Early Prediction of Sepsis from Clinical Data

Elena Wang, Master in Statistical Science;
Dr. Ricardo Henao; Dr. Eric Laber

Health Data Science (HDS) Fall 2022 Student Research Program

Introduction/Background

- Sepsis is nightmare! More than six million people die of sepsis annually, and U.S. hospitals spend more than 24 billion dollars on sepsis each year (13% of U.S. healthcare expenses).
- GOOD NEWS: Deaths are preventable!
- Goal: Investigating automated algorithms to detect sepsis labels more precisely by 2 major methods for over 70% incomplete observations:

  i) Logistic Regression w/o Missing Values

  ii) Gradient Importance Learning w/n Missing Values

Dataset Description:

- This EMR data sourced from PhysioNet of 40,336 Patients sent to ICU in 3 hospitals from 2009-2019
- Response variable: SepsisLabel

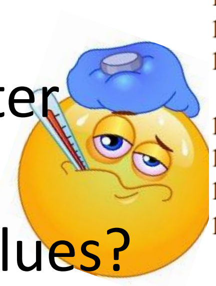
  1 if t≥tsepsis−6 and 0 if t<tsepsis−6
- Covariates: 8 vital physical signs, 26 laboratory values, and 6 demographic feature for each observation
- Below **Table1** has more details after important variables selection:

	Patients #	Sepsis Patients	Covariates #	Missingness Rate	Average Time Points/Patient
Training A	20,336	1790	36	73%	39
Testing B	20,000	1142	36	74%	38

Table 1: Summary of Dataset

Naïve Way for Brief View:

- Basic Logistic Regression after various time-series rearrangements of each patients' recordings and subsets attempting to adjust significant 2:98 imbalance problems and relative missing values imputation
- AUC could achieve≈0.65
- Question: Could we achieve better result or try more sophisticated method to execute within NA values?



Method1-Logistic Regression by Gradient Descent Update

- Built Logistic Regression with Gradient Descent update to decrease log loss in each iterations and computed a weighted cross entropy to trade off recall and precision by up- or down-weighting the cost of a positive error relative to a negative error
- Parameters for better performance: positive weight=50, batch size=12800, epochs=300, and learning rate=0.01
- Converged best results (**Figure1 & 2**): Low log loss (≈1.3)

  Relatively high AUC score (≈0.65)

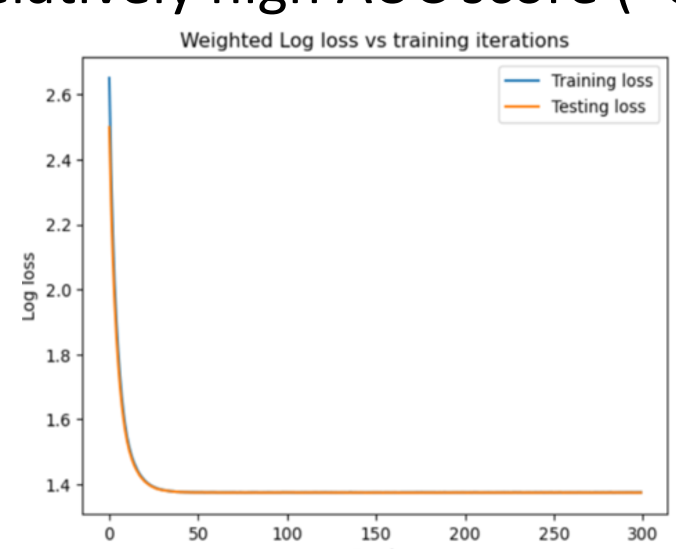


Figure1: Weighted Log Loss

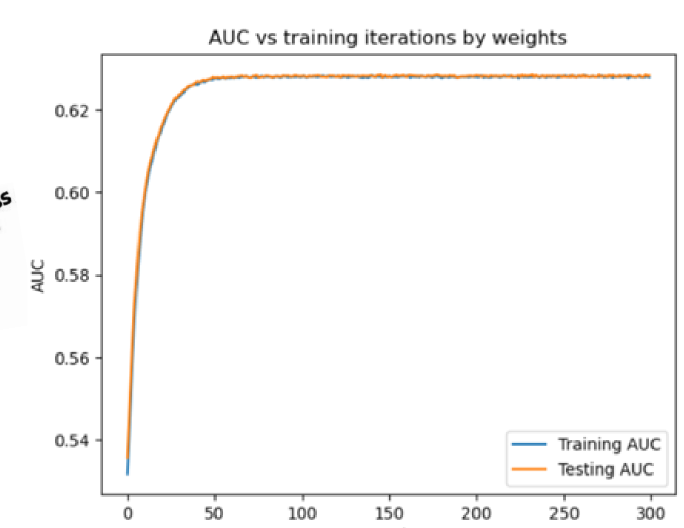


Figure2: AUC curve

Method2-Gradient Importance Learning for Incomplete Observations

Algorithm 1 Gradient Importance Learning (GIL).

