Kernelized Softmax

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

January 15, 2005

Abstract

We derive the kernelized version of Softmax (multiclass Logistic Regression).

Let $X \in \mathbb{R}^{n \times d}$, be a set of examples. Let $\vec{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{1, \ldots, l\}$, be a set of corresponding labels. We use A_i to denote the i^{th} row of A. Regularized Softmax learns parameters $W \in \mathbb{R}^{l \times d}$ so as to minimize

$$-\log P(\vec{y}|X,W) = \sum_{i=1}^{n} \log \left(\sum_{u=1}^{l} \exp(W_u X_i^T) \right) - \sum_{i=1}^{n} W_{y_i} X_i^T + \|W\|_{\text{Fro}}^2, \quad (1)$$

where $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm. The Representer Theorem (see Appendix B of (Rifkin, 2002)) gives us that we can rewrite the weight matrix as a weighted sum of example vectors. Let $C \in \mathbb{R}^{l \times n}$, $K = XX^T$, and $W = CX$. Hence,

$$-\log P(\vec{y}|X,W) = \sum_{i=1}^{n} \log \left(\sum_{u=1}^{l} \exp(K_i C_u^T) \right) - \sum_{i=1}^{n} K_i C_{y_i}^T + \|CX\|_{\text{Fro}}^2. \quad (2)$$

Define $Z_i = \sum_{u=1}^{l} \exp(K_i C_u^T)$ and $P_{iu} = \exp(K_i C_u^T)/Z_i$. The partial derivatives are

$$-\frac{\partial \log P(\vec{y}|X,W)}{\partial C_{uj}} = \sum_{i=1}^{n} K_{ij} P_{iu} - \sum_{i|y_i=u} K_{ij} + \lambda K C^T. \quad (3)$$

References

Rifkin, R. (2002). *Everything old is new again: A fresh look at historical approaches in machine learning*. Doctoral dissertation, Massachusetts Institute of Technology.