Maximum-Margin Logistic Regression

Jason D. M. Rennie
jrennie@gmail.com

February 10, 2005

The Regularized Logistic Regression (RLR) minimization objective is

$$J_{\mathrm{RLR}} = \sum_{i=1}^{n} \log(1 + \exp(-y_i \cdot \vec{x}_i^T \vec{w})) + \frac{\lambda}{2} \vec{w}^T \vec{w}, \tag{1}$$

where $\{\vec{x}_1, \ldots, \vec{x}_n\}$ are the training examples and $\{y_1, \ldots, y_n\}$ are the labels. The per-example (Logistic) loss is $g(z) = \log(1 + \exp(-z))$.

We make two modifications to RLR that improve its ability to generalize: (1) we use a generalized form the Logistic, and (2) we shift the Logistic by one. These two modifications have the effect of encouraging a margin yet ignoring examples that are predicted well by the model. This modified Logistic Regression, which we will call Maximum-Margin Logistic Regression (MMLR), can be viewed as an approximation to the Support Vector Machine.

Zhang and Oles discuss the Generalized Logistic[1] loss [1],

$$g(z, \gamma) = \frac{1}{\gamma} \log(1 + \exp(-\gamma z)). \tag{2}$$

γ is what we call the "sharpness." Define the sharpness of a function $f(x)$ as the maximum magnitude of the second derivative,

$$\mathrm{sharpness}(f) = \max_z \left| \frac{\partial^2 f(z)}{\partial z \partial z} \right|. \tag{3}$$

Define the closure $f(z) = g(z, \gamma)$. Then $\mathrm{sharpness}(f) = \frac{\gamma}{4}$.

At $\gamma = 1$, the Generalized Logistic loss is the Logistic loss; it is a re-scaled Logisitic for other values. Figure 1 shows graphs of the Logistic and Generalized Logistic. The Generalized Logistic is a smooth approximation of the Hinge loss. As $\gamma \to \infty$, sharpness increases without bound and the Generalized Logistic approaches the Hinge loss, $h(z) = \max(0, -z)$. Smaller values of γ yield increasingly smooth approximations of the hinge loss.

[1] In fact, Zhang and Oles discuss the shifted Generalized Logistic. We introduce the unshifted version here, then discuss the shifted version later.

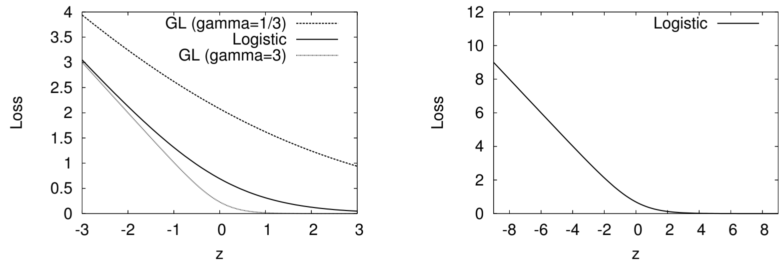

Figure 1: The left graphic shows the Logistic loss (middle) and the Generalized Logistic (GL) loss for $\gamma = 1/3$ (top) and $\gamma = 3$ (bottom). The right figure shows the Logistic loss, but the axes have been scaled by a factor of 3. Note the similarity of the scaled Logistic to the Generalized Logistic ($\gamma = 3$).

For our Maximum-Margin Logistic Regression, we use a shifted version of the Generalized Logistic Loss. We subtract one from z so that the "hinge" occurs at $z = 1$.

$$g_+(z, \gamma) = \frac{1}{\gamma} \log(1 + \exp(\gamma(1 - z))). \tag{4}$$

We use a large value of γ (e.g. $\gamma = 10$) so that our loss function approximates the Hinge loss. The minimization objective for MMLR is

$$J_{\text{MMLR}} = \frac{1}{\gamma} \sum_{i=1}^{n} \log(1 + \exp(\gamma(1 - z_i))) + \frac{\lambda}{2} \vec{w}^T \vec{w}, \tag{5}$$

where $z_i = y_i \cdot \vec{x}_i^T \vec{w}$. Optimization of the parameters can be done efficiently with first-order gradient descent-type techniques. Note that $\frac{\partial g_+(z,\gamma)}{\partial z} = -\frac{\exp(\gamma(1-z))}{1+\exp(\gamma(1-z))}$. The gradient of the objective is

$$\frac{\partial J_{\text{MMLR}}}{\partial w_j} = -\sum_{i=1}^{n} \frac{\exp(\gamma(1 - z_i))}{1 + \exp(\gamma(1 - z_i))} y_i x_{ij} + \lambda w_j. \tag{6}$$

Note that this model could be used as part of an iterative method for learning SVM parameters. Each round, γ is increased according to a pre-set schedule.

References

[1] T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4:5–31, 2001.