Interpretable Machine Learning for Time Series Data in an ICU Setting

Gareth Booth || Supervised by Dr. Sally Shrapnel

KEY QUESTION

Can we predict death with time series ICU data in an interpretable way?

MOTIVATION

- Interpretable ML could help doctors find important relationships
- Ample publicly available ICU data to perform machine learning on
- Many of open problems, especially in interpretability

DATA

Nature of ICU Data

3D data (sequence, days, features)

Patient ID	Day since admission	Feature 1	Feature ...
0	0	20	...
	1	30	...
1	0	25	…

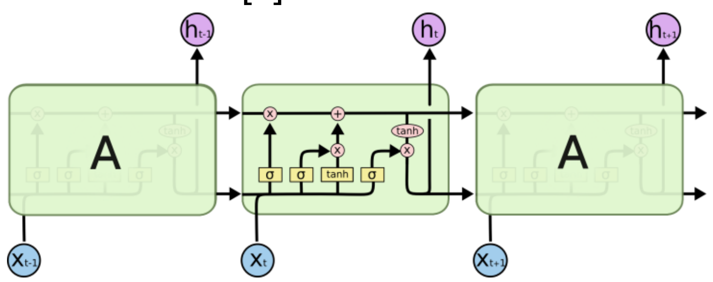
MIMIC-III [1]

- Publicly available dataset with 57,272 unique hospital admissions
- Due to the large amount of data, only use patients with complete records (I.e. no imputation)
- Final features: 3 demographics, 9 biomarkers, 4 comorbidities
- 2870 patient admissions, 15097 days of complete data, 31% deaths

MODELS

LSTM

Handles time series data by repeated application of a neural network, keeps track of state [2]



As well as LSTMs, want to use another model for a benchmark comparison

Random Forests

State of the art architecture [3], but can only handle 2D input

Suppress longitudinal input to predict:

- Death using first N days of data
- Death using last N days of data

ML RESULTS

Baseline: 69% for guessing discharge

RF using admission: 78% accuracy, f1 scores 0.64 for death.

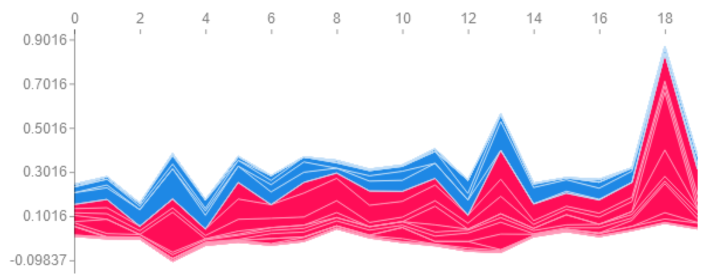
RF using final day: 84% accuracy, f1 scores 0.74 for death.

LSTM: 87% accuracy, f1 scores 0.75 for death, 0.91 for discharge.

LSTM accuracy is only 83% when using final day data, 86% for last 5 days

SHAP

- Model agnostic, local interpretability method [4]
- Uses 2D data. Well suited for RFs
- Remove sequence dimension to run LSTM in SHAP. Can plot these separate days together, see below



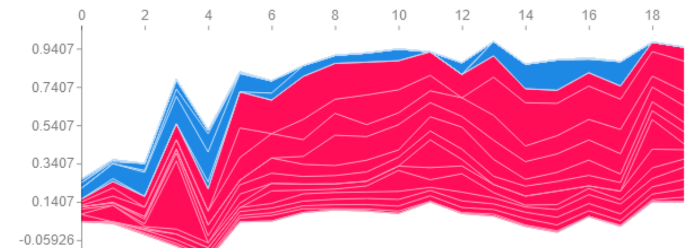
NEW SHAP

- Time series explanations for SHAP don't match the model's output!
- SHAP is good for explaining features, but what about explaining a patient's entire ICU stay?

Proposed a modified version of SHAP.

- Exploit the nature of LSTMs and their internal state
- Need to modify ML model to take and return internal state

The below shows the output from the new SHAP for the same data as above

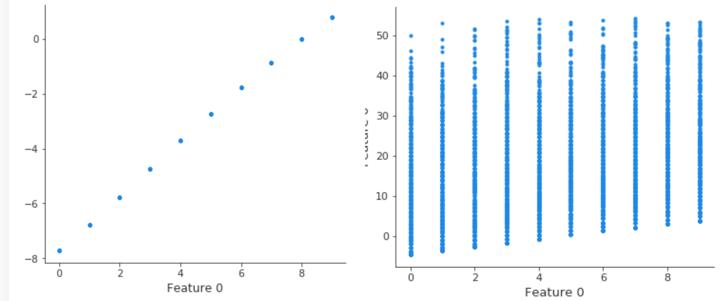


NEW SHAP RESULTS

Sanity Checks

Addition LSTM:

- Adds all numbers in a sequence
- What should the SHAP values be?



Counting LSTM:

- Counts the length of a sequence
- Do any features contribute?

Can extract 'hidden state' contribution using the SHAP scores from the previous element in the sequence.

Limitations

How does the previous state affect features in the current time step?

CONCLUSIONS

1. Modified SHAP seems promising for local explanations. Not a silver bullet
2. Lots more work in this space in the future

References
[1] A. E. Johnson, et al.,"Mimic-iii, a freely accessible critical care database,"Scientific Data,vol. 3, no. 1, 2016
[2] C. Olah. (2015, Aug) Understanding lstm networks. [Online]. Available:https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[3] S. M. Lundberg, et al. ,"Explainable AI for trees:From local explanations to globalunderstanding,"CoRR, vol. abs/1905.04610, 2019
[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems