Ordinal Smooth Hinge Classification

Jason D. M. Rennie
jrennie@gmail.com

February 22, 2005

1 Introduction

The Smooth Hinge Classification (SHC) minimization objective is

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

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 and $h(\cdot)$ is the smooth hinge loss function:

$$h(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \le 0 \\ \frac{1}{2}(1-z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \ge 1 \end{cases}. \tag{2}$$

Note that the derivative is zero to the right of the margin, one to the left of the margin and linearly interpolates between the two values within the margin:

$$h'(z) = \begin{cases} -1 & \text{if } z \le 0 \\ z-1 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \ge 1 \end{cases}. \tag{3}$$

See [2] for a discussion of the Smooth Hinge Loss function. We wish to extend this to multiple, ordinal labels, as we did for Logistic Regression in [1]. As before, we use $l-1$ thresholds, $\{\theta_1, \dots, \theta_{l-1}\}$ to represent the segments. We concern ourselves with the "all-threshold" objective for ordinal regression/classification and find no need to define "thresholds" at ±∞ as we did before.

2 All-Threshold

The All-Threshold Ordinal Smooth Hinge Classification (AOSHC) minimization objective is

$$J_{\mathrm{All}} = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i - 1} h(\vec{x}_i^T \vec{w} - \theta_k) + \sum_{k=y_i}^{l-1} h(\theta_k - \vec{x}_i^T \vec{w}) \right] + \frac{\lambda}{2} \vec{w}^T \vec{w}. \tag{4}$$

The partial derivative wrt each weight is

$$\frac{\partial J_{\text{All}}}{\partial w_j} = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i - 1} x_{ij} h'(\vec{x}_i^T \vec{w} - \theta_k) - \sum_{k=y_i}^{l-1} x_{ij} h'(\theta_k - \vec{x}_i^T \vec{w}) \right] + \lambda w_j. \tag{5}$$

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,

$$\frac{\partial J_{\text{All}}}{\partial \vec{w}} = \lambda \vec{w} - \sum_{k=1}^{l-1} X^T [\vec{s}(k) * h'(\vec{s}(k) * (\theta_k - X\vec{w}))], \tag{6}$$

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

$$\frac{\partial J_{\text{All}}}{\partial \theta_k} = \vec{1}^T [\vec{s}(k) * h'(\vec{s}(k) * (\theta_k - X\vec{w}))] \tag{7}$$

References

[1] J. D. M. Rennie. Ordinal logistic regression. http://people.csail.mit.edu/~jrennie/writing, February 2005.

[2] J. D. M. Rennie. Smooth hinge classification. http://people.csail.mit.edu/~jrennie/writing, February 2005.