|  |  |  |
| --- | --- | --- |
| **Fall 2024** | **Report #6 – 10/08/2024** | **Dingyi Nie** |

**——————————————————————————————————————**

**I. Task Achieved Last Week**

**——————————————————————————————————————**

* K-fold cross-validation and multiple runs on P12 to report average AUROC and deviation.
* Simple 4-feature analysis on MIMIC-III benchmark mortality prediction dataset.

**——————————————————————————————————————**

**II. Feedback and Interaction**

**——————————————————————————————————————**

* **Prof. Kuo’s Feedback**
  + Last week Prof. Kuo suggested we explore our 4 simple features analysis on more dataset and try to put everything together addressing the point that global statistics and first order dynamics matter more than fancy neural network models, write a paper and submit it to a medical journal.

**——————————————————————————————————————**

**III. Report**

**——————————————————————————————————————**

**K-fold cross-validation on P12 and multiple runs’ results**

Most previous works on the time series binary classification task of P12 mortality prediction have focused on proposing novel neural network structures and reported their results. Typically, AUROC is used as the main metric and an average AUROC along with its deviation (in a format of x±y) are reported for their models as well as each baseline. Note that how these studies define their multiple-run settings can vary.

For most neural network-based methods, since the model is stochastic and non-deterministic, most studies report multiple-run results using the exact same dataset split configuration and train a model from scratch each time. This keeps the comparison fair because all models are trained on the same training set and tested on the same evaluation set, avoiding any bias introduced by possible bad random splits that are imbalanced in data distribution.

However, since our classifier is XGBoost, which is a not-so-stochastic and mostly deterministic, multiple runs on same splits, even with different random seeds, typically yield the same or at least very similar results. K-fold cross-validation will make more sense here.

To address this gap, we either report deviation as 0 and directly compare to previously reported results of other models under the same experimental settings, or we’ll have to either train each baseline neural network model on our own again under the K-fold settings, which can be time consuming.

Anyway, I have done the K-fold experiment with P12 with k=5. The 5 runs yield this result:

86.53±0.02

**Analysis on MIMIC-III in-hospital mortality prediction task**

We run the same 4-feature analysis experiment on MIMIC-III as well. We adopt mimic3benchmark which selects a specific subset of features and a fixed split configuration, and it is used by SeFT with multiple models’ results reported.

For the whole dataset array shaped (N, L, C), for each channel in range(0, C-1), we capture 4 features:

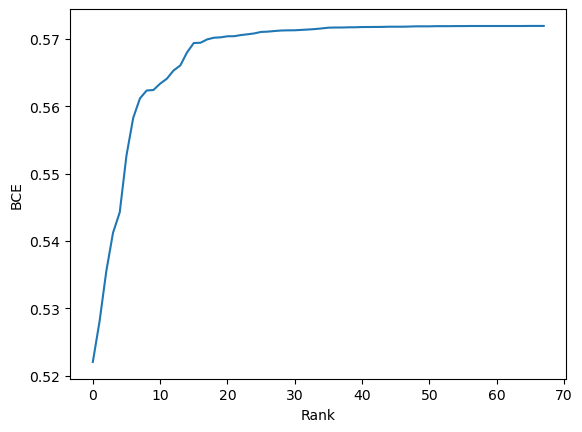
* Average of observed values
* Standard deviation of observed values
* Average of differenced observations (delta signal)
* Standard deviation of differenced observations (delta signal)

This will result in a feature array of size (N, 4C). For MIMIC-III dataset, C = 17 and N=~21,000.

