Ordered Probit Model with an Artificial Neural Network for Predicting Potential Consumer Ratings on Amazon Software Reviews

Jikhan Jeong, PhD

Department of Economics, Missouri University of Science and Technology

Version: Nov 10, 2024 (under major revision for resubmission to the journal)

Abstract

This paper introduces a novel approach for ordinal classification – an ordered probit model with an artificial neural network (OPANN) – which combines the classical ordered probit model framework with an artificial neural network (ANN). The critical difference lies in the ANN within the OPANN, which approximates a nonlinear function for the unobservable continuous dependent variable. This approach considers nonlinear interactions among variables and mitigates potential misspecification errors. This unique feature distinguishes the OPANN from the traditional ordered probit model (i.e., the benchmark model). In order to compare the predictive performance of the OPANN and the ordered probit model, this paper presents counterfactual predictive experiments to predict potential consumer ratings for Amazon software products without post-purchase information. Specifically, star ratings in online product reviews are ordinal variables that reflect the strength of consumer preferences. By accurately predicting potential consumer star ratings, a seller on a digital platform can better understand who is more likely to be satisfied with their product, thereby improving business decisions. In three predictive experiments, the OPANN consistently outperforms the classical ordered probit model, demonstrating its superior predictive power. The more uneven the class distribution in the dataset, the more significant the predictive performance gap between the OPANN and the ordered probit model. Additionally, the classical order probit model does not present predictive performance for the minority classes in three predictive experiments. The OPANN framework and approach can be used to predict star ratings in different products on different platforms and extended to predict ordinal responses in different domains.

Introduction

This paper proposes an ordered probit model with an artificial neural network (OPANN) for ordinal classification. The critical difference lies in the ANN within the OPANN, which approximates a nonlinear function for the unobservable continuous dependent variable. This approach considers nonlinear interactions among variables and mitigates potential misspecification errors. This unique feature distinguishes the OPANN from the traditional ordered probit model.

To the best of the author's knowledge, this paper is the first to combine the ordered probit model framework with an ANN for ordinal classification, highlighting the novelty and significance of the OPANN approach.

Methods

While the ordered probit model assumes a linear functional form for the unobservable continuous dependent variable y_i^*, the OPANN assumes a nonlinear functional form for y_i^* as follows:

$$y_i^* = f(x_i, \beta) + e_i, \text{ where } e_i \sim \text{i.i.d.} N(0,1) \text{ and } i = 1, \ldots, n$$

The above nonlinear function $f(x_i, \beta)$ is approximated by a standard ANN in the OPANN. Additionally, an ANN can capture the nonlinear relationships and interactions among independent variables.

The ANN in the OPANN has four layers: input, hidden, output, and log-likelihood function layers. For example, Fig. 1 shows the structure of an ANN with three hidden nodes in the hidden layer for the OPANN.

The hidden layer contains Q hidden nodes. Each node in the hidden layer receives linearly weighted independent variables plus a constant term from the input layer. These linearly weighted independent variables are then entered into the continuous nonlinear activation function σ(·) in each hidden node as follows:

$$h_{i,q} = \sigma(b_q + \sum_{k=1}^{K} w_{q,k}\, x_{i,q})$$

This paper uses the rectified linear unit (ReLU) as an activation function.

The fitted unobservable ordinal dependent variable $\widehat{y_i^*}$ is the linearly weighted output of the activation function in each node plus a constant term as follows:

