Loss Functions for Preference Levels: Regression with Discrete Ordered Labels

Jason D. M. Rennie
Massachusetts Institute of Technology
Comp. Sci. and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
jrennie@csail.mit.edu

Nathan Srebro
University of Toronto
Department of Computer Science
Toronto, Ontario M5S 3G4, Canada
nati@cs.toronto.edu

Abstract

We consider different types of loss functions for discrete ordinal regression, i.e. fitting labels that may take one of several discrete, but ordered, values. These types of labels arise when preferences are specified by selecting, for each item, one of several rating "levels", e.g. one through five stars. We present two general threshold-based constructions which can be used to generalize loss functions for binary labels, such as the logistic and hinge loss, and another generalization of the logistic loss based on a probabilistic model for discrete ordered labels. Experiments on the 1 Million MovieLens data set indicate that one of our construction is a significant improvement over previous classification- and regression-based approaches.

1 Introduction

In many systems, users specify preferences by selecting, for each item, one of several rating "levels", e.g. one though five "stars". When learning to predict further preferences, these rating levels serve as target labels (responses). This type of discrete ordered labels differs from more standard types of target labels encountered in other machine learning problems: binary labels (as in classification tasks), discrete, unordered labels (as in multi-class classification tasks) and continuous real-valued labels (as in typical regression tasks). Rating levels are discrete with a finite number of possibilities, like class labels in multiclass classification. However, unlike a standard multiclass classification setting, the labels are ordered—a rating of "three stars" is between a rating of "two stars" and a rating of "four stars".

Two obvious approaches for handling discrete ordinal labels are (1) treating the different rating levels as unrelated classes and learning to predict them as in a multiclass classification setting, and (2) treating them as a real-valued responses and using a standard regression setting with a loss function such as sum-squared error. However, neither of these reflects the specific structure of discrete ordinal labels.

1.1 Ordinal regression

In this paper we view fitting rating levels as a regression problem with discrete ordered labels. We view this as a generalization of binary regression (as in, e.g., logistic regression), which can be seen as a degenerate case in which only two levels, "positive" and "negative", are available. As with binary regression, we learn a real-valued *predictor* $z(x)$ (e.g. in linear binary regression, we would learn a linear function of the features), minimizing some *loss* $\text{loss}(z(x); y)$ on the target labels. Common choices for the loss function are the logistic loss (as in logistic regression), and the hinge loss (distance from the classification margin) used in Support Vector Machines. Here, we consider various generalizations to these loss functions suitable for multiple-level discrete ordinal labels.

Threshold-based approaches

Crammer and Singer [2002] suggest a generalization of the Perceptron algorithm for discrete ordinal labels: instead of the single threshold of the perceptron, they use $K-1$ thresholds to separate the real line to regions corresponding to K possible rating levels. Shashua and Levin [2003] suggest a similar generalization to Support Vector Machine (SVM): the single margin constraints (for each observation) of standard SVMs are replaced with a pair of margin constraints on the thresholds bounding the "correct" region (the region corresponding to the target label).

When slack is allowed, Shashua and Levin's approach can be seen as regularized regression with a specific generalization to the hinge loss, which we describe in Section 3.1 as the *immediate-threshold* generalization of the hinge loss. In Section 3.2 we discuss a different generalization, the *all-threshold* generalization, where constraints (and slack) are considered for all $K-1$ thresholds and not only those immediately bounding the correct region. We argue that such a generalization better penalizes predictions which violate multiple thresholds and present experimental evidence suggesting it might be more appropriate. We also discuss how other loss functions, such as the logistic loss or smooth variants of the hinge loss, can also be generalized in the same way.

Probabilistic approaches

Other than these threshold-based generalizations, we also suggest a different generalization to logistic regression, which we term "ordistic regression" (Section 4), that, like logistic regression, can be viewed as fitting a conditional probability model $P(y|x)$. We note that Chu and Ghahramani [2004] recently suggested a different generalization to the logistic

conditional model for discrete ordinal labels.

Probabilistic models for discrete ordinal response have also been studied in the statistics literature [McCullagh, 1980; Fu and Simpson, 2002]. However, the models suggested are much more complex, and even just evaluating the likelihood of a predictor is not straight-forward. On the other hand, in the ordistic model both the log-likelihood and its derivatives can be easily computed, using calculations similar to those used in standard (binary) logistic regression.

