TABULAR PLAYGROUND SERIES APR 2022

By Johnathan Andres

PROBLEM

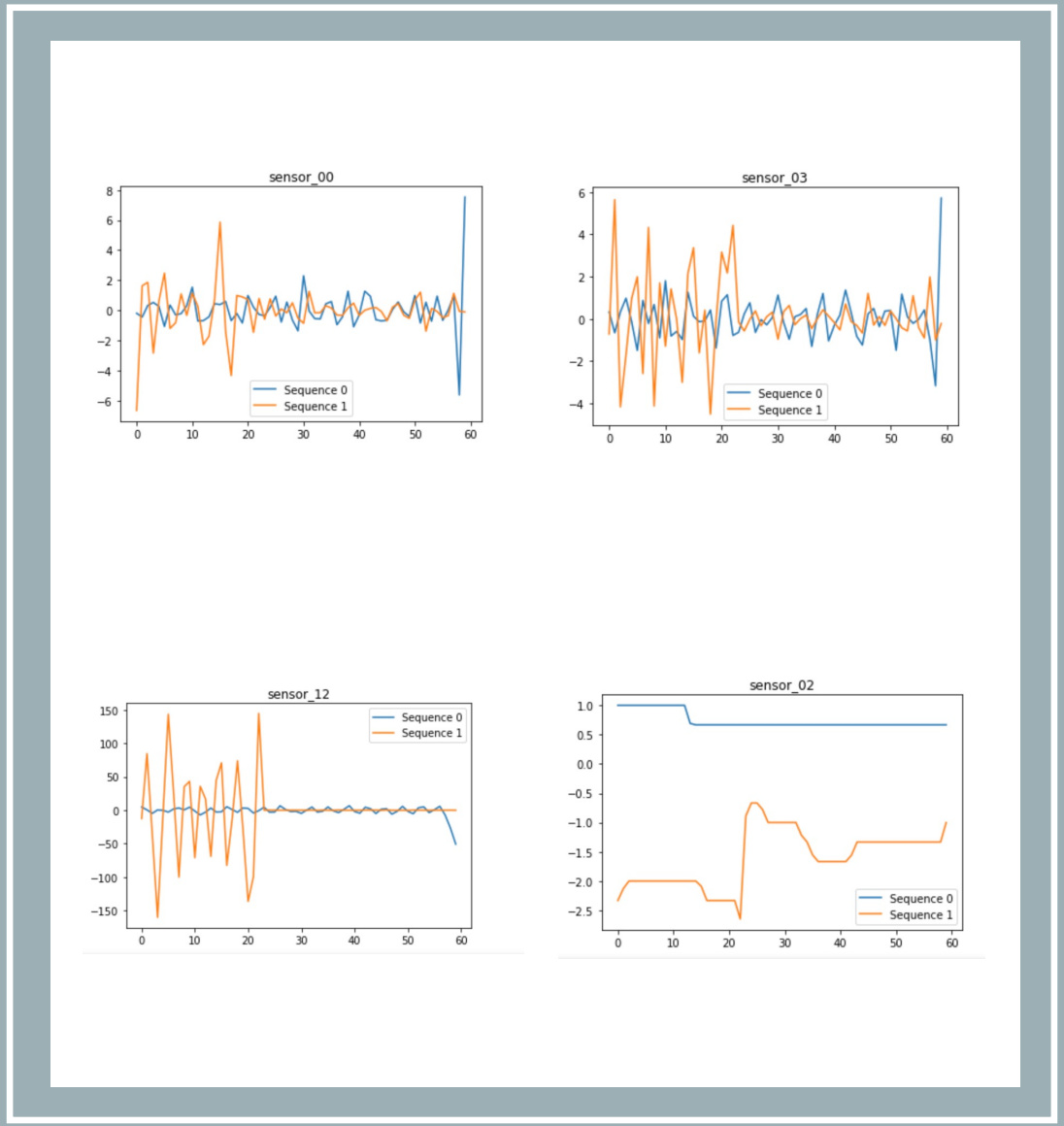
- You've been provided with thousands of sixty-second sequences of biological sensor data recorded from several hundred participants who could have been in either of two possible activity states. Can you determine what state a participant was in from the sensor data?
- Input/Feature: 12 sensor Features
- Output: 1 of two different states from state 0 or state 1.
- Metrics: Area under ROC Curve
- Data Details:
  - Training Set: ~26,000 60-second recordings of thirteen biological sensors for almost one thousand experimental participants
  - Test Set:~12,000 sequences, you should predict a value for that sequence's state. Over 700k data points.

	sequence	subject	step	sensor_00	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06	sensor_07	sensor_08	sensor_09	sensor_10
0	25968	684	0	2.427357	19.639706	1.00000	-1.466372	-1.289973	-4.207928	2.486339	-2.493893	8.0	-1.123555	-1.673048
1	25968	684	1	-4.950541	-21.747899	1.00000	0.983186	-0.569053	1.845924	-3.887978	1.727481	-2.9	0.395231	-0.882233
2	25968	684	2	1.136012	-10.756303	1.00000	1.016814	0.964157	2.454749	0.312386	1.154198	-5.6	1.114162	1.525273
3	25968	684	3	0.806028	6.504202	1.00000	-0.179646	0.969221	-1.035153	-0.457195	0.254962	-2.7	-0.588873	0.608761
4	25968	684	4	1.288253	5.552521	1.00000	-0.493805	-1.036124	-1.126402	2.008197	-0.730534	0.0	0.899566	-1.259615
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
733075	38185	773	55	0.211747	2.005252	-1.33282	0.695575	-0.161327	-1.193717	0.421676	0.869466	0.0	-1.536850	0.388101
733076	38185	773	56	-0.826121	-2.468487	-1.33282	0.381416	0.144745	1.060583	-0.765938	0.288550	0.2	-1.956647	-0.032158
733077	38185	773	57	0.755023	1.469538	-1.33282	-1.253097	-0.414802	0.007479	0.907104	-1.556489	0.4	4.341763	0.150273
733078	38185	773	58	-0.187017	0.714286	-1.33282	0.077876	1.323245	0.159312	-0.397996	0.306870	0.1	-1.013728	-0.608616
733079	38185	773	59	-0.414992	-2.858193	-1.33282	1.061062	-0.264150	-0.449514	-0.601093	1.621374	-1.0	-3.650289	-0.147107

733080 rows × 16 columns

DATA VISUALIZATION

- Here is where I want to get a feel for the data.
- These are 4 features of 2 sequences that represent different states.
- Gives a little intuition on what might be a more impactful feature as compared to others.



PRE-PROCESSING

- Training on 100 sequences. (Computing power)
- Over 720 variable given a sequence.
- Pre-Processing
  - Fit data into LSTM formatting
    - 3-D tensor of shape (samples, time step, features)

FORMULATION

- Model
  - LSTM(Long-short term memory) layer
    - Useful for longer time step data sets.
  - Dense relu layer
  - Dense sigmoid layer(Classification)
  - Loss: Binary Cross-entropy
  - Optimizer: Adam (Stochastic Gradient Descent)

RESULTS/FUTURE

- Main Result
  - Tracked Loss and AUC over epochs
  - Approximately AUC=.46
  - Train on a much lower epoch level because of set size.
- Future
  - Check model performance on Test Data Set.
  - Use a much larger training Data Set for the model.
  - Compare possibly to a model using a CNN.
  - Try other optimizers and metrics such as accuracy.