Ordinal Regression

Rahul Ramachandran
cs21btech11049

Rishit D
cs21btech11053

October 8, 2023

Contents

1 Ordinal Regression: Summary

Ordinal random variables are based on classfifcation random variables but with a stochastic ordering. For instance, support for various political parties is considered a classfication problem whereas support for various parts of the political spectrum is an ordinal regression problem. Consider k categories which are the qualitative representation of the ordering. We would want to determine the probabilities of landing in various categories based on a set of independent features.
Assume the independent variables to be represented through $\mathbf{x}$ with the probability of being in category i as $\pi_i(\mathbf{x})$. Now define cumulative probabilities $\gamma_i(\mathbf{x})$ as follows:

$$\gamma_i(\mathbf{x}) = \sum_{a=1}^{i} \pi_a(\mathbf{x}) \tag{1}$$

Our ordinal regression model is based on the General Linearized Model, represented as follows:

$$g(\gamma_i(\mathbf{x})) = \theta_i - \beta^T \mathbf{x} \tag{2}$$

where θ_i represents the cut-point for category i, β represents a vector of unknown parameters, $\mathbf{x}$ represents the vector of independent features and $g(x)$ is the link function for the GLM. Note that link functions are monotone and map $(0, 1)$ to $(-\infty, \infty)$.

1.1 Examples

For instance, defining the link function $g(x)$ as the logit function would give the proportional odds model as follows:

$$ln\left(\frac{\gamma_i(\mathbf{x})}{1-\gamma_i(\mathbf{x})}\right) = \theta_i - \beta^T \mathbf{x} \tag{3}$$

Note that the ratio of two odds for two input vectors x_1 and x_2 is independent of the category i, hence the name.

Similarly, defining the link function $g(x)$ as the complementary log-log function would generate the proportional hazards model, described as follows:

$$ln(-ln(1-\gamma_i(\mathbf{x}))) = \theta_i - \beta^T\mathbf{x} \tag{4}$$

Here, defining the survivor function $S(t;x)$ as $1-\gamma_i(\mathbf{x})$ would ensure that the ratio of the survivor functions remains constant and is independent of the category, giving us the proportional hazard model.

1.2 Differences with Multiclass Classification

Consider a multinomial logistic regression problem where the probability of sample x residing in category i, captured by the output variable Y.

$$\pi_i(\mathbf{x}) = p(Y=i|\mathbf{x}) = \frac{e^{\beta_i^T\mathbf{x}}}{\sum_{a=1}^{k} e^{\beta_a^T\mathbf{x}}} \tag{5}$$

One may note that the $\mathbf{x} = (1, x_1, ..., x_D)^T$ here also includes 1 as its first entry and correspondingly the first term of the parameter vector $\beta = (\beta_0, \beta_1, ..., \beta_D)^T$ is the intercept term, β_0.

Firstly, the ordinal regression model uses cumulative probabilities $\gamma_i(\mathbf{x})$ rather than the probability masses $\pi_i(x)$ used in the softmax model.

Secondly, we observe the linear estimators

$$\eta_i = \beta_i^T\mathbf{x} \tag{6}$$

for each case. In the case of the softmax model, every set of parameters β_i for category i is not necessarily the same. But in case of the ordinal regression model, we have

$$\eta_i = \beta_{i0} + \beta_{i1}x_1 + ... + \beta_{iD}x_D = \theta_i - \beta_*^T\mathbf{x} \tag{7}$$

where we notice that β_{i0} serves the role of the intercept θ_i and that β_* remains constant over all categories i. Hence, we only have $D+k$ parameters as compared to $(D+1)k$ parameters in a multinomial logistic regression. Since our β_* remains constant over all categories, the linear estimators η_i form a set of parallel hyperplanes as opposed to a multinomial logistic regression where there are no such constraints placed on hyperplanes to distinguish samples. Intuitively, this is logical as ordinal regression must maintain stochastic ordering. Consider two hyperplanes intersecting in an ordinal regression setting; this would imply a switch in ordering at some point x_0 in the input space, violating the ordinality of categories.

1.3 Log Likelihood and Log Odd Differences

In both cases, given n input points $\{\mathbf{x}_i\}_{1..n}$, the likelihood function would be given as:

$$\mathcal{L} = p(Y|\beta_1, ..., \beta_k) \tag{8}$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{k} \pi_j(x_n)^{\mathbb{1}\{j=Y_i\}} \tag{9}$$

Similarly, the log likelihood l would be defined as:

$$l = ln\left(\mathcal{L}\right) \tag{10}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{j,Y_i} ln\left(\pi_j(x_n)\right) \tag{11}$$

For multiclass logistic regression, l takes the following form:

$$l = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{j,Y_i}\left(\beta_j x_i - ln\left(\sum_{j=1}^{k} e^{\beta_j^T x_i}\right)\right) \tag{12}$$

Assuming $g(x)$ is the link function for ordinal regression models and using the fact that $\pi_j(\mathbf{x}) = \gamma_j(\mathbf{x}) - \gamma_{j-1}(\mathbf{x})$, we get the following expression for l:

$$l = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{j,Y_i}\left(ln(g^{-1}(\theta_j - \beta_*^T(x_i)) - g^{-1}(\theta_{j-1} - \beta_*^T(x_i)))\right) \tag{13}$$

When using proportional odds model, the above expression reduces to:

$$l = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{j,Y_i}\left(ln\left(\frac{e^{\theta_j}}{e^{\theta_j} + e^{\beta_*^T x_i}} - \frac{e^{\theta_{j-1}}}{e^{\theta_{j-1}} + e^{\beta_*^T x_i}}\right)\right) \tag{14}$$

In case of multiclass classfication, our log-odds are defined as follows:

$$ln\left(\frac{\pi_i(\mathbf{x})}{1-\pi_i(\mathbf{x})}\right) = \beta_i^T\mathbf{x} - \ln\left(\sum_j e^{\beta_j^T\mathbf{x}}\right) \tag{15}$$

But in the case of ordinal regression (in this case, specifically proportional odds), we define log-odds in terms of cumulative probabilities $\gamma_i(\mathbf{x})$ as:

$$ln\left(\frac{\gamma_i(\mathbf{x})}{1-\gamma_i(\mathbf{x})}\right) = \ln\left(\frac{g^{-1}(\theta_j - \beta_*^T x_i)}{1 - g^{-1}(\theta_j - \beta_*^T x_i)}\right) \tag{16}$$

$$= \theta_j - \beta_*^T x_i \tag{17}$$

2 Likelihood & Parameter Estimation

In all of our following derivations, we consider the proportional-odds model. Consider a single input vector $\mathbf{x}$ and its associated output $Y = (n_1, n_2, ..., n_k)^T$. Define R_i and Z_i as following:

$$R_i = \sum_{j=1}^{i} n_i \tag{18}$$

$$Z_i = \frac{R_i}{N} \tag{19}$$

where $N = \sum_{j=1}^{k} n_i$.

We interpret our partial likelihood for ordinal regression as follows (we drop the $\mathbf{x}$ for convenience):

$$\mathcal{L}(\mathbf{x}) = \prod_{i=1}^{k} \pi_i(\mathbf{x})_i^n \tag{20}$$

$$= \gamma_1(\mathbf{x})^{n_1} \prod_{i=2}^{k} (\gamma_i(\mathbf{x}) - \gamma_{i-1}(\mathbf{x}))^{n_i} \tag{21}$$

$$= \gamma_1(\mathbf{x})^{R_1} \prod_{i=2}^{k} (\gamma_i(\mathbf{x}) - \gamma_{i-1}(\mathbf{x}))^{R_i - R_{i-1}} \tag{22}$$

$$= \left(\prod_{i=2}^{k} \frac{\gamma_{i-1}^{R_{i-1}(\mathbf{x})} \times (\gamma_i(\mathbf{x}) - \gamma_{i-1}^{R_i - R_{i-1}}(\mathbf{x}))}{\gamma_i^{R_i}(\mathbf{x})} \right) \gamma_k^{R_k}(\mathbf{x}) \tag{23}$$

$$= \prod_{i=2}^{k} \frac{\gamma_{i-1}^{R_{i-1}(\mathbf{x})} \times (\gamma_i(\mathbf{x}) - \gamma_{i-1}^{R_i - R_{i-1}}(\mathbf{x}))}{\gamma_i^{R_i}(\mathbf{x})} \tag{24}$$

$$= \prod_{i=2}^{k} \left(\frac{\gamma_{i-1}}{\gamma_i} \right)^{R_{i-1}} \left(\frac{\gamma_i - \gamma_{i-1}}{\gamma_i} \right)^{R_i - R_{i-1}} \tag{25}$$

$$= \prod_{i=2}^{k} \left(\frac{\gamma_{i-1}}{\gamma_i - \gamma_{i-1}} \right)^{R_{i-1}} \left(\frac{\gamma_i - \gamma_{i-1}}{\gamma_i} \right)^{R_i} \tag{26}$$