1.2 Other approaches

We briefly mention another approach suggested for handling discrete ordinal ratings. Herbrich *et al.* [2000] suggest extracting from the rating levels binary comparison relationships on the rated items and thus mapping the problem to a partial ordering problem. Herbrich *et al.* then study a generalized SVM for learning from binary comparison relationships. A drawback of this approach is the number of order constraints on T items with observed labels can be of order T^2, even though the original input to the problem (the observed labels) is only linear in T.

1.3 Specific contribution

The main contribution of this paper is studying, in a systematic way, different loss functions for discrete ordinal regression. Since our main interest is in how to handle discrete ordinal labels, we focus on regularized linear prediction in a simple learning setting, which we clarify in Section 2. In Section 2.1 we review various loss functions for binary labels and discuss their properties. In Section 3 we present the immediate-threshold and all-threshold constructions mentioned above, using the loss functions from the previous sections as building blocks. In Section 4 we present the ordistic model which generalizes the logistic. In Section 5 we compare the various methods through experiments using the different loss functions, and compare them also to standard multiclass and sum-squared-error regression approaches.

We have already used the immediate-threshold and all-threshold generalizations of the hinge-loss in our work on collaborative prediction using Maximum Margin Matrix Factorizations [Srebro *et al.*, 2005]. Here, we present these constructions in detail and more generally, as well as the ordistic model.

2 Preliminaries

Since our main object of interest is how to handle discrete ordinal labels, we focus on a simple learning setting in which we can demonstrate and experiment with various loss functions. We are given a training set $(x^t, y^t)_{t=1\ldots T}$ of T rated items, where for each item, $x^t \in \mathbb{R}^d$ is a feature vector describing the item and y^t is the rating level for the item. We want to predict preferences of future items. We do so by learning a *prediction mapping* $z(x) : \mathbb{R}^d \to \mathbb{R}$ such that for an item with feature vector x, $z(x)$ corresponds as well as possible to the appeal of the item (i.e. is high if the item is highly rated and low otherwise). We investigate different *loss functions* $\text{loss}(z; y)$ for measuring the goodness of the correspondence between $z(x^t)$ and the target rating level y^t.

In this paper, we focus on L_2-regularized linear prediction, where $z(x) = w'x + w_0$ is a linear (or more accurately, affine) function of $x \in \mathbb{R}^d$, parametrized by a weight vector $w \in \mathbb{R}^d$ and a bias term $w_0 \in \mathbb{R}$. We seek a linear predictor that minimizes a trade-off between the overall training loss and the (Euclidean) norm of the weights:

$$J(w) = \sum_t \text{loss}(w'x^t + w_0; y^t) + \frac{\lambda}{2}|w|^2 \qquad (1)$$

where λ is a trade-off parameter set using cross-validation.

2.1 Binary Regression

We first review common loss functions used with binary labels (i.e. in a binary classification setting), where $y \in \pm 1$. These serve as a basis for our more general loss functions for discrete ordinal labels. We go into some detail regarding aspects of these loss functions which will be relevant in our constructions in the following sections.

Zero-one error

Our objective in binary regression is to be able to correctly predict a binary label. The obvious way of predicting a binary label from a real-valued predictor $z(x) = w'x + w_0$ is thresholding the predictor, and predicting $\text{sign}(z(x))$. The simplest conceivable loss function is a loss function counting the number of prediction errors:

$$\text{loss}(z; y) = \begin{cases} 0 & \text{if } yz > 0 \\ 1 & \text{if } yz \leq 0 \end{cases} \qquad (2)$$

However, this simple loss function is problematic for several reasons:

- It is not convex, and minimizing it is a difficult (in fact, NP-hard) optimization problem.
- It is not continuous, let alone differentiable, and so even local optimization is difficult.
- It is insensitive to the magnitude of z, and so also to the magnitude of w. Regularizing w is therefore meaningless, as shrinking w and w_0 towards zero would yield the same error, but with a regularization term approaching zero.

Margin

The third problem can be addressed by requiring not only that z predict y correctly, but that it does so with a margin:

$$\text{loss}(z; y) = \begin{cases} 0 & \text{if } yz \geq 1 \\ 1 & \text{if } yz < 1 \end{cases} \qquad (3)$$

This modified loss function is sensitive to the magnitude of z, and therefore also to the magnitude of w. Summing this loss function corresponds to counting the number of violations of the constraints $y(w'x + w_0) \geq 1$. Rewriting these constraints as $y(\frac{w'}{|w|}x + \frac{w_0}{|w|}) \geq \frac{1}{|w|}$, we can interpret $\frac{1}{|w|}$ as a geometrical margin around the separating hyperplane, specified by its normal $\frac{w'}{|w|}$. Minimizing the loss (3) as well as the L_2 regularizer $|w|$ can therefore be interpreted as maximizing the separation margin $M = \frac{1}{|w|}$ while minimizing the number of training points not classified correctly with a margin of at least M.

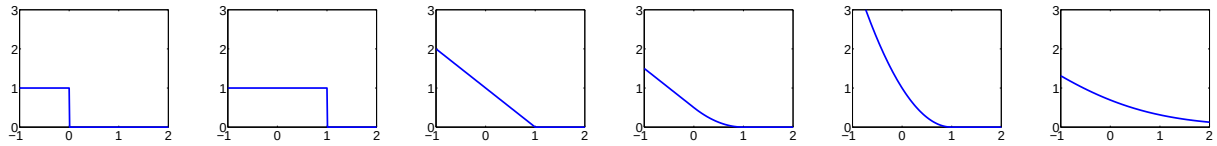

Figure 1: Different margin penalty functions $f(yz)$ (left to right): (1) sign agreement, (2) margin agreement, (3) hinge, (4) smooth hinge, (5) modified least squares, (6) logistic.

Hinge loss

Minimizing the margin loss (3) might be a good ideal, but this loss function is still non-convex and non-continuous. The common approach to large-margin classification is therefore to minimize the *hinge loss*:

$$\text{loss}_h(z;y) = h(yz) \tag{4}$$

where $h(z)$ is the *hinge function*:

$$h(z) = \max(0, 1-z) = \begin{cases} 0 & \text{if } z \geq 1 \\ 1-z & \text{if } z < 1 \end{cases} \tag{5}$$

This is the loss function typically minimized in soft-margin Support Vector Machine (SVM) classification. In the context of SVMs, the hinge loss is usually written as a sum over margin violations ξ^t included in the constraints $y(w'x + w_0) \geq 1 - \xi^t$.

An important property of the hinge-loss is that it is an upper bound on the zero-one misclassification error (2), and thus large-margin generalization error bounds bounding its value on examples not in the training set also bounds the value of the zero-one misclassification error, which is perhaps the true object of interest.

Smoothed hinge loss

Other loss functions share properties of the hinge, but are easier to minimize since they have a smooth derivative. We introduce "smooth" hinge loss as an approximation to the hinge that is easier to minimize:

$$h(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ (1-z)^2/2 & \text{if } 0 < z < 1 \\ 0.5 - z & \text{if } z \leq 0 \end{cases} \tag{6}$$

Modified least squares

Zhang and Oles [2001] suggest a different loss function with a smooth derivative, in which the hinge function is replaced with a truncated quadratic:

$$h(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ (1-z)^2 & \text{if } z < 1 \end{cases} \tag{7}$$

The modified least squares loss based on (7) is much more sensitive to outliers and large errors then the hinge loss using (5) or smoothed hinge loss using (6).

The margin error (3), which we might want to view as a non-convex "ideal", does not pay any attention to the magnitude of the error, and penalizes all errors equally. This allows for a few outliers fairly cheaply, but leads to a non-convex objective. The hinge loss, as well as the smoothed hinge, introduce a linear dependence on the magnitude of the error, but such a linear (at least) dependence is unavoidable in a convex loss function. The modified least squares goes beyond this necessary dependence on the magnitude of the error, and introduces an unnecessary (from the point of view of convexity) quadratic dependence, further deviating from the zero/one margin error.

Logistic regression

Another common loss function, which can also be written as a function of the classification margin yz, is the logistic loss:

$$\text{loss}_g(z;y) = g(yz) \tag{8}$$

$$g(z) = \log(1 + e^{-z}) \tag{9}$$

The logistic loss can be viewed as a negative conditional log-likelihood $\text{loss}_g(z;y) = -\log P(z|y)$ for a logistic conditional model $P(y|z) \propto e^{yz}$ (i.e. Y is a Bernoulli random variable, with natural parameter z). The predictor $z(x) = w'x + w_0$ minimizing the summed logistic loss is thus the maximum conditional likelihood estimator among the parametric class of conditional models $P(y|x) \propto e^{y(w'x+w_0)}$. Introducing an L_2 regularizer as in (1) corresponds to maximum a-posteriori (MAP) estimation with a Gaussian prior on the weight vector w.

