A Comparison of McCullagh's Proportional Odds Model to Modern Ordinal Regression Algorithms

Jason D. M. Rennie
jrennie@gmail.com

October 24, 2006*

Abstract

We introduce McCullagh's Proportional Odds as the foundation for modern Ordinal Regression approaches. Proportional Odds introduced the ideas of (1) mapping examples to the real number line, and (2) segmenting the real number line using a set of thresholds. We compare against two modern approaches to Ordinal Regression which use the framework established by Proportional Odds and find some surprising similarities.

1 Proportional Odds

McCullagh's Proportional Odds model (1980) assumes that:

- Examples are represented as d-dimensional real-valued feature vectors, $x \in \mathbb{R}^d$,
- Each example has an underlying score, defined by the dot-product between the feature vector and a weight vector, $w \in \mathbb{R}^d$,
- Each example is associated with a discrete, ordinal label, $y \in \{1, \ldots, l\}$,
- The real number line is segmented via a set of threshold values, $-\infty \equiv \theta_0 < \theta_1 < \theta_2 < \cdots < \theta_{l-1} < \theta_l \equiv +\infty$, and
- Each label, y, is associated with a segment of the real number line, (θ_{y-1}, θ_y).

Define $g(z) \equiv \frac{1}{1+e^{-z}}$, the sigmoid function. Proportional Odds defines the cumulative likelihood of an example being associated with a label less-than-or-equal-to j (for $j \leq l-1$) as the sigmoid function,

$$P_{\text{PO}}(y \leq j|x) = g(\theta_j - w^T x) = \frac{1}{1 + \exp(w^T x - \theta_j)}. \quad (1)$$

*Updated October 30, 2006.

By definition, $P(y \leq l|x) = 1$. Note that $P(y \leq 1|x) \equiv P(y = 1|x)$. We calculate other likelihoods by taking cumulative differences,

$$P_{\text{PO}}(y = j|x) = P_{\text{PO}}(y \leq j|x) - P_{\text{PO}}(y \leq j-1|x) \qquad \text{for } j \geq 2, \tag{2}$$
$$= \frac{1}{1 + \exp(w^T x - \theta_j)} - \frac{1}{1 + \exp(w^T x - \theta_{j-1})}.$$

Parameters (w, θ) are learned via maximum likelihood. One can "modernize" Proportional Odds by introducing a Gaussian prior on the weight vector (w), which would serve to regularize the predictor. Also, the kernel trick (Wikipedia, 2006) can be used to yield a non-linear predictor without increasing memory consumption.

2 Immediate Thresholds with a Logistic Loss

Rennie and Srebro (2005) use a framework for Ordinal Regression which has similarities to Proportional Odds. In particular, consider Immediate Thresholds (IM) with a Logistic loss[1]. The "loss" function is

$$\text{Loss}_{\text{IM}}(j|x) = \log\left[1 + \exp(w^T x - \theta_j)\right] + \log\left[1 + \exp(\theta_{j-1} - w^T x)\right] \tag{3}$$
$$= -\log\left[g(\theta_j - w^T x)\right] - \log\left[g(w^T x - \theta_{j-1})\right] \tag{4}$$

Note that we can artificially reproduce this from Proportional Odds by treating each $y = j$ event as two events: $y \leq j$ and $y > j - 1$. Treating these two events as independent, we get an (unnormalized) likelihood[2] of

$$P_{\text{IM}}(y = j|x) \propto \frac{1}{\left[1 + \exp(w^T x - \theta_j)\right]\left[1 + \exp(\theta_{j-1} - w^T x)\right]}, \tag{5}$$

which corresponds exactly with the IM loss. Note that we can similarly artificially reproduce All Thresholds (w/ Logistic loss) by splitting the $y = j$ event into $l - 1$ events.

2.1 Comparison with Proportional Odds

Figure 1 compares the loss functions for Proportional Odds and Immediate Thresholds using $z \equiv w^T x$ as a single parameter (as well as Gaussian Processes, discussed in section 3). The two plots (for PO and IM) look strikingly similar. In fact, they are identical save for a small constant difference of 0.0185 (PO loss

[1] Shashua and Levin's "fixed margin" approach (2003) is equivalent to what Rennie and Srebro (2005) call Immediate Thresholds with a Hinge loss.

[2] The loss function for a probabilistic model is the negative log probability.

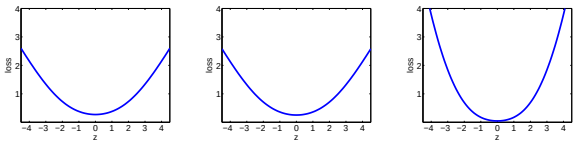

Figure 1: Loss functions for (left) Proportional Odds, (center) Immediate Thresholds, and (right) Gaussian Processes. The x-axis is the predictor output, $z \equiv w^T x$; the y-axis is the loss (or negative log-likelihood). The thresholds are at $\theta_j = 2$ and $\theta_{j-1} = 2$.

