Information Fusion 14 (2013) 199–207

Contents lists available at SciVerse ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Statistical models and learning algorithms for ordinal regression problems

Takafumi Kanamori

Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan

ARTICLE INFO

Article history:
Received 17 January 2011
Received in revised form 24 May 2012
Accepted 25 May 2012
Available online 9 June 2012

Keywords:
Ordinal regression
Location models
Location-scale models
Conditional probability

ABSTRACT

In this study, we propose a learning algorithm for ordinal regression problems. In most existing learning algorithms, the threshold or location model is assumed to be the statistical model. For estimation of conditional probability of labels for a given covariate vector, we extended the location model to apply ordinal regressions. We present this learning algorithm using the squared-loss function with the location-scale models for estimating conditional probability. We prove that the estimated conditional probability satisfies the monotonicity of the distribution function. Furthermore, we have conducted numerical experiments to compare these proposed methods with existing approaches. We found that, in its ability to predict labels, our method may not have an advantage over existing approaches. However, for estimating conditional probabilities, it does outperform the learning algorithm using location models.

1. Introduction

In this paper, we study a learning algorithm for *ordinal regression* problems. Suppose that labeled training samples are observed. Then, the main task is predicting labels of unseen samples. In ordinal regressions, the labels denote preference–rankings; thus, a total order is defined over the set of labels. Usually, there are a finite number of labels in ordinal regressions. For instance, insurance companies attempt to rank their costumers on risk–low, average, or high–based on each costumer's financial status and credit history among various other factors. As a result, ordinal regressions are regarded as an intermediate problem of classification and regression. However, unlike regression problems, the quantitative differences among labels do not have the absolute meaning. For example, the quantitative differences among "good," "average," and "bad" depend on individual samples. Therefore, the mapping these qualitative labels against specific quantities such as 1, 2, and 3 may produce irrelevant results. Information in training samples only presents the frequency of labels on each covariate.

In most learning algorithms for ordinal regression problems, the so-called *location (or threshold) model* is used for statistical inference. In estimation of location model parameters, variants of online algorithms, support vector machines (SVMs), or Gaussian process models are applied. A brief introduction of this algorithm is presented in Section 2.

Although the location model is commonly used in ordinal regression problems, it is not necessarily flexible enough to represent the complex structure of probability distributions behind the training samples. The location model uses a fixed dispersion parameter of label probability. However, in the real-world data, uncertainty of labels may depend on the covariate. For example, a label will almost certainly appear at a covariate vector, while the label distribution may be closer to the uniform distribution at the other covariate. The location model cannot deal with such heteroscedasticity. Since label prediction requires only an estimation of the median of label probability, some existing algorithms can more successfully predict the labels in ordinal regression problems. However, these algorithms do not perform well in estimating label probability because of the lack of flexibility in the location model.

In this paper, we study the learning algorithms for estimating conditional probability of ordered labels, given the covariate vector. To improve existing estimating methods, we extended the location model to the *location-scale model*. This is a popular model for statistical data analysis in economics and medicine [1,2]. In the location-scale model, the *scale function* is incorporated to adjust the dispersion of the label probability. When dispersion of label probability is not homogeneous, the location-scale model is useful for fitting observed samples. In our numerical studies, we have shown that, for estimation of conditional probability, location-scale models outperform location models. On the other hand, for label prediction of the benchmark dataset, location models perform better than location-scale models. For estimation of conditional probability, we use the observed information to capture the whole structure of probability distribution; hence, the flexible model with appropriate regularization is well suited for such statistical analysis. On the other hand, for label prediction, we must only estimate the median of the label distribution. In SVM-based algorithms using the location model, analysis of training samples focuses on estimating this median. As a result, prediction accuracy is high, although these approaches fail to provide reliable estimations of label probability.

E-mail address: kanamori@is.nagoya-u.ac.jp

1566-2535/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.inffus.2012.05.006

The paper is organized as follows: In Section 2, we formulate the ordinal regression problems, and introduce some of the existing approaches. Statistical models for ordinal regressions are considered in section 3. In Section 4, we investigate loss functions for probability estimation. In Section 5–7, we propose the learning algorithm, conduct numerical studies, and present our conclusions.