As discussed above, logistic regression corresponds to maximum *conditional* likelihood estimation for a *conditional* parametric model $P(y|x)$. It is worth noting that this parametric family of conditional models $P(y|x) \propto e^{y(w'x+w_0)}$ is exactly the family of conditional distributions $P(y|x)$ for joint distributions $P(y,x)$ where $X|Y$ follows a multivariate spherical Gaussian distribution with variance which does not depend on Y, and a mean which does depend on Y, i.e.:

$$P(x|y) \propto e^{\frac{1}{2\sigma^2}|x-\mu_y|^2} \tag{10}$$

where $\mu_{-1}, \mu_1 \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}$, as well as the prior probability $P(Y=1)$ are the parameters of the joint distribution model.

For our purposes below, it would also be useful to view the conditional model $P(y|z) \propto e^{yz}$ similarly, as the conditional distribution arising from a joint distribution $P(y,z)$ in which $P(y=1) = \frac{1}{2}$ and $Z|Y \sim N(1,Y)$, i.e.:

$$P(z|y) \propto e^{\frac{1}{2}|z-y|^2}. \tag{11}$$

Loss as a function of classification penalties We note that all loss functions discussed here can be viewed as penalties $\text{loss}(y;z) = f(yz)$ imposed on the *classification margins* yz, and differ only in the *margin penalty function* $f(\cdot)$ used. The various margin penalty functions discussed are shown in Figure 1.

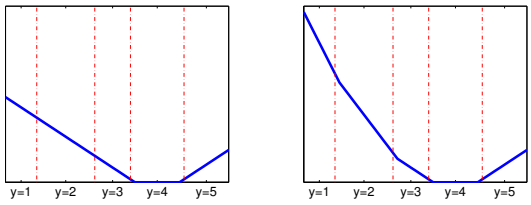

Figure 2: Threshold-based loss constructions: (left) immediate-threshold, and (right) all-threshold. Both plots show a generalization of the hinge loss for an example with label $y = 4$.

2.2 Discrete Ordinal Regression

So far, we discussed loss functions for binary labels $y = \pm 1$. However, the main topic of this paper is loss functions for discrete ordinal labels. In this scenario, the labels y can take on K distinct values which we denote $\{1, 2, \ldots, K\}$.

In the next section, we present different constructions for generalizing margin-based loss function for binary labels, of the form $\text{loss}(z; y) = f(yz)$, to discrete ordinal labels. We do so by introducing $K - 1$ thresholds, instead of the single threshold (zero) in the binary case.

We also introduce a generalization of the logistic loss in which the joint probabilistic model $P(y, z)$ is generalized to a mixture of K unit-variance Gaussians (instead of a mixture of two unit-variance Gaussians as in the binary case).

2.3 Beyond Feature-Based Regression

Although the main setting we focus on is a feature-based linear regression setting, the loss functions we describe are applicable in more general settings where a real-valued predictor needs to be related to a discrete ordinal label. In fact, our study originated in matrix-completion approaches to collaborative prediction (completing unobserved entries in a partially observed matrix of user preferences), where the "predictors" are entries in a proposed matrix [Srebro *et al.*, 2005].

We focus on linear regression using explicit features, but we note that all methods discussed here can also be "kernelized", as in Support Vector Machines. Both the immediate-threshold and all-threshold constructions with a hinge loss can also be seen as generalizations of SVMs and can be stated as quadratic programs to which optimization techniques typically employed in learning SVMs apply (in fact, Shashua and Levin [2003] introduce the immediate-threshold construction in the context of SVMs).

