Ordinal Logistic Regression

Jason D. M. Rennie
jrennie@gmail.com

February 16, 2005

1 Introduction

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 $\{\vec{x}_1, \dots, \vec{x}_n\}$, $x_i \in \mathbb{R}^d$, are the training examples and $\{y_1, \dots, y_n\}$, $y_i \in \{+1, -1\}$, are the labels. We wish to extend this to multiple, ordinal labels. In other words, we still want a single weight vector, but we want to segment the real line into l sections, one for each label. We use $l-1$ thresholds, $\{\theta_1, \dots, \theta_{l-1}\}$ to represent the segments. We use θ_0 and θ_l to denote $-\infty$ and $+\infty$ (respectively). Label $k \in \{1, \dots, l\}$ corresponds to the segment (θ_{k-1}, θ_k).

Shashua and Levin introduced the idea of applying large-margin classifiers to the problem of ordinal classification (also known as "ranking" or "rating") [2]. We discuss their "fixed-margin" formulation; we call it "immediate-threshold." We also discuss a formulation, introduced by Srebro et. al., where loss is incurred for all thresholds, not only the neighboring ones [3]; we call this "all-threshold." Whereas earlier works used the Hinge loss, we use the Logistic loss here. We also discuss the Generalized Logistic loss, which provides a continuum between the Logistic and the Hinge.

2 Immediate-Threshold

Define $h(z) := \log(1+\exp(z))$. Then the minimization objective for the immediate-threshold version of Ordinal Logistic Regression is

$$J_{\mathrm{Imm}} = \sum_{i=1}^{n} h(\theta_{y_i-1} - \vec{x}_i^T \vec{w}) + h(\vec{x}_i^T \vec{w} - \theta_{y_i}) + \frac{\lambda}{2} \vec{w}^T \vec{w}. \tag{2}$$

Note that $h(\theta_0 - \vec{x}_i^T \vec{w}) = h(\vec{x}_i^T \vec{w} - \theta_l) = 0 \; \forall i, \vec{w}$. Also, note that we have defined h so that the thresholds appear are oriented as they are on the real number line with respect to outputs of correctly classified examples. For example, $\vec{x}_i^T \vec{w} < \theta_1$ if $y_i = 1$, and x_i is correctly classified.

We can use gradient descent-type methods to learn this model. For that, we need to be able to calculate the gradient. Note that $\frac{\partial h(z)}{\partial z} = \exp(z)/(1+\exp(z))$. Define $g(z) := (1 + \exp(-z))^{-1} = \frac{\partial h(z)}{\partial z}$. Then, the partial derivative wrt each weight is

$$\frac{\partial J_{\text{Imm}}}{\partial w_j} = \sum_{i=1}^{n} x_{ij} \left[g(\vec{x}_i^T \vec{w} - \theta_{y_i}) - g(\theta_{y_i - 1} - \vec{x}_i^T \vec{w}) \right] + \lambda w_j, \tag{3}$$

or, written more compactly using matrix notation,

$$\frac{\partial J_{\text{Imm}}}{\partial \vec{w}} = X^T \left[g(X\vec{w} - \theta_{\vec{y}}) - g(\theta_{\vec{y}-1} - X\vec{w}) \right] + \lambda \vec{w}, \tag{4}$$

where $\theta_{\vec{y}} = \{\theta_{y_1}, \ldots, \theta_{y_n}\}$ and functions are applied element-wise. The partitial derivative wrt each threshold is

$$\frac{\partial J_{\text{Imm}}}{\partial \theta_k} = \sum_{i|y_i - 1 = k} g(\theta_k - \vec{x}_i^T \vec{w}) - \sum_{i|y_i = k} g(\vec{x}_i^T \vec{w} - \theta_k). \tag{5}$$

3 All-Threshold

Note that with the above Immediate-Threshold formulation, there is no guarantee that the thresholds will be ordered. In the case of one or more under-represented labels, one can almost be assured that for the optimal parameter setting, there will be some $i < j$ such that $\theta_i > \theta_j$. The formulation we describe next, All-Threshold, imposes additional penalties which ensure that the thresholds are ordered, $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{l-1}$.

We define $h(z)$ and $g(z)$ as before. Then the minimization objective for the all-threshold version of Ordinal Logistic Regression is

$$J_{\text{All}} = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i - 1} h(\theta_k - \vec{x}_i^T \vec{w}) + \sum_{k=y_i}^{l-1} h(\vec{x}_i^T \vec{w} - \theta_k) \right] + \frac{\lambda}{2} \vec{w}^T \vec{w}. \tag{6}$$

The partial derivative wrt each weight is

$$\frac{\partial J_{\text{All}}}{\partial w_j} = \sum_{i=1}^{n} \left[\sum_{k=y_i}^{l-1} x_{ij} g(\vec{x}_i^T \vec{w} - \theta_k) - \sum_{k=1}^{y_i - 1} x_{ij} g(\theta_k - \vec{x}_i^T \vec{w}) \right] + \lambda w_j. \tag{7}$$

We can also write this compactly using matrix notation. Define $\vec{s}(k)$ such that $s_i(k) = \begin{cases} +1 & \text{if } k \geq y_i \\ -1 & \text{if } k < y_i \end{cases}$. Then,

$$\frac{\partial J_{\text{All}}}{\partial \vec{w}} = \sum_{k=1}^{l-1} X^T [\vec{s}(k) * g(\vec{s}(k) * (X\vec{w} - \theta_k))] + \lambda \vec{w}, \tag{8}$$

where $*$ denote element-wise multiplication. Using our definition for $\vec{s}(k)$, the partial derivative wrt each threshold is

$$\frac{\partial J_{\text{All}}}{\partial \theta_k} = -\vec{1}^T[\vec{s}(k) * g(\vec{s}(k) * (X\vec{w} - \theta_k))] \tag{9}$$

4 Generalized Logistic Loss

So far we have used the Logistic Loss, $h(z) = \log(1 + \exp(z))$. Zhang and Oles (§2, page 6) discuss the Generalized Logistic loss[1],

$$h_+(z) = \frac{1}{\gamma}\log(1 + \exp(\gamma(z-1))), \tag{10}$$

which effectively scales the x- and y-axes according to γ [4]. See Rennie for additional discussion [1]. An important property of the Generalized Logistic is that its limit as $\gamma \to \infty$ is the (reflected) Hinge loss. We are concerned with only the relative values of $h_+(z)$, so we can discard the outside multiplicative constant. Additionally, since we learn (via 10-fold cross-validation, or some such techique) the regularization parameter, λ (which effectively controls the magnitude of z), we can dismiss the multiplication of z by a constant. We are left with

$$h^*(z) = \log(1 + \exp(z - \gamma)), \tag{11}$$

which we call the Shifted Logistic loss. As $\gamma \to \infty$, the parameters learned using this loss will tend to the parameters learned using the Hinge loss[2]. A benefit of this over the Generalized Logistic loss is that it will tend to be more stable for gradient descent-type algorithms that use function curvature estimates for line search. As discussed in [1], sharpness(h_+) $= \gamma/4$; yet sharpness(h^*) $= 1/4$. So, we think the Shifted Logistic loss will be easier to use with gradient descent-type optimization algorithms.

Updating Immediate-threshold and All-threshold Ordinal Logistic Regression to use the Shifted Logistic loss requires only small changes. However, care must be taken to ensure that signs are correct—γ should always have the same sign as the threshold.

4.1 Immediate-Threshold

The updated objective is

$$J^*_{\text{Imm}} = \sum_{i=1}^{n} h(\theta_{y_i-1} + \gamma - \vec{x}_i^T\vec{w}) + h(\vec{x}_i^T\vec{w} - \theta_{y_i} - \gamma) + \frac{\lambda}{2}\vec{w}^T\vec{w}. \tag{12}$$

[1] Traditionally, the exponent is negated in the Logistic and Generalized Logistic; we break from tradition. But, notice that the difference is only surface-deep. The important properties of the Logistic loss do not change; $h(z)$ is simply the reflection about the y-axis of the traditionally defined Logistic loss.

[2] Assuming that the regularization parameter is learned in a way that is not tied to the magnitude of the parameters, e.g. minimization of cross-validation error on the training data.

The partial derivatives follow analogously.

4.2 All-Threshold

The updated objective is

$$J^*_{\text{All}} = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i - 1} h(\theta_k + \gamma - \vec{x}_i^T \vec{w}) + \sum_{k=y_i}^{l-1} h(\vec{x}_i^T \vec{w} - \theta_k - \gamma) \right] + \frac{\lambda}{2} \vec{w}^T \vec{w}. \quad (13)$$

The partial derivatives follow analogously.

References

[1] J. D. M. Rennie. Maximum-margin logistic regression. http://people.csail.mit.edu/∼jrennie/writing, February 2005.

[2] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems 15*, 2003.

[3] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

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