QUESTION 2

ABHAY SHANKAR K: CS21BTECH11001 AND KARTHEEK TAMMANA: CS21BTECH11028

(I) **(a)**

- The paper proposes two models for ordinal data, namely the proportional odds model and the proportional hazards model.
- The proportional odds model is a generalisation of the logistic regression model for ordinal data. Here, the cumulative odds of the response variable $Y \leq j$ are given by
$$\boldsymbol{\kappa}_j = \kappa_j \exp(-\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x})$$
with
$$\boldsymbol{\kappa}_j = \frac{\gamma_j}{1-\gamma_j}$$
and
$$\gamma_j = \sum_{i=1}^{j} \pi_j$$
where π_j is the probability of the j'th category of the response variable Y, and $\mathbf{x}$ is the covariant vector.

  The paper also defines the cumulative odds ratio $\boldsymbol{\kappa}_{j,j+1} = \frac{\boldsymbol{\kappa}_j}{\boldsymbol{\kappa}_{j+1}}$, which is the odds of the j'th category over the $j+1$'th category.
- The proportional hazards model considers a hazard function $\lambda(t) =$, which expresses the probability of failure at time t, of the form
$$\lambda(t) = \lambda_0(t) \exp\left(-\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}\right)$$

  From this, we define the Survival function $S(t) = \exp\left(\Lambda_0 \exp\left(-\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}\right) dt\right)$ with $\Lambda_0(t) = -\int_0^t \lambda_0(t)dt$ which represents the probabiltiy of surviving beyond time t.

  For discrete data, we define γ_j to be the cumulative hazard function, and can write the logarithm of the Survival function as
$$\ln\left[-\ln\left(1-\gamma_j(\mathbf{x})\right)\right] = \theta_j - \boldsymbol{\beta}^T\mathbf{x}$$
which is known as the complementary log-log transform.
- The paper proposes a generalised empirical logit transform, as a generalisation of the two models. The quantity $Z_i = \Sigma_j w_j \tilde{\lambda}_{ij}$, with weights
$$w_j \propto R_{.j}(n - R_{.j})(n_{.j} + n_{.j+1})$$
and the logit transform
$$\tilde{\lambda}_{ij} = \ln\left(\frac{R_{ij}+\frac{1}{2}}{n_i - R_{ij} + \frac{1}{2}}\right)$$
where the R terms are various cumulatives of empirical data, is called the generalised empirical logit transform for the i'th group.
- The paper also discusses
  - The properties of the two models, proposing a few alternative link functions.
  - Invariances of the models under reversal of the ordering.
  - Asymptotic properties of the two models.
  - Parameter estimation for both models.

Date: October 8, 2023.

– Application of the models to real data.

(b) Differences between Ordinal Regression, Multiclass classification and Linear Regression.

	Ordinal	Multiclass	Linear
Response variable	Ordinal	Categorical	Continuous
Link function	Logit	Softmax	Identity
Loss function	Cross-entropy	Cross-entropy	Mean squared error
Optimisation technique	GD	GD	Closed form
Categories	Categories have relative order	Classes distinct	No classes

They all have different likelihood functions also:

- Ordinal regression: Given in next part.
- Logistic Multiclass classification:

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[\pi \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right)\right]^{t_n} \left[(1-\pi)\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\right)\right]^{1-t_n}$$

- Linear Regression:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \boldsymbol{\beta}) = \mathcal{N}\left(\mathbf{t}|\mathbf{w}^{\mathbf{T}}\boldsymbol{\phi}(\mathbf{x}_n), \boldsymbol{\beta}^{-1}\right)$$

Where all the variables have the usual meaning.

In the ordinal regression models, the cumulative odds of the output variable are expressed as exponential in the weights $\boldsymbol{\beta}$ and the covariant vector $\mathbf{x}$.

In the logistic regression model, the probability distribution of the output is expressed as exponential in $\boldsymbol{\beta}^T\mathbf{x}$.

In the linear regression model, the probability distribution of the output is linear in the weights.

(II) Question: Parameter Estimation

Solution: Revising the notation from the paper, we have the probabilities of the k ordered categories of the response variable Y given by $\{\pi_1, \ldots, \pi_k\}$, as a function of the covariant vector $\mathbf{x}$, and their cumulative probabilities given by $\gamma_j = \sum_{i=1}^{j} \pi_j$. The cumulative odds are thus $\boldsymbol{\kappa}_j = \frac{\gamma_j}{1-\gamma_j}$.

We then have the likelihood function $\boldsymbol{\kappa}_j = \kappa_j \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}\right)$, which we can reframe as

$$\begin{aligned} \frac{\gamma_j}{1-\gamma_j} &= \exp\left(\theta_j - \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}\right) \\ \implies \gamma_j &= \frac{1}{1+\exp\left(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} - \theta_j\right)} \end{aligned} \tag{1}$$

where we set $\theta_0 = 0$. We also define $R_j = \sum_{i=1}^{j} n_i$.

We have the likelihood function:

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta}) &= \prod_{j=1}^{k} \pi_j^{n_j} \\
&= \pi_1^{n_1} \prod_{j=1}^{k-1} (\gamma_{j+1} - \gamma_j)^{n_{j+1}} \\
&= \pi_1^{n_1} \prod_{j=1}^{k-1} (\gamma_{j+1} - \gamma_j)^{R_{j+1} - R_j} \\
&= \prod_{n=1}^{k-1} \left(\frac{\gamma_j}{\gamma_{j+1}}\right)^{R_j} \left(1 - \frac{\gamma_j}{\gamma_{j+1}}\right)^{R_{j+1} - R_j}
\end{aligned} \tag{2}
$$

Taking the logarithm, we have

$$
-\ln p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta}) = \sum_{j=1}^{k} \left[R_j \ln\left(\frac{1 + \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} - \theta_j\right)}{1 + \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} - \theta_{j+1}\right)}\right) - (R_{j+1} - R_j)\left(\ln\left(\frac{e^{-\theta_{j+1}} - e^{-\theta_j}}{1 + \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} - \theta_{j+1}\right)}\right) + \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}\right)\right] \tag{3}
$$

We can use gradient descent to find the optimal weights $\boldsymbol{\beta}$ or intervals $\boldsymbol{\theta} = \begin{pmatrix}\theta_1 & \dots & \theta_{k-1}\end{pmatrix}^T$. However, the paper does not provide the gradient of the likelihood function due to its complexity.

(III) Code

Refer `code/q2.ipynb`