3 Threshold-Based Constructions

To extend binary loss to the case of discrete ordinal regression, we introduce $K - 1$ thresholds $\theta_1 < \theta_2 < \cdots < \theta_{K-1}$ partitioning the real line to K segments. The exterior segments are semi-infinite, and for convenience we denote $\theta_0 = -\infty$ and $\theta_K = +\infty$. Each of the K segment corresponds to one of the K labels and a predictor value of $\theta_{y-1} < z < \theta_y$ (i.e. in the yth segment) corresponds to a rating of y. This generalizes the binary case, where we used a single threshold, namely zero, separating the real line into a semi-infinite segment, namely $z < 0$, corresponding to negative labels $y = -1$, and a semi-infinite segment, $z > 0$ corresponding to positive labels. In fact, the bias term in the binary setting can be viewed as specifying a threshold. The $K - 1$ thresholds replace this single bias term / threshold.

We describe two different constructions for loss functions based on such thresholds. The constructions differ in how predictors outside the segment segment corresponding to the "correct" label (or too close to its edges) are penalized. Both constructions are based on combining penalties for threshold violation, where each threshold violation is penalized using some margin penalty function $f(\cdot)$.

3.1 Immediate-Threshold

For the immediate-threshold construction, we consider, for each labeled example (x, y), only the two thresholds defining the "correct" segment (θ_{y-1}, θ_y), and penalize violations of these thresholds:

$$\text{loss}(z; y) = f(z - \theta_{y-1}) + f(\theta_y - z), \qquad (12)$$

where $z = z(x)$ is the predictor output for the example. Figure 2 gives an example visualization for $y = 4$. Note that if f is an upper bound on the binary classification zero-one error, then the immediate-threshold loss is an upper bound on the ordinal regression zero-one error. The immediate-threshold loss is ignorant of whether multiple thresholds are crossed.

3.2 All-Threshold

In a multi-class classification setting, all mistakes are equal; there is no reason to prefer one incorrect label over another. This is not true in ordinal regression where there is a distinct ordering to the labels. It is better to predict '4' than '1' if the true label is '5'. This is evident in the evaluation criterion. We use mean absolute error for ordinal regression, which counts the sum of distances between the true and predicted labels. Immediate-threshold bounds zero-one error, but not (necessarily) mean absolute error.

We introduce a construction that bounds mean absolute error. The all-threshold loss is a sum of all threshold violation penalties. If the binary loss function bounds zero-one error, then the all-threshold loss bounds mean absolute error. Define $s(l; y) = \begin{cases} -1 & \text{if } l < y \\ +1 & \text{if } l \geq y \end{cases}$. Then the all-threshold loss is

$$\text{loss}(z; y) = \sum_{l=1}^{T-1} f\Big(s(l; y)(\theta_l - z)\Big). \qquad (13)$$

where $f(\cdot)$ is some margin penalty function. Figure 2 gives an example visualization for $y = 4$. Note that the slope of the loss increases each time a threshold is crossed. Thus, solutions are encouraged that minimize the number of thresholds that are crossed.

3.3 Learning Thresholds

Fitting an ordinal regression models involves fitting the parameters of the predictor, e.g. $z(x) = w'x$, as well as the thresholds $\theta_1, \ldots, \theta_{K-1}$. Learning the thresholds, rather than

fixing them to be equally spaced, allows us to capture the different ways in which users use the available ratings, and alleviates the need for per-user rating normalization. In a setting in which multiple users are considered concurrently, e.g. in collaborative prediction, a different set of thresholds can be learned for each user.

4 A Probabilistic Generalization

Recall that the logistic loss for binary labels can be viewed as the negative conditional log-likelihood $\text{loss}_l(y;z) = -\log P(y|z)$ for a conditional model $P(y|z) \propto e^{yz}$ corresponding to a mixture-of-Gaussians joint distribution $P(y,z)$, as described in Section 2.1. Here, we generalize the logistic by generalizing this probabilistic model to a mixture of K Gaussians, resulting in a similar simple form for the conditional model $P(y|z)$. We refer to this model, and the resulting loss, as the "ordistic" model.