Input: $\mathcal{X}, \mathcal{Y}, \mathcal{M}, \mathbf{W}_{enc}, \mathbf{W}_{inf}, \pi_\theta, Q_\nu, \alpha_\theta, \alpha_\nu, \alpha, E$

Begin:

1: Initialize $\mathbf{W}_{enc}$ and $\mathbf{W}_{inf}$, actor π_θ and critic Q_ν

2: Sample $\mathbf{x}$ from $\mathcal{X}$ and obtain the corresponding label $\mathbf{y}$ from $\mathcal{Y}$

3: Obtain the feature $\zeta \leftarrow f_{enc}(\mathbf{x}|\mathbf{W}_{enc})$ and prediction $\hat{\mathbf{y}} = f_{inf}(\zeta|\mathbf{W}_{inf})$ from the encoding and inference layers, respectively

4: $\mathbf{s} \leftarrow (\mathbf{x}, \mathbf{m}, \zeta, \hat{\mathbf{y}})$

5: **for** $iter$ in $1 : max\_iter$ **do**

6: Obtain importance from a behavioral policy $\mathbf{a} = \beta(\mathbf{s}|\pi_\theta)$

7: Train the encoding layer following $\mathbf{W}'_{enc} \leftarrow \mathbf{W}_{enc} - \alpha\boldsymbol{\Delta} \cdot (\mathbf{x}^\top \odot \mathbf{a}^\top)$ as in (4)

8: Train the inference layers following regular gradient descent, *i.e.*, $\mathbf{W}'_{inf} \leftarrow \mathbf{W}_{inf} - \alpha(\partial E / \partial \mathbf{W}_{inf})_{SGD}$

9: Obtain the prediction following the updated weights $\hat{\mathbf{y}} \leftarrow f(\mathbf{x}|\mathbf{W}'_{enc}, \mathbf{W}'_{inf})$

10: Obtain the reward $r \leftarrow R(\mathbf{s}, \mathbf{a})$

11: Get a new sample $\mathbf{x}'$ from $\mathcal{X}$ and obtain the corresponding label $\mathbf{y}'$ from $\mathcal{Y}$

12: Obtain the feature $\zeta' \leftarrow f_{enc}(\mathbf{x}'|\mathbf{W}'_{enc})$ and prediction $\hat{\mathbf{y}}' = f_{inf}(\zeta'|\mathbf{W}'_{inf})$ from the encoding and inference layers, respectively

13: $\mathbf{s}' \leftarrow (\mathbf{x}', \mathbf{m}', \zeta', \hat{\mathbf{y}}')$

14: Update the actor π_θ and critic Q_ν using $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ following (5)

15: $\mathbf{s} \leftarrow \mathbf{s}', \mathbf{W}_{enc} \leftarrow \mathbf{W}'_{enc}, \mathbf{W}_{inf} \leftarrow \mathbf{W}'_{inf}$

16: **end for**

- Gradient Importance Learning (GIL) method to train multilayer perceptrons (MLPs) and long short-term memories (LSTMs) to directly perform inference from inputs containing missing values
- Reinforcement learning (RL) to adjust gradients used to train these models back-propagation, which allows the model to exploit the underlying information behind missingness patterns
- Tabular analysis is designated: training and testing sets separated by normal (all non-sepsis) and abnormal (all sepsis)

Conclusion

- **System Satisfactory**: Good fitness of assumptions and comprehensive algorithms for Logistic Regression
- **Limitations**: 1. Missing entries leads poor performance by simply generation of estimation as the assumptions may do not satisfy the real-world applications, such as, patient time-series characteristics and a huge proportion of missingness rate or a small sample size

  2. Imputation error could limit the model capabilities
- **Future Improvements**: Simplify and tune GIL model to solve overfitting drawbacks

References

Gaoqitong. (n.d.). Gaoqitong/gradient-importance-learning: Gradient importance learning for incomplete observations, Gao et al., ICLR 2022. GitHub. Retrieved December 1, 2022, from https://github.com/gaoqitong/gradient-importance-learning

Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. D. (2019, August 5). Early prediction of sepsis from clinical data: The PHYSIONET/computing in cardiology challenge 2019. Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019 v1.0.0. Retrieved December 1, 2022, from https://physionet.org/content/challenge-2019/1.0.0/

Acknowledgements: Health Data Science at Duke is supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002553. The Duke Protected Analytics Environment (PACE) program is supported by Duke's Clinical and Translational Science Award (CTSA) grant (UL1TR001117), and by Duke University Health System. The CTSA initiative is led by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health.

Duke AI HEALTH

AI Health Poster Showcase

Duke University, December 6, 2022