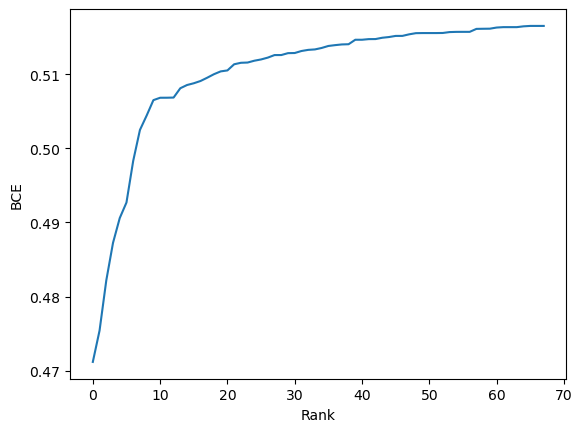
MIMIC-III benchmark defines a ~6:1 train-test ratio. If in a specific channel of an episode there are no observed value, use the global average on train set to impute the average observation. For the other 3 features, impute with 0.

By fitting an XGBoost classifier directly to the (N, 4C) training feature array, we get a 86.27% test AUROC with the best parameters searched (n\_estimators=100, max\_depth=5, learning\_rate=0.1).

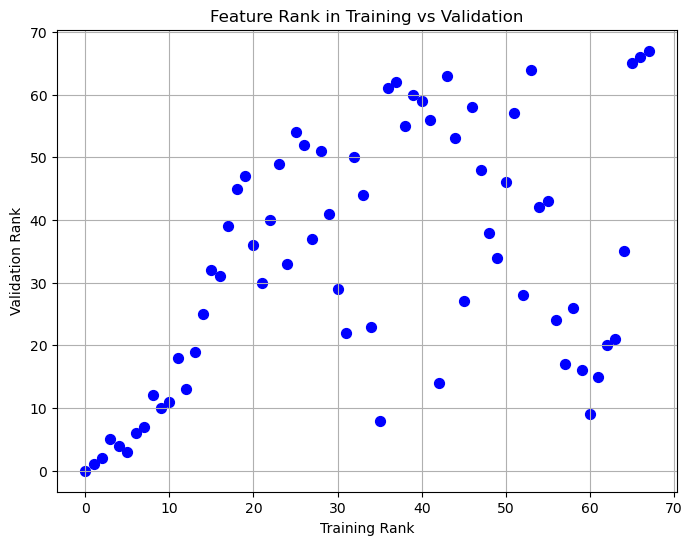
DFT rank elbow curve on train set:



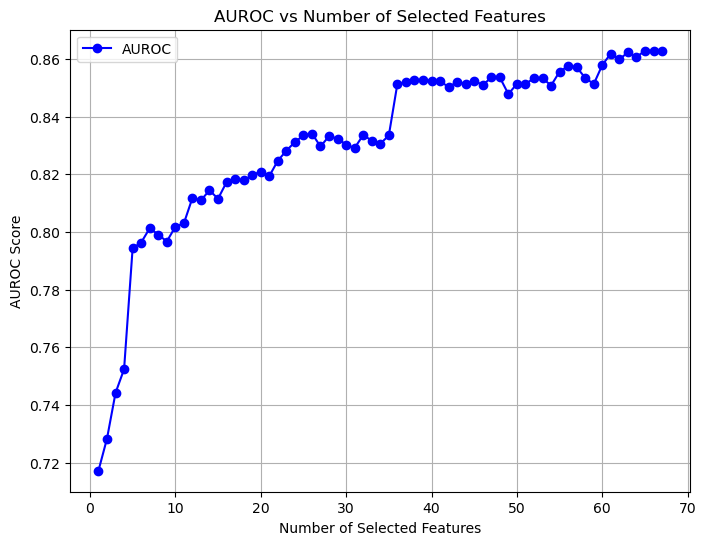
DFT elbow curve on valid set:



Joint feature rank scatter plot:



Select top-k ranked features and fit a classifier again on selected features, we can plot a curve of test AUROC vs number of selected features:



This curve shows that: with number of selected features set according to the elbow point, the performance is not maximized, which is not common. Details to be discussed.

**——————————————————————————————————————**

**IV. Next Steps**

——————————————————————————————————————

* Apply same analysis on the rest 2 common datasets: P19 and PAM.

**——————————————————————————————————————**

**V. Milestone**

——————————————————————————————————————

* Codes are available here: <https://github.com/d9sus4/GL-TS>