Unlike the threshold-based constructions which are parametrized by $K-1$ thresholds, the ordistic model is parametrized by the K means $\mu_1 < \mu_2 < \cdots < \mu_K$, and possibly also the K prior probabilities $p_1, \ldots, p_K > 0$, $\sum p_i = 1$. Considering a joint probability model in which $P(y) = p_y$ and $Z|Y \sim N(\mu_Y, 1)$, the conditional distribution $P(y|z)$ is:

$$P(y|z) = \frac{p_y e^{-(z-\mu_y)^2/2}}{\sum_i p_i e^{-(z-\mu_i)^2/2}} = \frac{\exp\left(\mu_y z + (\pi_y - \mu_y^2/2)\right)}{\sum_i \exp\left(\mu_i z + (\pi_i - \mu_i^2/2)\right)} \tag{14}$$

where $\pi_i = \log p_i$. If we simplify the model by fixing $p_i = \frac{1}{K}$, these terms drop from the conditional distribution. The ordistic loss, with parameters (μ_i, π_i) is obtained by considering the negative log-likelihood:

$$\text{loss}_o(y;z) = -\log P(y|z) = -\log \frac{e^{\left(\mu_y z + (\pi_y - \mu_y^2/2)\right)}}{\sum_i e^{\left(\mu_i z + (\pi_i - \mu_i^2/2)\right)}}$$

If all means $\mu_1, \ldots, \mu_K$ are allowed to vary freely, regularizing the weight vector w is meaningless, since rescaling it can be corrected by rescaling the means. In order to impose a fixed scale, we fix the extreme means to $\mu_1 = -1$ and $\mu_K = 1$. Unlike threshold-based models, a bias term w_0 is meaningful since the location of the extreme means is fixed. The overall number of parameters for an ordistic model with fixed $p_i = \frac{1}{K}$ is therefore the same as the number of parameters in the threshold constructions ($K-1$ parameters for the $K-2$ means and the bias term, or $K-1$ parameters for the thresholds, in addition to the weight vector). Allowing priors introduces $K-1$ additional parameters.

The ordistic degenerates to the logistic when $K = 2$.

4.1 Difference from soft-max multi-class classification

A common generalization of logistic regression to multi-class classification is to learn a predictor $z_i(x) = w_i' x + w_{i0}$ for each class and fit a soft-max conditional model in which $P(y|x) \propto z_y(x) = w_y' x + w_{y0}$. This conditional model corresponds to a joint distribution $P(y,x)$ where $X|Y$ follows a unit-variance spherical Gaussian distribution with mean $\mu_Y^o \in \mathbb{R}^d$. This model differs from the ordistic model in that the means of the K Gaussians are allowed to fall in arbitrary positions in $\mathbb{R}^d$. On the other hand, in the ordistic model, we model Z, rather then X, as a Gaussian mixture. An alternate view of the ordistic model would be to view X as a Gaussian mixture, but in this case, all means would be constrained to be co-linear, since the same weight vector w is used for all labels. This constraint captures the core difference between a standard softmax model and the ordistic model: the ordistic model constrains the different labels to correspond to different extents along the same direction (hence collinear means), rather then arbitrary variations in different directions.

4.2 Derivatives

As with the logistic model, the derivatives of the ordistic loss have a simple form. In order to simplify the derivatives, it will be useful to refer to expectation and probabilities with respect to the joint probabilistic (Gaussian mixture) model for (Z,Y) discussed above. The gradient with respect to z, from which the gradient with respect to the weights can be calculated, is:

$$\begin{aligned}\frac{\partial \text{loss}_o(z;y)}{\partial z} &= -\mu_y + \frac{\sum_i \mu_i \exp\left(\mu_i z + (\pi_i - \mu_i^2/2)\right)}{\sum_i \exp\left(\mu_i z + (\pi_i - \mu_i^2/2)\right)} \\ &= -\mu_y + \sum_i \mu_i \Pr\left(Y = i | Z = z; \mu, \pi\right) \\ &= -\mu_y + \mathbf{E}_{\mu,\pi}\left[\mu_Y | Z = z\right] \end{aligned} \tag{15}$$

Similarly, the derivative with respect to the log-priors π_i and the means μ_i can be calculated as:

$$\begin{aligned}\frac{\partial \text{loss}_o(z;y)}{\partial \pi_i} &= \Pr\left(Y = i | Z = z; \mu, \pi\right) - \delta_{y,i} \\ \frac{\partial \text{loss}_o(z;y)}{\partial \mu_i} &= \left(\Pr\left(Y = i | Z = z; \mu, \pi\right) - \delta_{y,i}\right)(z - \mu_i)\end{aligned} \tag{16}$$

where $\delta_{y,i}$ is one if $y = i$ and zero otherwise.

