Bayesian Hierarchical Ordinal Regression

Ulrich Paquet, Sean Holden, and Andrew Naish-Guzman

Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK
ulrich.paquet, sean.holden, andrew.naish-guzman@cl.cam.ac.uk

Abstract. We present a Bayesian approach to ordinal regression. Our model is based on a hierarchical mixture of experts model and performs a soft partitioning of the input space into different ranks, such that the order of the ranks is preserved. Experimental results on benchmark data sets show a comparable performance to support vector machine and Gaussian process methods.

1 Introduction

Many applications in Machine Learning require the prediction of ordered categories, and thereby ask of us to bridge the gap between regression and classification problems. *Ordinal regression*, or ranking, often arise when a judgment of preference is made. In collaborative filtering, for example, we seek to predict a consumer's rating of a novel item on an ordinal scale such as *good* $>$ *average* $>$ *bad*, using past ratings of similar items. The problem shares properties with classification since the targets are discrete and finite, but also with regression estimation by the existence of an ordering in the target space.

In this paper we adopt a Bayesian approach to the ordinal regression problem, based on the *hierarchical mixture of experts* (HME) model (Jordan & Jacobs, 1994; Waterhouse et al. 1996). The HME model consists of a hierarchy of 'experts', where each expert models some data-generating process on a subset of the data. We simplify each expert to an indicator function, such that an expert is responsible for labeling a pattern with a certain rank on a subset of the input space. The ordering of the targets is imposed by a left-to-right assignment of ranks to experts in a binary HME tree.

2 Learning From Examples

We are given a data set $\mathcal{D}$ of independent and identically distributed examples of real-valued input vectors $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ and corresponding targets $\mathbf{y} = \{y_n\}_{n=1}^N$. The targets come from a space $\mathcal{Y}$ consisting of a finite number of ranks, $\mathcal{Y} = \{1, \ldots, R\}_>$. The subscript $>$ denotes that there is an ordering between the ranks, and can be interpreted as 'preferred to'. For simplicity we use integers to indicate the ordered set of ranks, but any labels will do. Given a new example $\mathbf{x}_*$ and the observed data, we wish to determine the probability distribution of its rank, $P(y_* = r|\mathbf{x}_*, \mathcal{D})$.

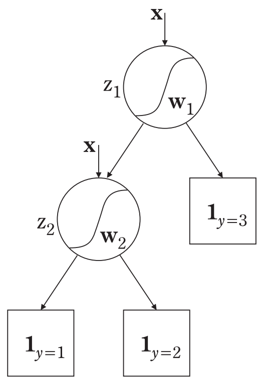

Fig. 1. A binary mixture of experts tree for ordinal regression. The expert (leaf) nodes are indicator functions, each responsible for labeling one possible rank. Here $\mathbf{1}_A$ is one if A is true, and zero otherwise. The gating nodes indicate the probability of following the left—or conversely right—branch down the tree to a rank. The structure of the HME tree, with a left-to-right assignment of ranks to the 'experts', encapsulates the ordinal regression problem.

3 Hierarchical Mixture of Experts for Ordinal Regression

We formulate the distribution of the ordinal target variables with a binary mixture of experts tree. Figure 1 illustrates such a tree, where the leaves, called 'experts', are component distributions of the targets. The non-leaf nodes, called 'gates', form coefficients that mix the experts. Each gate is conditioned on an input variable and indicates the probability of following its left—or conversely right—branch down the tree; consequently the gates perform a soft partitioning of the input space. This soft partitioning is used as our ordinal regression model.

We associate a binary variable z_i with each gate, and set it to one if the left branch is followed from the ith gate. The parameters of the model are the real-valued weight vectors of the gates, which we indicate with $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^{I}$. The experts are labeled with discrete labels $1, \ldots, R$, and we require the experts to be indicator functions. Hence, given expert r, the probability that it labeled $(\mathbf{x}, y)$ is one if $y = r$, and zero otherwise. With a left-to-right assignment of ranks to the experts, the structure of the HME tree and the resulting partitioning of the input space impose a natural ordering on the targets. In this paper we restrict ourselves to complete binary trees, although a more judicious choice of tree structure, based on evidence maximization, can be made.