Throughout the paper, we have used the following notations: Let $\mathcal{X}$ be the set of input vectors and $\mathcal{Y}=\{1,\dots,K\}$ be the set of output labels. The known order $1<2<\cdots<K$ is defined as ranking for the set of labels. The conditional probability distribution of labels $y\in\mathcal{Y}$ given $\mathbf{x}\in\mathcal{X}$ is expressed as $\Pr(Y\leqslant y|\mathbf{x})$, and let μ be a probability measure defined on $\mathcal{X}$. The training samples are denoted as $(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_n,y_n)\in X\times\mathcal{Y}$. Let $\mathcal{Y}'$ be $\{1,\dots,K-1\}=\mathcal{Y}\setminus\{K\}$ for the simplicity of notation. The indicator function $[[A]]$ is defined as

$$[[A]]=\begin{cases}1 & A \text{ is true},\\ 0 & \text{otherwise}.\end{cases} \tag{1}$$

2. Brief introduction of existing learning methods

In ordinal regression problems, the main purpose is to predict the label $y\in\mathcal{Y}$ of given input $\mathbf{x}\in\mathcal{X}$ based on the training samples $(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_n,y_n)\in\mathcal{X}\times\mathcal{Y}$. There are a number of works on ordinal regression problems [3–14]. Some approaches come from the statistical community, and the others come from the machine learning community. In this section, we introduce some of the existing methods.

A popular statistical model for label prediction in ordinal regression is called the *threshold model*. In this model, the label prediction $\hat{y}$ for $\mathbf{x}\in\mathcal{X}$ is given as:

$$\hat{y}(\mathbf{x})=\min\{y\in\mathcal{Y}|b_y\geqslant f(\mathbf{x})\}. \tag{2}$$

The thresholds $b_1,\dots,b_K\in\mathbb{R}$ should satisfy the monotonicity, $b_1\leqslant b_2\leqslant\cdots\leqslant b_{K-1}<b_K:=\infty$. The K-th threshold b_K is fixed to ∞ for convenience. We need to estimate the function $f:\mathcal{X}\to\mathbb{R}$ and the thresholds $b_1,b_2,\dots,b_{K-1}$ for label prediction. To remove a redundant degree of freedom in the threshold model, we impose a constraint, for example $f(\mathbf{x}_0)=0$ for a fixed $\mathbf{x}_0\in\mathcal{X}$.

The perceptron ranking (PRank) algorithm is an on-line algorithm using the model $f(\mathbf{x})=\langle\mathbf{w},\mathbf{x}\rangle$, where $\langle\mathbf{w},\mathbf{x}\rangle$ is the inner product of $\mathbf{w}$ and $\mathbf{x}$ [4]. The absolute loss function is defined over the set of labels. In the PRank algorithm, parameters are updated so as to reduce the accumulate loss $\sum_{t=1}^{T}|y_t-\hat{y}_t|$, where $\hat{y}_t$ is the prediction of y_t at the t-th round of the on-line process. Note that the label y is regarded as a numerical value in PRank algorithm. In summary, PRank is an on-line algorithm that minimizes empirical approximation of the expected absolute loss $E[|y-\hat{y}(\mathbf{x})|]$ with respect to the population distribution. It is widely accepted that the minimum solution for absolute loss provides the median of the random variable. Therefore, what PRank estimates is the median value of the labels under the population distribution.

In some recent works on ordinal regression problems, the support vector machine (SVM) [15,16] has been applied. SVM is one of the most popular learning methods for binary classification problems. The purpose of binary classification is prediction of the output binary label for a given input $\mathbf{x}$ based on the training samples $(\mathbf{x}_1,z_1),\dots,(\mathbf{x}_n,z_n)\in\mathcal{X}\times\{-1,1\}$. SVM with the linear kernel provides the estimator in the form of $b-\langle\mathbf{w},\mathbf{x}\rangle$. Prediction of a binary label is given by the sign of $b-\langle\mathbf{w},\mathbf{x}\rangle$ for the given $\mathbf{x}$, i.e., $\hat{y}(\mathbf{x})=2[[b\geqslant\langle\mathbf{w},\mathbf{x}\rangle]]-1$. The estimator is given as the optimal solution of

$$\min_{\mathbf{w},b}\ \frac{1}{2}\mathbf{w}^{\top}\mathbf{w}+C\sum_{i=1}^{n}\ell_{\mathrm{h}}(z_i(b-\langle\mathbf{w},\mathbf{x}_i\rangle)), \tag{3}$$

