification. We have a set of data points, $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$, and a set of labels, $\vec{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{+1, -1\}$. We would like to learn weights for a Logistic Regression model,

$$P(\vec{y}|X) = \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(\sum_j w_j f_j(x_{ij}, y_i)\right), \tag{1}$$

where the $\{w_j\}$ are the model parameters and the $\{f_j\}$ are feature functions. The normalization constant is $Z_i = \exp\left(\sum_j w_j f_j(x_{ij}, 1)\right) + \exp\left(\sum_j w_j f_j(x_{ij}, -1)\right)$. For optimization purposes, this model is better written in log-form:

$$\log P(\vec{y}|X) = \sum_i \sum_j w_j f_j(x_{ij}, y_i) - \sum_i \log Z_i. \tag{2}$$

The first term is a linear function of the weights. Linear functions are both convex and concave. The sum of two convex functions is convex; the sum of two concave functions is concave. So, if $\log Z_i$ is a convex function of the weights, then the entire model is a concave function of the weights (due to the minus sign). Consider a simplified model with a single weight and feature function. $\log Z_i$ is convex for the simplified model if

$$\frac{\log\left(e^{w_1 c_1} + e^{w_1 c_2}\right) + \log\left(e^{w_2 c_1} + e^{w_2 c_2}\right)}{2} \geq \log\left(e^{\frac{c_1}{2}(w_1+w_2)} + e^{\frac{c_2}{2}(w_1+w_2)}\right) \tag{3}$$

A little bit of reversible manipulation leaves us with:

$$\frac{1}{2}\left(e^{w_1c_1+w_2c_2}+e^{w_1c_2+w_2c_1}\right)\geq e^{\frac{1}{2}(w_1c_1+w_2c_2+w_1c_2+w_2c_1)}, \tag{4}$$

which is true due to the convexity of exponentiation. This generalizes to any number of weights/features. Hence, $\log Z_i$ is convex and the entire model is concave.

Now, we consider a mixture of Logistic Regression models,

$$\log P(\vec{y}|X)=\sum_i \log\left(e^{\sum_j w_j f_j(x_{ij},y_i)}+e^{\sum_j v_j f_j(x_{ij},y_i)}\right)-\sum_i \log Z_i. \tag{5}$$

Immediately we see that this will generally not be concave. As we have seen, the log of a sum of exponentials is convex. A sum of convex functions is convex, so the first term is convex. But, $\log Z_i$ is, at best, convex, thus giving us the difference between two convex terms, which is generally not convex. One sum of logarithms is ominous, but two are real trouble.