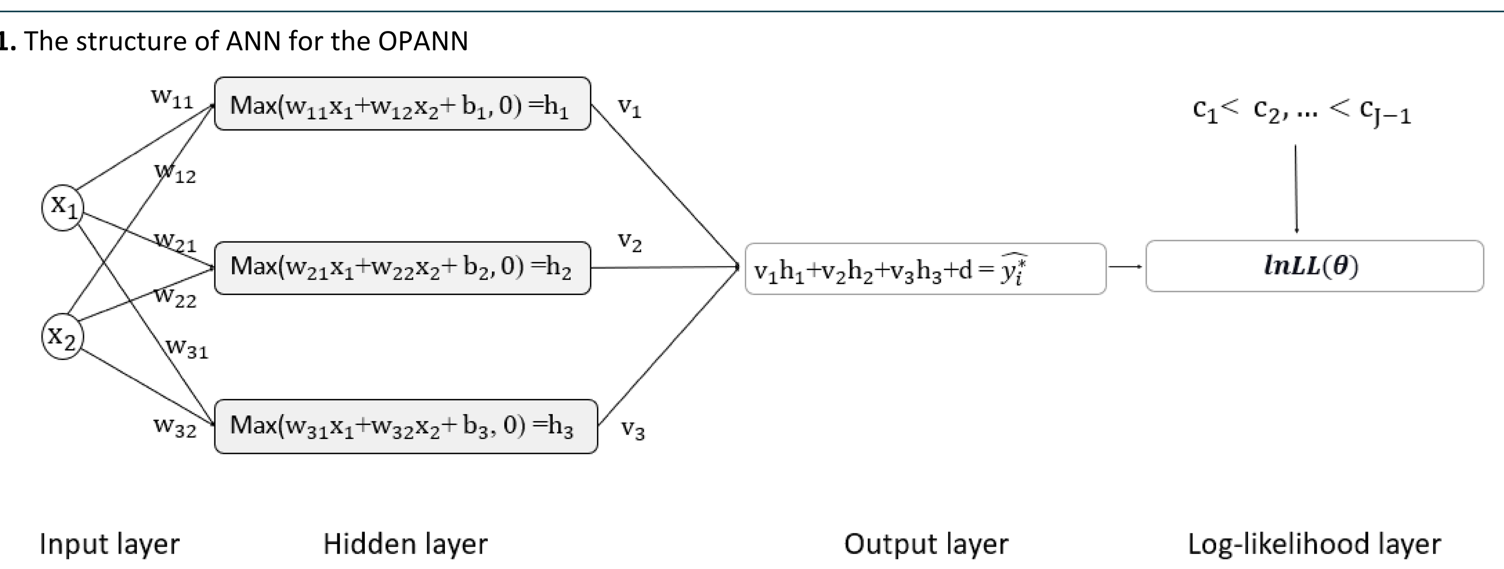
$$\widehat{y_i^*} = d + \sum_{q=1}^{Q} v_q\, h_{i,q}$$

The probability of the observable ordinal dependent variable for 'i' in the OPANN model is:

$$P(y_i = 1|x_i) = P(-\infty < y_i^* \leq c_1|x_i) = \Phi(c_1 - x_i\beta)$$
$$P(y_i = 2|x_i) = P(c_1 < y_i^* \leq c_2|x_i) = \Phi(c_2 - \widehat{y_i^*}) - \Phi(c_1 - \widehat{y_i^*})$$
$$\vdots$$
$$P(y_i = J|x_i) = P(c_{J-1} < y_i^* \leq \infty|x_i) = 1 - \Phi(c_{J-1} - \widehat{y_i^*})$$

where Φ denotes the cumulative distribution function for a standard normal distribution.

Figure 1. The structure of ANN for the OPANN



The log-likelihood (LL) function takes the fitted unobservable ordinal dependent variable $\widehat{y_i^*}$ from the output layer and initial J-1 cutting points. The cutting points follow a strictly ascending order. The LL function for the n sample is:

$$lnLL(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{J} I(y_i = j)\, lnP(y_i = j|x_i)$$

The weights and cutting points are estimated by backpropagation. The stochastic grading descent (SGD) method is used to estimate the unknown parameters while minimizing the negative log-likelihood. Initial values of cut points are generated from standard normal distribution. To mitigate potential grading explosion problems during optimization, the fitted unobservable ordinal dependent variable $\widehat{y_i^*}$ and cut points are bound to a two-sided confidence interval of 99.9%. The predicted observed ordinal dependent variable $\widehat{y_i}$ is determined based on the ordered probit model framework. The PyTorch is primarily used for computation.