5 Experiments

To determine the appropriateness of the different constructions discussed earlier, we conducted experiments on a well-known collaborative filtering data set. We implemented the two threshold-based constructions discussed in Section 3. We also implemented multi-class classification and sum-squared error regression constructions to compare against.

We used the "1 Million" MovieLens data set for evaluation. The data set contains 1,000,209 rating entries, made by 6040 users on 3952 movies. Similar to the work of Crammer and Singer [2002] and Shashua and Levin [2003], we used the ratings of the top 200 users as "features" to predict the ratings of the remaining users. To deal with "missing" ratings, we subtracted the user's mean rating and filled-in empty values with zero. We used the remaining 5840 users' ratings as labels for ordinal regression. For each user, we used one randomly selected rating for testing, another for validation and the remaining ratings for training. We limited our experiments to

	Multi-class Test MAE	Imm-Thresh Test MAE	All-Thresh Test MAE	
Mod. Least Squares	0.7486	0.7491	**0.6700**	(1.74e-18)
Smooth Hinge	0.7433	0.7628	**0.6702**	(6.63e-17)
Logistic	0.7490	0.7248	**0.6623**	(7.29e-22)
	Multi-class Test ZOE	Imm-Thresh Test ZOE	All-Thresh Test ZOE	
Mod. Least Squares	0.5606	0.5807	**0.5509**	(7.68e-02)
Smooth Hinge	0.5594	0.5839	**0.5512**	(1.37e-01)
Logistic	0.5592	0.5699	**0.5466**	(2.97e-02)

Table 1: Mean absolute error (MAE) and zero-one error (ZOE) results on MovieLens. For each construction/loss and error type, we selected the regularization parameter with lowest validation error. Numbers in parentheses are p-values for all-threshold versus the next best construction. As a baseline comparison, simple sum-squared-error (L2) regression achieved test MAE of 1.3368 and test ZOE of 0.7635.

the top 2000 movies to ensure a minimum of 10 ratings per movie. This gave us test and validation sets of size 5,840 and a training set of 769,659 ratings.

For each method (combination of construction method and margin penalty), and range of values of the regularization parameter λ, we fit weight and threshold vectors for each user by minimizing the convex objective (1) using conjugate gradient descent. We calculated mean absolute error (MAE) and zero-one error (ZOE) between predicted and actual ratings on the validation set and used the regularization parameter with the lowest validation set MAE/ZOE for test set evaluation.

Table 1 shows test MAE and ZOE for various constructions and margin penalty functions. Across all penalty functions, all-threshold yields the lowest MAE. The differences are highly significant according to a nonparametric, two-tailed binomial test—the largest p-value is 6.63e-17. Interestingly, all-threshold also yields lower ZOE, although the comparison with multi-class classification is not conclusive (p-values around 0.03–0.1). The difference in ZOE compared to immediate-threshold is highly significant, with p-values at most 7.84e-06 (not shown).

Results indicate that the choice of construction is more important than penalty function—all-threshold with the worst-performing penalty function yields lower MAE and ZOE than the best non-all-threshold combination. However, it appears that the logistic loss tends to work best; in particular, the differences in MAE between logistic and other penalty functions (for the all-threshold construction) are significant at the $p = 0.01$ level (largest p-value is 9.52e-03) according to the two-tailed binomial test.

6 Conclusion

We presented a study of discrete ordinal regression in a general, loss function based framework. We began with binary classification and described two threshold-based constructions as generalizations of binary loss functions. We also discussed an alternate generalization of logistic regression. We conducted experiments on MovieLens using the two threshold-based constructions and found that the all-threshold construction outperformed multi-class classification and simple regression methods, as well as the immediate-threshold construction.

Acknowledgements

Jason Rennie was supported in part by the DARPA CALO project.

References

[Chu and Ghahramani, 2004] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. Technical report, University College London, 2004.

[Crammer and Singer, 2002] Koby Crammer and Yoram Singer. PRanking with ranking. In *Advances in Neural Information Processing Systems 14*, 2002.

[Fu and Simpson, 2002] Limin Fu and Douglas G. Simpson. Conditional risk models for ordinal response data: simultaneous logistic regression analysis and generalized score tests. *Journal of Statistical Planning and Inference*, pages 201–217, Nov 2002.

[Herbrich *et al.*, 2000] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

[McCullagh, 1980] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

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

[Srebro *et al.*, 2005] Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

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