is larger). Why is this? Consider the difference between loss functions:

$$\begin{aligned}\Delta \equiv \text{Loss}_{\text{IM}}(j|x) + \log P(y=j|x) &= \log\left[\frac{g(\theta_j - w^T x) - g(\theta_{j-1} - w^T x)}{g(\theta_j - w^T x) g(w^T x - \theta_{j-1})}\right] \\ &= \log\left[\frac{1 + \exp(w^T x - \theta_{j-1}) - 1 - \exp(w^T x - \theta_j)}{[1+\exp(w^T x - \theta_j)][1+\exp(w^T x - \theta_{j-1})]}\right. \\ &\qquad \left.\frac{}{[1+\exp(w^T x - \theta_j)][1+\exp(\theta_{j-1} - w^T x)]}\right] \\ &= \log\left[\exp(w^T x - \theta_{j-1}) - \exp(w^T x - \theta_j)\right] - \log[1+\exp(w^T x - \theta_{j-1})] \\ &\qquad + \log[1+\exp(\theta_{j-1} - w^T x)]\end{aligned} \tag{6}$$

Now, consider the partial derivative of the difference with respect to w,

$$\frac{\partial \Delta\text{Loss}}{\partial w} = x - \frac{x\exp(\theta_{j-1} - w^T x)}{1+\exp(\theta_{j-1} - w^T x)} - \frac{x\exp(w^T x - \theta_{j-1})}{1+\exp(w^T x - \theta_{j-1})} = 0 \tag{7}$$

A change in the weight vector, w, does not affect the difference in loss functions. This explains the fact that in Figure 1 the PO and IM plots differ by a constant amount. Now, consider the partial derivative with respect to a threshold, θ_j,

$$\frac{\partial \Delta\text{Loss}}{\partial \theta_j} = \frac{\exp(w^T x - \theta_j)}{\exp(w^T x - \theta_{j-1}) - \exp(w^T x - \theta_j)} = \frac{1}{\exp(\theta_j - \theta_{j-1}) - 1} \tag{8}$$

Note that $\theta_j \geq \theta_{j-1}$; i.e. the partial derivative is non-negative. We see that as the difference in thresholds increases, so does the difference in loss functions. However, this effect disappears quickly as the difference between the thresholds increases. If we had compared Proportional Odss and Immediate Thresholds with a different upper threshold (θ_j), we would have found a different constant difference between the loss functions.

We conclude that if the thresholds are fixed a priori, then the two models (Proportional Odds and Immediate-Thresholds) behave identically. However, the models are affected differently by the settings of the threshold values.

3 Gaussian Processes

Chu and Ghahramani (2004) introduce a Gaussian Process framework for Or-

dinal Regression. Similar to Proportional Odds, they map each example to the real line via a function and segment the real line via a set of thresholds. They also use a probabilitistic model which is normalized over the set of labels. They posit an "ideal" likelihood[3],

$$P_{\text{ideal}}(y = j|w^T x) = \begin{cases} 1 & \theta_{j-1} < w^T x \leq \theta_j \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

and assume a Gaussian prior over the data point location to account for noise in the data. They integrate the "ideal" likelihood over the Gaussian prior to arrive at the likelihood,

$$P(y = j|x) = \int P_{\text{ideal}}(y = j|w^T x + \delta)\mathcal{N}(\delta; 0, \sigma^2)d\delta = \Phi(z_1) - \Phi(z_2), \tag{10}$$

where $z_1 = \frac{\theta_j - w^T x}{\sigma}$, $z_2 = \frac{\theta_{j-1} - w^T x}{\sigma}$, $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(z; 0, 1)dz$ and $\mathcal{N}(\delta; 0, \sigma^2)$ denotes a Gaussian PDF with random variable δ, zero mean and σ^2 variance.

3.1 Comparison with Proportional Odds

Recall that the likelihood for Proportional Odds is the difference between two sigmoids (2). The likelihood for Chu and Ghahramani's model is the difference between two Gaussian CDFs. If we replace the Gaussian PDF in (10) with the derivative of the sigmoid, we arrive at the Proportional Odds likelihood,

$$P(y = j|x) = \int P_{\text{ideal}}(y = j|w^T x + \delta)g(\delta)(1 - g(\delta))d\delta \tag{11}$$

$$= g(\theta_j - w^T x) - g(\theta_{j-1} - w^T x). \tag{12}$$

The effect of the standard deviation (σ) in the Gaussian can be achieved by using a more general version of the sigmoid (with a scaling parameter). A similar effect can be achieved by including a prior (regularizer) on the weight vector (w). I.e. at its core, Chu and Ghahramani's model is Proportional Odds with a swap of the sigmoid for a Gaussian cumulative distribution. As can be observed in Figure 1, Proportional Odds imposes an approximately linear penalty on margin violations, whereas Chu and Ghahramani's model imposes an approximately quadratic penalty. As a result, Chu and Ghahramani's model may be overly sensitive to outliers.

References

Chu, W., & Ghahramani, Z. (2004). *Gaussian processes for ordinal regression* (Technical Report). University College London.

[3]For simplicity and consistentcy, we use a linear predictor here. Chu and Ghahramani's framework allows for general (non-linear) predictors.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, *42*, 109–142.

Rennie, J. D. M., & Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. *Proceedings of the 22nd International Conference on Machine Learning*.

Shashua, A., & Levin, A. (2003). Ranking with large margin principle: Two approaches. *Advances in Neural Information Processing Systems 15*.

Wikipedia (2006). Kernel trick. http://en.wikipedia.org/wiki/Kernel_trick.