2a) Summary:

Ordinal data:

oridanal data is a type of catogerical data that represents values with natural order or ranking. The categories have a meaningful sequence, but the intervals between the catogeries are nort necessarily well defined.

Ex:

① Education level

② Response to a statement in IITH Course feed back form (strongly agree, Agree, Neutral, disagree, Strongly disagree) (Likert scale ratings)

This paper focuses on the cases where there is a single response measured on ordinal scale.

The paper proposes a family of models by focusing on proportional odds Model & Proportional Hazards Model & observing their general form.

$$\text{link}\{\gamma_j(x)\} = \theta_j - \beta^T x, \longrightarrow (2.1)$$

where $\gamma_j(x) = P(Y \le j \mid x)$, $\theta_j = \log k_j$ & link is the logit function in proportional odds Model & Complimentary log-log function in case of Hazards

In General, any other montonically increasing functi mapping the unit interval $(0,1)$ onto $(-\infty, \infty)$ can be used as a link function. Some of the examples given in paper are inverse Cauchy function, arctan & the log-log function $(\log(-\log(\gamma_j)))$

All linear models of form (2.1) are qualitatively similar and for any given data set, the fits are often indistingui shable. Selection of an appropriate function should therefore be based primarily on ease of interpretation. Further properties of these models are discussed deeply in the paper.

All the models advocated in this paper share the property that the categories can be thought of as contiguous intervals on some continuous scale. However, its existence is not required for model interpretation. Complex covariate structures & Alternative models like log-linear & asymmetric models were discussed. Reference & discussions of this paper in other coveted scientific

papers were given for further exploration on ordinal regression.

Ordinal Regression Vs Multi Class Classification:

ordinal regression models used cumulative probability distributions, to estimate the probability of an observation belonging to or below a specific category. The coefficients in ordinal regression models represent how predictor variables influence the odds of moving to a higher category relative to a reference category.

Multi class classification models the likelihood of an observation belonging to a specific class out of a set of mutually exclusive classes. In multi class classification odds ratios are not typically used as they are in ordinal regression. Instead you typically compute & interpret class probabilities directly. The class with the highest probability is predicted as the outcome.

The likelihood function of ordinal regression (with logit link) is given as:

$$P(Y \leq j|x) - P(Y \leq j-1|x)$$

If the Y's distribution is continuous

we know that,

$$\log\left[\gamma_j(x) / 1 - \gamma_j(x)\right] = \theta_j - \beta^T x$$

where $\gamma_j(x) = P(Y \leq j | x)$

$$\frac{\gamma_j(x)}{1 - \gamma_j(x)} = e^{\theta_j - \beta^T x}$$

$$\frac{1}{\gamma_j(x)} - 1 = \frac{1}{e^{\theta_j - \beta^T x}}$$

$$\frac{1}{\gamma_j(x)} = \frac{1}{e^{\theta_j - \beta^T x}} + 1$$

$$\gamma_j(x) = \frac{e^{\theta_j - \beta^T x}}{1 + e^{\theta_j - \beta^T x}}$$

$$= \frac{1}{1 + e^{-(\theta_j - \beta^T x)}}$$

$$P(Y \leq j | x) = \sigma(\theta_j - \beta^T x)$$

(θ_j → category threshauld; $\beta^T x$ → linear combinations of x)

~~In Mult~~

~~$P(y = k | x) =$~~

In Multiclass classification using probablistic Generative models

$$P(y = k | x) = \frac{P(x | C_k) P(C_k)}{\sum_{j=1}^{k} P(x | C_j) P(C_j)}$$

$$= \frac{\exp(a_k)}{\sum_{j=1}^{k} \exp(a_j)}$$

$$a_k = \ln\{p(x|C_k)\,p(C_k)\}$$

~~Odds ra~~ In discriminant analysis :

$$p(y=k|x) = \frac{\exp(w_k^T x)}{\sum_{i=1}^{K} \exp(w_i^T x)}$$

Observation : The coefficients of x is same ~~for~~ in ordinal regression (β is same $\forall j$), but in case of multiclass classification each class has ~~same~~ its own corresponding coefficients.

odds ratio :

in ordinal regression (with logit link)

$$= \frac{p(Y \le j|x)}{1 - p(Y \le j|x)}$$

$$p(Y \le j|x) = \pi_1 + \pi_2 + \cdots \pi_j$$

where $\pi_i = p(y=i|x)$

and in case of multiclass classification the odds will be $\frac{p(y=k|x)}{1-p(y=k|x)}$.

Differences b/w Ordinal & linear regression :

1) Nature of ~~yes~~ target variable is ordinal for ordinal regression, ~~but~~ continuous & numerical for linear & any categorical type for logistic regression.

~~Apart from the~~

Ordinal regression models the cumulative probabilities of an observation falling into or below a particular category. Linear regression models the expected value of the target as a linear combination of coefficient parameters.

Ordinal regression assumes the proportional odds or parallel line assumption, which means that the effect of predictor variables on the odds of being in a higher category is consistent across all catogeries.

(2b) $\text{link}\{\gamma_j(x)\} = \theta_j - \beta^T x$.

In the proportional odds assumption, link is the logit function

$$\log\left[\frac{\gamma_j(x)}{1-\gamma_j(x)}\right] = \theta_j - \beta^T x \longrightarrow (1)$$

where $\gamma_j(x) = \pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)$

& $\pi_j(x)$ is $P(Y=j|x)$

$\beta \longrightarrow$ The parameters

From (1), we can say

$$\frac{\gamma_j(x)}{1-\gamma_j(x)} = e^{(\theta_j - \beta^T x)}$$

$$\gamma_j(x) = \frac{1}{1+e^{-(\theta_j - \beta^T x)}}$$

From definetions, we can say that

$$\pi_j(x) = \gamma_j(x) - \gamma_{j-1}(x).$$

The likelihood function for being in category -j (for n-inputs)

$$L(\beta) = \prod_{i=1}^{n} (\pi_j(x_i))^{i} \prod_{i=1}^{k} (1 - \pi_j(x_i))^{n-1}$$

when there are k catogeries $L(\beta)$ transforms into following

$$L(\beta) = \prod_{j=1}^{K} \prod_{i=1}^{n} \left[[\pi_j(x_i)]^i [1 - \pi_j(x_i)]^{n-i} \right]$$

by applying log on both sides

$$\log[L(\beta)] = \sum_{j=1}^{K} \sum_{i=1}^{n} \left[i \log[\pi_j(x_i)] + (n-i) \log[1 - \pi_j(x_i)] \right.$$

Then β substitute $\pi_j(x)$ with $\gamma_j(x) - \gamma_{j-1}(x)$

and

$$\& \quad \gamma_j(x) = \frac{1}{1 + e^{-(\theta_j - \beta^T x)}}$$

we should use newton raphson method to iteratively estimate the parameters β that gives the parameters that maximizes the log-likelihood. Update eqn for the above is

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1} \frac{\partial}{\partial \beta} \log[L(\beta)]$$