Empirical Application

This paper's Amazon software product review dataset is secondary data (Ni et al. 2019). The modified total dataset comprises 6,173 reviews for 929 software in 328 brands, collected from 6,173 reviewers from January 1, 2018, to September 26, 2018, through data mining from the original secondary data (459,436 reviews). The study assumes that post-purchase information is unavailable to a seller who wants to predict potential consumers' star ratings. Table 1 describes key variables.

Table 1. Description of variables (* Note: t_i denotes a day when a consumer 'i' searches for software)

Variable	Description
Overall	Reviewer i's star rating for the software p at t_i* (dependent variable)
Price	Price for software
Asin_n_rev	Number of other reviewers' prior reviews of the software p before t_i
Asin_n_reviewers	Number of other reviewers who wrote a review for the software p before t_i
Asin_verified_share	Share of purchase-verified reviews in the prior reviews for the software p before t_i
User_n_rev	Reviewer i's number of prior reviews before t_i
User_reviewed_asin	Reviewer i's number of reviewed software before t_i
Day	Day of the week at t_i for the fixed effect. Saturday is the base group.

The paper demonstrates three different ordinal classification experiments to evaluate the predictive performance of the OPANN in various degrees of imbalanced class distributions: a five-star rating, three-class (positive with five- and four-star ratings, neutral with three-star ratings, and negative with two- and one-star ratings), and modified three-class (positive with five-star ratings, mild with four-, three-, and two-star ratings, and negative with one-star ratings).

Table 2. Class distribution in the five-star rating classification

Star-ratings	Total Data	Total Train	Train	Valid	Test
1	1650 (26.729%)	1338 (26.905%)	1027 (27.148%)	311 (26.134%)	312 (26.00%)
2	349 (5.654%)	283 (5.691%)	216 (5.710%)	67 (5.630%)	66 (5.500%)
3	436 (7.063%)	360 (7.239%)	266 (7.031%)	94 (7.899%)	76 (6.333%)
4	675 (10.935%)	560 (11.261%)	432 (11.420%)	128 (10.756%)	115 (9.583%)
5	3063 (49.619%)	2432 (48.904%)	1842 (48.692%)	590 (49.580%)	631 (52.583%)
Total	6173	4973	3783	1190	1200
Period	2018-01-01 2018-09-26	2018-01-01 2018-06-12	2018-01-01 2018-04-24	2018-04-25 2018-06-12	2018-06-13 2018-09-26

The predictive performance of the OPANN outperforms that of the classical ordered probit model in counterfactual predictive experiments. The predictive performance gap between the OPANN and the ordered probit model is larger when the dataset is less imbalanced.

Table 3. Prediction results for the **five-star rating** classification in macro-average (MA)

	Hyperparameters	Accuracy	MA Precision	MA Recall	MA F1 Score
OPANN	Hidden dim 12 Epoch 400 Learning rat 0.00001	0.3175	0.2072	0.1981	**0.1593**
Ordered Probit		0.5300	0.2093	0.2068	0.1564

Table 4. Prediction results for the **three-class** classification in macro-average (MA)

	Hyperparameters	Accuracy	MA Precision	MA Recall	MA F1 Score
OPANN	Hidden dim 12 Epoch 300 Learning rate 0.00001	0.6008	0.4922	0.3405	**0.3027**
Ordered Probit		0.6275	0.4381	0.3412	0.2749

Table 5. Prediction results for the modified three-class classification in macro-average (MA)

	Hyperparameters	Accuracy	MA Precision	MA Recall	MA F1 Score
OPANN	Hidden dim 20 Epoch 400 Learning rate 0.00001	0.4883	0.3236	0.3450	**0.3083**
Ordered Probit		0.5283	0.3374	0.3425	0.2567

Conclusions

The OPANN demonstrates better predictive performance in three classification experiments than the ordered probit model. The classical ordered probit model does not show predictive performance for the minority class.

The OPANN can be applied to predict star ratings in online reviews for different product groups across various digital platforms, as well as ordinal responses in other domains such as credit ratings and stock price predictions. Additionally, the OPANN framework can be extended to data-driven discrete choice modeling and classification using structured and unstructured data (e.g., text, image, and video), given deep learning's success in classifying unstructured data.

References

1. Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 188-197).