Modified LATTE

dr.marquis.jue.hou

January 2024

1 Introduction

1.1 LATTE Loss

Sequential modeling module: communicating the visit embeddings along time through a Bi-GRU layer:

$$\boldsymbol{\Gamma}:\left\{\boldsymbol{\Phi}\left(\mathscr{D}_i\right),\boldsymbol{\psi}\left(\mathscr{D}_i\right)\right\}\mapsto\left\{\mathbf{F}\left(1,\mathscr{D}_i\right),\ldots,\mathbf{F}\left(T_i,\mathscr{D}_i\right)\right\},\mathbf{F}\left(t,\mathscr{D}_i\right)\in\mathbb{R}^q.$$

Incident predictor: separating the logistic regression transformations for incidence and silver-standard label models with link $\operatorname{expit}(x)=1/\left(1+e^{-x}\right)$, for incidence

$$G_Y\left\{\mathbf{F}\left(t,\mathscr{D}_i\right)\right\}=\operatorname{expit}\left\{\beta_{Y,0}+\boldsymbol{\beta}_Y^{\top}\mathbf{F}\left(t,\mathscr{D}_i\right)\right\},$$

for silver-standard label

$$G_S\left\{\mathbf{F}\left(t,\mathscr{D}_i\right)\right\}=\operatorname{expit}\left\{\beta_{S,0}+\boldsymbol{\beta}_S^{\top}\mathbf{F}\left(t,\mathscr{D}_i\right)\right\}.$$

Kernel-weighted losses: Incorporating Distance to Incidence. The outcomes around a given time t may provide useful modality information for the prediction model at that time t. For cumulative incidence, the features $\mathbf{D}_i^t$ may follow different patterns depending on the distance between t and the onset time. We translate the distance factor into a kernel weighting, where

$$\begin{aligned}
w_{Y,i}^t&=w_{\min}+\exp\left\{-d_{Y,i}(t)^2/\left(2h^2\right)\right\},d_{Y,i}(t)=\min\left\{|u-t|:Y_i^u=1\right\},\\
w_{S,i}^t&=w_{\min}+\exp\left\{-d_{S,i}(t)^2/\left(2h^2\right)\right\},d_{S,i}(t)=\min\left\{|u-t|:S_i^u\geq\kappa\right\}.
\end{aligned}$$

Here, h is a bandwidth hyper-parameter, κ is a threshold for silver-standard label S_i^u above which incidence Y_i^u is likely active, and we set $\min\{\emptyset\} = +\infty$. A minimum weight of $w_{\min}$ ensures the stability of the loss. With standardized kernel weighting, we construct the supervised and unsupervised loss functions,

$$\mathcal{L}_{\mathrm{SL}}(\mathbf{\Theta}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{t=1}^{T_i} w_{Y,i}^t \left[Y_i^t \log\left\{G_Y \circ \mathbf{F}\left(t, \mathscr{D}_i\right)\right\} + (1-Y_i^t)\log\left\{1 - G_Y \circ \mathbf{F}\left(t, \mathscr{D}_i\right)\right\}\right]}{\sum_{t'=1}^{T_i} w_{Y,i}^{t'}}$$

$$\mathcal{L}_{\mathrm{UL}}(\mathbf{\Theta}) = -\frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{t=1}^{T_i} w_{S,i}^t \left[S_i^t \log\left\{G_S \circ \mathbf{F}\left(t, \mathscr{D}_i\right)\right\} + (1-S_i^t)\log\left\{1 - G_S \circ \mathbf{F}\left(t, \mathscr{D}_i\right)\right\}\right]}{\sum_{t'=1}^{T_i} w_{S,i}^{t'}}$$

1.2 Ordinal Data Loss

Assume the labels Y can take on K distinct values which we denote $1, 2, \ldots, K$ and $Z_i^t = \beta_{Y,0} + \boldsymbol{\beta}_Y^\top \mathbf{F}\left(t, \mathscr{D}_i\right)$.

1.2.1 Threshold-Based Loss Function

Introduce $K-1$ thresholds $\theta_1 < \theta_2 < \cdots < \theta_{K-1}$, and denote $\theta_0 = -\infty$ and $\theta_K = +\infty$.

Define

$$s(l; y) = \begin{cases} -1 & \text{if } l < y \\ +1 & \text{if } l \geq y \end{cases}$$

Then the Threshold-Based Loss Function is

$$\mathcal{L}_{\mathrm{Threshold}}(Z_i^t; Y_i^t) = \sum_{l=1}^{K-1} f\left(s(l; Y_i^t)\left(\theta_l - Z_i^t\right)\right)$$

where $f(\cdot)$ is some margin penalty function.

Thus

$$\mathcal{L}_{\mathrm{SL}}(\mathbf{\Theta}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{t=1}^{T_i} w_{Y,i}^t \mathcal{L}_{\mathrm{Threshold}}(Z_i^t; Y_i^t)}{\sum_{t'=1}^{T_i} w_{Y,i}^{t'}}$$

1.3 Ordistic

Introduce K means $\mu_1 < \mu_2 < \cdots < \mu_K$, and the K prior probabilities $p_1, \ldots, p_K > 0$, $\sum p_i = 1$.

$$\mathcal{L}_{\text{Ordistic}}(Z_i^t; Y_i^t) = -\log P(Y_i^t = k \mid Z_i^t) = -\log \frac{\exp\{\mu_k Z_i^t + (\pi_k - \mu_k^2/2)\}}{\sum_{j=1}^{K} \exp\{\mu_j Z_i^t + (\pi_j - \mu_j^2/2)\}}$$

$$\mathcal{L}_{\text{SL}}(\mathbf{\Theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{t=1}^{T_i} w_{Y,i}^t \mathcal{L}_{\text{Ordistic}}(Z_i^t; Y_i^t)}{\sum_{t'=1}^{T_i} w_{Y,i}^{t'}}$$

2 Simulation Data Generation

2.1 Embeddings

The data of embeddings follows a standard normal distribution.

2.2 Other Variable

Other:1 Column is generated by R function $seq(0, by = 0.5)$.

Other:2 - Other:6 Variables follow a standard normal distribution.

Other:7 - Other:10 Variables follow a Poisson distribution, whose mean value is 0.1.

2.3 Code Variable

Code Variables follow a Poisson distribution, whose mean value is 0.1.

2.4 How to generate Y

γ is matrix has 4 columns.

$$v = 0.5 \times (1.1)^{0:(other.num + code.num - 1)}$$
$$\gamma = \begin{bmatrix} 0 & v \times 0.45 & v \times 0.5 & v \times 0.55 & v \times 0.6 \end{bmatrix}$$

where $other.num$ is the number of other variables and $code.num$ is the number of code variables

1. Generate X by merging Other and Code variables and Get IX by adding an intercept column.

2. Apply exponential models to transform the linear results of IX and γ.
3. Calculate cumulative density.
4. Generate random numbers that follow the uniform distribution and compare them to the cumulative density to get Y.