The probability of y having rank r, given $\mathbf{x}$, is equal to the probability that expert r was responsible for generating the target. Equivalently it is equal to the probability of correctly setting the binary indicator variables z_i to form a path from the root to the rth 'expert',

$$P(y = r|\mathbf{x}, \mathbf{W}) = \prod_{i:\text{root}\to r} P(z_i|\mathbf{x}, \mathbf{w}_i). \tag{1}$$

We use notation $i : \text{root} \to r$ to indicate that the product is taken over the gates on the unique path from the root to the rth expert, and note that summing (1) over all ranks give unity. By defining $\sigma(a) = 1/(1 + e^{-a})$, the probability of following the left branch from the ith gate is

$$P(z_i = 1|\mathbf{x}, \mathbf{w}_i) = \sigma(\mathbf{w}_i^\top \mathbf{x}).$$

Throughout this paper, we implicitly augment input vectors with a bias clamped at 1. From (1), the likelihood of observing the entire data set is

$$P(\mathcal{D}|\mathbf{W}) \equiv P(\mathbf{y}|\mathbf{X}, \mathbf{W}) = \prod_{n=1}^{N} \prod_{i:\mathrm{root} \to y_n} P(z_{in}|\mathbf{x}_n, \mathbf{w}_i). \tag{2}$$

3.1 The Posterior

A probabilistic formulation—often prone to overfitting, as in the familiar case of supervised learning—can be found by maximizing the likelihood (2) with respect to the model parameters $\mathbf{W}$. We rather use the usual Bayesian approach of making predictions by computing the expected value of $P(y_* = r|\mathbf{x}_*, \mathbf{W})$ for a new example $\mathbf{x}_*$ with respect to the posterior distribution of $\mathbf{W}$. For the purpose of obtaining this posterior distribution from Bayes' theorem, we place a Gaussian prior on each gate's parameter vector,

$$p(\mathbf{w}_i|\alpha_i) = \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left\{ - \frac{\alpha_i}{2} \mathbf{w}_i^\top \mathbf{w}_i \right\},$$

and combine it with the likelihood (2), normalized by the evidence. The hyperparameter α_i controls the width of the prior.

The weight vector of gate i, conditioned on the observed data, is independent of the parameters of the other gates, and only dependent on the examples that were labeled by its left and right subtrees. As a notational convenience, let $\mathcal{T}_i$ indicate the set of experts that are leaves in the subtree with gate i as root. Define $\mathcal{D}_i$ to be the subset of examples associated with $\mathcal{T}_i$. From Bayes' theorem, the posterior distribution of each gate's parameters is

$$p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i) = \frac{P(\mathcal{D}_i|\mathbf{w}_i) p(\mathbf{w}_i|\alpha_i)}{p(\mathcal{D}_i|\alpha_i)} \tag{3}$$

$$\propto \prod_{n:y_n \in \mathcal{T}_i} \sigma(\mathbf{w}_i^\top \mathbf{x}_n)^{z_{in}} (1 - \sigma(\mathbf{w}_i^\top \mathbf{x}_n))^{1 - z_{in}} \exp\left\{ - \frac{\alpha_i}{2} \mathbf{w}_i^\top \mathbf{w}_i \right\}. \tag{4}$$

The full posterior is simply the product over all individual gate posterior distributions, $p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) = \prod_{i=1}^{I} p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)$.[1]

3.2 Inference

To determine the rank of a new example $\mathbf{x}_*$, we marginalize over the posterior distribution of the weights, given the observed data:

$$P(y_* = r|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\alpha}) = \int P(y_* = r|\mathbf{x}_*, \mathbf{W})\, p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha})\, d\mathbf{W}$$

$$= \prod_{i:\mathrm{root} \to r} \int P(z_i|\mathbf{x}_*, \mathbf{w}_i)\, p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)\, d\mathbf{w}_i. \tag{5}$$

[1] Ideally we want $p(\mathbf{W}|\mathcal{D}) = \int p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}|\mathcal{D})\, d\boldsymbol{\alpha}$, a matter that we shall touch on in Section 3.3.