Our partial log-likelihood would be defined as:

$$l(\mathbf{x}) = \sum_{i=2}^{k} R_{i-1} ln \left(\frac{\gamma_{i-1}}{\gamma_i - \gamma_{i-1}} \right) + R_i ln \left(\frac{\gamma_i - \gamma_{i-1}}{\gamma_i} \right) \tag{27}$$

$$= \sum_{i=2}^{k} R_{i-1} \phi_{i-1} - R_i f(\phi_{i-1}) \tag{28}$$

$$= N \left(\sum_{i=2}^{k} Z_{i-1} \phi_{i-1} - Z_i f(\phi_{i-1}) \right) \tag{29}$$

$$\tag{30}$$

where we define:

$$\phi_i = ln \left(\frac{\gamma_i}{\gamma_{i+1} - \gamma_i} \right) \tag{31}$$

$$f(t) = ln(1 + e^t) \tag{32}$$

Consider the gradient of $l(\mathbf{x})$ with respect to β_* which can be computed as follows:

$$\nabla_{\beta_*} l(\mathbf{x}) = N\left(\sum_{i=2}^{k} Z_{i-1}\nabla_{\beta_*}\phi_{i-1} - Z_i \nabla_{\beta_*} f(\phi_{i-1})\right) \tag{33}$$

$$= N\left(\sum_{i=2}^{k} \left(Z_{i-1} - Z_i f'(\phi_{i-1})\right)\nabla_{\beta_*}\phi_{i-1}\right) \tag{34}$$

$$= N\left(\sum_{i=2}^{k} \left(Z_{i-1} - Z_i \frac{\gamma_{i-1}}{\gamma_i}\right)\gamma_i \mathbf{x}\right) \tag{35}$$

$$= N\left(\sum_{i=2}^{k} \left(\gamma_i Z_{i-1} - \gamma_{i-1} Z_i\right)\mathbf{x}\right) \tag{36}$$

Assume we have P input points x_p**.** Now, our gradient of our final log-likelihood function would be:

$$\nabla_{\beta_*} l = \sum_{p=1}^{P} \nabla_{\beta_*} l(x_p) \tag{37}$$

2.1 Other Derivations

We derive $\nabla_{\beta_*}\gamma_j$ as follows:

$$ln\left(\frac{\gamma_j}{1-\gamma_j}\right) = \theta_j - \beta_*^T \mathbf{x} \tag{38}$$

$$\nabla_{\beta_*}\left(ln\left(\frac{\gamma_j}{1-\gamma_j}\right)\right) = \nabla_{\beta_*}\left(\theta_j - \beta_*^T \mathbf{x}\right) \tag{39}$$

$$\left(\frac{1-\gamma_j}{\gamma_j}\right)\left(\frac{(1-\gamma_j)-(-\gamma_j)}{(1-\gamma_j)^2}\right)\nabla_{\beta_*}\gamma_j = \mathbf{x} \tag{40}$$

$$\nabla_{\beta_*}\gamma_j = \gamma_j(1-\gamma_j)\mathbf{x} \tag{41}$$

We derive $f'(t)$ as:

$$f(t) = ln(1+e^t) \tag{42}$$

$$f'(t) = \frac{e^t}{1+e^t} \tag{43}$$

Note we can derive the following partial derivatives:

$$\frac{\partial \phi_j}{\partial \gamma_j} = \frac{\gamma_{j+1}}{\gamma_j(\gamma_{j+1}-\gamma_j)} \tag{44}$$

$$\frac{\partial \phi_j}{\partial \gamma_{j+1}} = \frac{-1}{\gamma_{j+1}-\gamma_j} \tag{45}$$

Using the above we can calculate $\nabla_{\beta_*}\phi_j$ through exact differentials as follows:

$$\nabla_{\beta_*}(\phi_j) = \frac{\partial \phi_j}{\partial \gamma_j}\nabla_{\beta_*}(\gamma_j) + \frac{\partial \phi_j}{\partial \gamma_{j+1}}\nabla_{\beta_*}(\gamma_{j+1}) \tag{46}$$

$$= \gamma_{j+1}\mathbf{x} \tag{47}$$