Smooth Hinge Classification

Jason D. M. Rennie
jrennie@gmail.com

February 22, 2005

1 Smooth Hinge Loss

In earlier writing [2, 1], we discussed alternate loss functions that might be used for classification. We continue our discussion here by introducing yet another loss function, the Smooth Hinge. Recall that the (Shifted) Hinge loss function is defined as

$$\text{Hinge}(z) = \max(0, 1 - z). \tag{1}$$

In our eyes, there are two key properties of the Hinge. The first is that it is zero for values greater than 1. Thus, a classification model using this loss does not incur gain for pushing examples far from the decision boundary; model loss for an example decreases until the example is a pre-set "margin" distance from the boundary. The second important property is that the loss is linear for outliers. This encourages the model to balance the *number* of outliers per class; a model with a linear outlier loss treats all outliers equally (distance away from the boundary is unimportant). This is unlike the least squares model, which balances the distances of outliers—a single negative class outlier 1,000 units across the boundary balances 1,000 positive class outliers each one unit across the boundary. The Hinge loss also encourages sparsity in the dual variables due to the sharp corner at one. We find this property less compelling; if anything, we would prefer sparsity in the primal variables (feature selection).

A difficulty with the Hinge loss is that direct optimization is, at best, difficult, due to the discontinuity in the derivative at $z = 1$. We propose to use a smooth version of the Hinge. This smooth verison shares important similarities to the Hinge. The "Smooth Hinge" is zero for $z \geq 1$ and has constant (negative) slope for $z \leq 0$. For $1 > z > 0$, it smoothly transitions between a zero slope and a constant negative slope. We define the Smooth Hinge as

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

Note that the function is smooth—the derivative is continuous:

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

Clearly this is not the only smooth verison of the Hinge loss that is possible. However, it is a canonical one that has the important properties we discussed; it is also sufficiently smooth that classification models using the Smooth Hinge can be solved using classic gradient-descent-type algorithms.

2 Classification Objective and Gradient

Here we give the classification and objective for binary classification. Let $X \in \mathbb{R}^{n \times d}$ be the training examples, one per row. Let $\vec{y} \in \{+1, -1\}^n$ be the corresponding labels. Then, Smooth Hinge Classification (SHC) finds a weight vector, $\vec{w} \in \mathbb{R}^d$, to minimize

$$J_{\text{SHC}} = \vec{1}^T h(\vec{y} * X^T \vec{w}) + \frac{\lambda}{2} \vec{w}^T \vec{w}, \tag{4}$$

where $*$ is element-wise product. The gradient is

$$\frac{\partial J_{\text{SHC}}}{\partial \vec{w}} = X^T (\vec{y} * h'(\vec{y} * X^T \vec{w})) + \lambda \vec{w}. \tag{5}$$

Since the objective is convex, any local minimum we find is the global minimum. Also, since the gradient is continuous, we expect that traditional gradient-descent type optimization algorithms will be able to efficiently optimize the objective.

3 The Generalized Smooth Hinge

As we mentioned earlier, the Smooth Hinge is one of many possible smooth verison of the Hinge. Here we detail a family of smoothed Hinge loss functions which includes the Smooth Hinge discussed above. One desirable property of the Hinge is that it encourages a margin of exactly one. This is a result of a constant derivative > 0 for $z < 1$. We can more closely approximate this property by using a more quickly increasing function in the $0 < z < 1$ interval:

$$h_\alpha(z) = \begin{cases} \frac{\alpha-1}{\alpha} - x & \text{if } z \leq 0 \\ \frac{1}{\alpha+1} x^{\alpha+1} - x + \frac{\alpha-1}{\alpha} & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1 \end{cases}. \tag{6}$$

$$h'_\alpha(z) = \begin{cases} -1 & \text{if } z \leq 0 \\ x^\alpha - 1 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1 \end{cases}. \tag{7}$$

Figure 1 shows the Hinge, the Smooth Hinge and the Generalized Smooth Hinge ($\alpha = 3$) Losses.

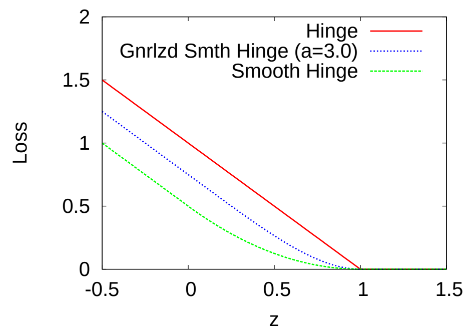

Figure 1: Shown are the Hinge (top), Generalized Smooth Hinge ($\alpha = 3$) (middle), and Smooth Hinge (bottom) Loss functions. Note that all three are zero for $z \geq 1$ and have constant slope of -1 for $z \leq 0$.

References

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

[2] J. D. M. Rennie. Modified regularized least squares classification. http://people.csail.mit.edu/~jrennie/writing, February 2005.