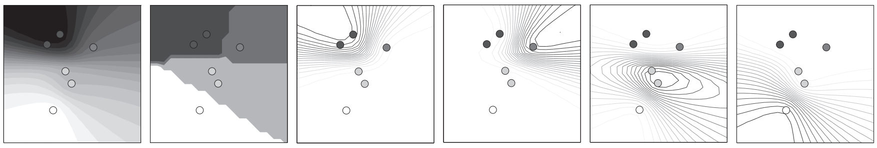

Fig. 2. An example showing four ranks. Shown from left to right is the expected rank; most probable rank; posterior probabilities of ranks 1 to 4.

Figure 2 illustrates a toy problem with four ranks, and the respective posterior probabilities of each rank.

It is not possible to perform the integration in (5) analytically, so we make a Laplace approximation (MacKay, 1992) to each $p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)$. Laplace's method involves a quadratic approximation of the log-posterior around its mode: the negative logarithm of the posterior (3) is maximized over $\mathbf{w}_i$ to give the most probable weight vector $\mathbf{w}_{\mathrm{MP}_i}$. We find $\mathbf{w}_{\mathrm{MP}_i}$ by setting the first derivative of $-\ln p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)$ to zero and solving with a standard Newton-Raphson method. The second-order Taylor expansion of $-\ln p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)$ around its maximum $\mathbf{w}_{\mathrm{MP}_i}$ allows us to approximate the posterior with a Gaussian distribution with mean $\mathbf{w}_{\mathrm{MP}_i}$ and variance-covariance matrix $\mathbf{A}_i^{-1}$. Here $\mathbf{A}_i$ is the Hessian, the matrix of second derivatives $-\nabla^2 \ln p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i)$ evaluated at the most probable parameter values $\mathbf{w}_{\mathrm{MP}_i}$. This leads to an approximation of (5) with

$$P(y_* = r|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\alpha}) \simeq \prod_{i:\mathrm{root}\to r} \int P(z_i|\mathbf{x}_*, \mathbf{w}_i)\, \mathsf{Normal}(\mathbf{w}_i; \mathbf{w}_{\mathrm{MP}_i}, \mathbf{A}_i^{-1})\, d\mathbf{w}_i. \quad (6)$$

The probability $P(z_i = 1|\mathbf{x}_*, \mathbf{w}_i) = \sigma(\mathbf{w}_i^\top \mathbf{x}_*)$ has a linear dependence on the weight parameter through the scalar $a_i = \mathbf{w}_i^\top \mathbf{x}_*$, and hence the dimensionality of the integral can be reduced by finding the probability density $p(a_i|\mathbf{x}_*, \mathcal{D}) = 1/\sqrt{2\pi s_i^2} \cdot \exp\{-(a_i - a_{\mathrm{MP}_i})^2/2s_i^2\}$ with the mean and variance given by $a_{\mathrm{MP}_i} = \mathbf{w}_{\mathrm{MP}_i}^\top \mathbf{x}_*$ and $s_i^2 = \mathbf{x}_*^\top \mathbf{A}_i^{-1} \mathbf{x}_*$ respectively. The marginalized output, where each of the integrals in the product (6) is effectively $P(z_i|\mathbf{x}_*, \mathcal{D}_i, \alpha_i)$, is therefore

$$P(z_i = 1|\mathbf{x}_*, \mathcal{D}_i, \alpha_i) = \psi(a_{\mathrm{MP}_i}, s_i^2) \equiv \int \sigma(a_i)\, \mathsf{Normal}(a_i; a_{\mathrm{MP}_i}, s_i^2)\, da_i.$$