where $\ell_{\mathrm{h}}(z)=\max\{0,1-z\}$ is the hinge loss function and C is a positive constant adjusting the trade-off between regularization and fitting terms. Let $\Pr(Z=1|\mathbf{x})$ be the conditional probability of the binary label. Suppose that the decision boundary $\{\mathbf{x}|\Pr(Z=1|\mathbf{x})=1/2\}\subset\mathcal{X}$ is realized by the specified model. Then, as the sample size tends to infinity, the estimator will satisfy the following,

$$\Pr(Z=1|\mathbf{x})>1/2\Rightarrow b>\langle\mathbf{w},\mathbf{x}_i\rangle,\ \ \Pr(Z=1|\mathbf{x})<1/2\Rightarrow b<\langle\mathbf{w},\mathbf{x}_i\rangle, \tag{4}$$

with high probability.

In ordinal regressions, we need to estimate the function $f(\mathbf{x})$ and thresholds $b_y,\ y\in\mathcal{Y}'$. [7] proposed SVM for ordinal regression problems, and elucidated their method from the perspective of binary classification problems by introducing the concept referred to as *label reduction* [9]. Let $y^{(k)}$ be the reduced label of $y\in\mathcal{Y}$ at $k\in\mathcal{Y}'$ defined as

$$y^{(k)}=2[[y\leqslant k]]-1=\begin{cases}1 & y\leqslant k\\ -1 & y>k\end{cases}. \tag{5}$$

The reduced label is obtained by splitting the label at order k. The reduced label of the training sample y_i is denoted as $y_i^{(k)}$. For a fixed $k\in\mathcal{Y}'$, the loss function to estimate the decision boundary of $\Pr(y^{(k)}=1|\mathbf{x})$ is given by (3) with $z_i=y_i^{(k)}$. In the threshold model, there are $K-1$ thresholds $b_1,\dots,b_{K-1}$ to be estimated. Thus, the loss function summed up over $k\in\mathcal{Y}'$ is used for the parameter estimation:

$$\frac{1}{2}\mathbf{w}^{\top}\mathbf{w}+C\sum_{k\in\mathcal{Y}'}\sum_{i=1}^{n}\ell_{\mathrm{h}}(y_i^{(k)}(b_k-\langle\mathbf{x}_i,\mathbf{w}\rangle)). \tag{6}$$

As a statistical model of $f(\mathbf{x})$, a reproducing kernel Hilbert space (RKHS) is frequently used to fit the regression function to a complicated data structure [16]. When sample size tends to infinity, the sign of $b_k-f(\mathbf{x})$ for given $\mathbf{x}$ is expected to be the sign of $\Pr(Y\leqslant k|\mathbf{x})-1/2$, since the equality $\Pr(Y^{(k)}=1|\mathbf{x})=\Pr(Y\leqslant k|\mathbf{x})$ holds. As a result, the prediction (2) will provide an approximate value of label $\hat{y}(\mathbf{x})$ satisfying the following:

$$\Pr(Y\leqslant\hat{y}(\mathbf{x})-1|\mathbf{x})<\frac{1}{2}\leqslant\Pr(Y\leqslant\hat{y}(\mathbf{x})|\mathbf{x}). \tag{7}$$

Therefore, the predicted label provides an approximation of the median value of the labels. In a strict sense, the above expression will not hold even in the limit of sample size, since the loss function (6) leads estimation bias in general. Intuitively, however, in both PRank and SVM-based learning methods, the prediction target is considered to be the median of the labels under the population distribution.

[9] proposed a general framework for ordinal regressions with the reduction technique. [8] studied "order learning,", proposing a mapping from label to binary, which is the same method as that used by [9]. In their algorithms, an extended SVM is applied for label prediction under the threshold models.

3. Statistical models for ordinal regression problems

We shall now compare the two statistical models for ordinal regression problems: the location models and the location-scale models.

3.1. Location models

Let $F:\mathbb{R}\to[0,1]$ be a distribution function on $\mathbb{R}$ satisfying $F(0)=1/2$, and Z be a random variable obeying $F(z)$. Then, $f(\mathbf{x})+Z$ has the distribution function $\Pr(f(\mathbf{x})+Z\leqslant z)=F(z-f(\mathbf{x}))$. For given thresholds $b_1\leqslant b_2\leqslant\cdots\leqslant b_{K-1}<b_K=\infty$, the random variable Y is defined as

$$Y=\min\{y\in\mathcal{Y}|b_y\geqslant f(\mathbf{x})+Z\}. \tag{8}$$