Strategies to increase trust in machine learning models Based on experiments with wearable sensor data

Sagar Sen, 30.04.2021

Outline

- Physiological signals the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Physiological signals Interpreting breathing patterns







Physiological signals Interpreting breathing patterns to predict effort



input output estimated



In one afternoon's work: 95% accuracy

Physiological signals Interpreting breathing patterns to *predict* effort



Physiological signals Interpreting breathing patterns to classify apnea







mmmmmm

Physiological signals Can I explain how these predictions happen?



Mapping to an Interpretable Domain Diagrams Natural Language Interpretable Input

Outline

- Physiological signals the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Machine learning models **Exploring explanations with LIME for TIME**



LSTM layer Historic size: 100 points Hidden neurons: 50

Architectures

- **Bi-directional LSTM**
- Variational Auto-encoder
- Transformer models
- Dense feedforward neural network



Sparse Linear Explanations in LIME



$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f,g,\pi_x) = \sum_{z,z'\in\mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \tag{2}$$

Interpretable data representations

Super-pixel - Presence or absence of a continuous patch of similar pixels essentially a tensor with three colour channels per pixel.



Binary vector indicating presence or absence of words - even though some classifiers use word embeddings such as bag of words

Prediction probabilities	atheism	(
atheism 0.58 christian 0.42	Posting 0.15 Host 0.14 NNTP 0.11 edu 0.04 have 0.01 There 0.01	

Christian Text with highlighted words From: johnchad@triton.unm.edu (jchadwic) Subject: Another request for Darwin Fish Organization: University of New Mexico, Albuquerque Lines: 11 NNTP-Posting-Host: triton.unm.edu Hello Gang, There have been some notes recently asking where to obtain the DARWIN fish. This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Raw input data in time domain with LSTM Explaining the prediction of one point



sitive	rc_t-34	0.78
2	rc_t-33	0.79
•	rc_t-32	0.80
2	rc_t-31	0.80
	rc_t-30	0.81
0	rc_t-29	0.83
2	rc_t-28	0.84
2	rc_t-27	0.85
2	rc_t-26	0.86
L	rc_t-25	0.85
2	rc_t-24	0.84
	rc_t-23	0.84
	rc_t-22	0.83
2	rc_t-21	0.82
2		0.01

Work done with intern Gutama Ibrahim this summer

Can machine learning models match our perception of reality?







With more input feature engineering Explaining prediction of mean from frequency and amplitude

EXPLANATION FOR DATA POINT : 194, ITS TRUE LABEL: 0.7047546 and predicted label 0.63825523853302

C→



Looking at mean output instead of point in a sequence

Outline

- Physiological signals the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Conclusion

- 1. Current state of the art in explainable AI works better for classification
- 2. All machine learning models had very high training and test accuracy but are the explanations palatable?
- 3. Input and output features can be modified before training such that the explanation is closer to our perception of reality. (e.g. a cylinder)
- 4. Machine learning models need to be explored such that they use meaningful input features to produce physically meaningful output features. (e.g. amplitude and frequency)
- 5. Machine learning models that use seemingly meaningless inputs to get very high accuracy need to be discarded.

A flowchart for trustworthy AI models