The integral of a sigmoid times a Gaussian is approximated by $\psi(a_{\mathrm{MP}_i}, s_i^2) \simeq \sigma(\kappa(s_i^2) \cdot a_{\mathrm{MP}_i})$, with $\kappa(s_i^2) = 1/\sqrt{1 + \pi s_i^2/8}$ (MacKay, 1992), so that we make a final prediction with

$$P(y_* = r|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\alpha}) \simeq \prod_{i:\mathrm{root}\to r} \sigma(\kappa(s_i^2) \cdot a_{\mathrm{MP}_i})^{z_i} [1 - \sigma(\kappa(s_i^2) \cdot a_{\mathrm{MP}_i})]^{1-z_i}.$$

3.3 Finding Values for Hyperparameters $\boldsymbol{\alpha}$

The preferred Bayesian treatment for hyperparameters such as $\boldsymbol{\alpha}$ is to integrate them out of any predictions with $p(\mathbf{w}_i|\mathcal{D}_i) = \int p(\mathbf{w}_i|\mathcal{D}_i, \alpha_i) p(\alpha_i|\mathcal{D}_i)\, d\alpha_i$. We will assume rather that the hyperparameter posterior $p(\alpha_i|\mathcal{D}_i)$ is sharply peaked around its most probable value α_{MP_i}, so that $p(\mathbf{w}_i|\mathcal{D}_i) \simeq p(\mathbf{w}_i|\mathcal{D}_i, \alpha_{\mathrm{MP}_i})$. The hyperparameters which maximize the posterior $p(\alpha_i|\mathcal{D}_i)$ need to be found; by assuming a non-informative hyperprior over α_i, this task amounts to maximizing the likelihood term (or *evidence*: the denominator in (3)). The log of the evidence as a function of α_i is $\ln p(\mathcal{D}_i|\alpha_i) = \frac{d}{2}\ln\alpha_i - \frac{\alpha_i}{2}\mathbf{w}_{\mathrm{MP}_i}^\top \mathbf{w}_{\mathrm{MP}_i} - \frac{1}{2}\ln|\mathbf{A}_i| + c$. Following MacKay (1992), maximizing the log-evidence with respect to α_i leads to $\alpha_{\mathrm{MP}_i} = (d - \alpha_i \mathsf{Trace}(\mathbf{A}_i^{-1}))/\mathbf{w}_{\mathrm{MP}_i}^\top \mathbf{w}_{\mathrm{MP}_i}$, which we use as a re-estimation formula for α_i. The Hessian and most probable weights are recomputed, and the process repeated until convergence of α_i.

3.4 Nonlinear Decision Boundaries

Nonlinearity is introduced to the model with a fixed set of basis functions, and we replace $\mathbf{w}^\top\mathbf{x}$ by $\sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$. For simplicity, we let the basis functions be shared over all the gates. For practical results, we use radial basis functions, $\phi_m(\mathbf{x}) = \exp\{-\frac{1}{2h^2}||\mathbf{x} - \boldsymbol{\mu}_m||^2\}$, and keep one basis function fixed at unity (the bias). The basis function centres are set by a k-means clustering on each rank. The M basis functions used in each gate are the collection of all basis functions over the ranks. The width h of the basis functions is set to twice the average spacing between the cluster centres. We defer other methods of implementing the gates to Sec. 5.

4 Experimental Results

The proposed HME approach to ordinal regression was evaluated on benchmark data sets from Chu & Ghahramani (2004), who have discretized the targets from the data sets, normally used for metric regression, into 5 and 10 ordinal ranks using equal-length binning. The data were partitioned into training and test sets, with a repartitioning performed 20 times on each data set.[2]

We evaluate the accuracy by taking the most likely rank as the predicted rank $\hat{y}_n$, and comparing it to the true rank y_n. If there are N' elements in the test set, the *mean zero-one error* averages the number of incorrect predictions with $\frac{1}{N'}\sum_{n=1}^{N'} \mathbf{1}_{\hat{y}_n \neq y_n}$. For the nonlinear case we added 10 basis functions per rank to the set of basis functions used. Table 1 shows the averages over 20 trials, along with the standard deviation. The first three columns are taken from Chu & Ghahramani (2004), who have compared Gaussian processes with Gaussian basis functions to the support vector machine (SVM) approach of Shashua & Levin (2003). Both a MAP estimation with Laplace approximation (MAP) and Expectation Propagation algorithm with variational bound (EP) was used as inference techniques to implement the Gaussian process. The HME model with both linear and nonlinear gates gives comparable performance.

[2] The datasets and partitions are downloadable from www.gatsby.ucl.ac.uk/∼chuwei/ordinalregression.html.

Table 1. The test results of five algorithms. The data sets used, with (attributes, training instances, test instances), are **Di.** Diabetes (2, 30, 13); **Py.** Pyrimidines (27, 50, 24); **Tr.** Triazines (60, 100, 86); **Wi.** Wisconsin Breast Cancer (32, 130, 64); **St.** Stocks Domain (9, 600, 350); **Ab.** Abalone (8, 1000, 3177).

Mean zero-one error (5 equal-length bins)					
Data	SVM	GP (MAP)	GP (EP)	HME (linear)	HME (nonlinear)
Di.	57.31±12.09%	54.23±13.78%	54.23±13.78%	51.54±6.16%	57.69±15.28%
Py.	41.46±8.49%	39.79±7.21%	36.46±6.47%	46.25±8.32%	47.71±8.16%
Tr.	54.19±1.48%	52.91±2.15%	52.62±2.66%	56.80±8.50%	55.12±4.55%
Wi.	70.78±3.73%	65.00±4.71%	65.16±4.65%	74.61±4.83%	68.36±2.91%
St.	10.81±1.70%	11.99±2.34%	12.00±2.06%	19.26±1.80%	14.43±2.16%
Ab.	21.58±0.32%	21.50±0.22%	21.56±0.36%	21.91±0.30%	21.91±0.30%
Mean zero-one error (10 equal-length bins)					
Di.	90.38±7.00%	83.46±5.73%	83.08±5.91%	76.54±7.27%	80.77±9.50%
Py.	59.37±7.63%	55.42±8.01%	54.38±7.70%	64.79±8.60%	60.83±9.21%
Tr.	67.91±3.63%	63.72±4.34%	64.01±3.78%	68.37±5.65%	69.30±4.37%
Wi.	85.86±3.78%	78.52±3.58%	78.52±3.51%	88.75±4.11%	79.53±4.53%
St.	17.79±2.23%	19.90±1.72%	19.44±1.91%	32.00±3.82%	23.87±2.24%
Ab.	44.32±1.46%	42.60±0.91%	42.27±0.46%	43.14±0.52%	42.56±1.27%

5 Conclusion and Future Work

We have described a novel Bayesian approach to ordinal regression, based on a hierarchical mixture of experts tree. The model was made analytically tractable with a Laplace approximation to the parameter posterior: future work will involve using Markov-chain Monte Carlo methods to average (integrate) predictions over the posterior distribution. The gates can equally well be impemented with Gaussian processes, a matter worthy of investigation.

Ulrich Paquet is supported by a Commonwealth Scholarship, and expresses his thanks to the Commonwealth Scholarship Commission. Andrew Naish-Guzman holds a Millennium Scholarship.

References

Chu, W. & Ghahramani, Z. (2004) Gaussian processes for ordinal regression. Technical report, Gatsby Computational Neuroscience Unit, University College London.

Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation*, *4*(5), 698–714.

Shashua, A. & Levin, A. (2003). Ranking with large margin principle: two approaches. In *Advances in Neural Information Processing Systems 15*, (pp. 937–944). MIT Press

Waterhouse, S. R., MacKay, D. J. C., & Robinson, A. J. (1996). Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems 8*, (pp. 351–357). MIT Press.