Global Model Explanation for Time Series

Anonymous Author(s)
Affiliation
Address
email

Abstract

This is a novel technique to find global explanations in binary classification machine learning models by finding the salience of features. The explanation can be made on categorical, continuous, and time series data. Coefficients from a Cox Proportional Hazards regression explain the effect of variables upon the probability of an in-class response for a score output from the black box model. The analysis is conducted on a long short-term memory (LSTM) network.

1 Introduction

The method is derived using the assumption of an underlying Markov process and methods developed in the field of Survival Analysis. The stochastic counting process is uses in-class or censored observations to derive a non-parametric statistic, out-of-class observations are truncated. The stochastic process is an observation changing from out-of-class to in-class over the indexed value of the model score. The index set used is the score output from the binary classification model.

2 Background

Methods of explanations for time series black box models include visualizations of the changes of the internal state of the model over sequences of input and is conducive to what-if analysis [Strobelt et al., 2018]. Another method is learned prototypes generated from the latent space of the model [Gee et al., 2019]. Thirdly, sensitivity analysis [Tabatabaee et al., 2012] varies features over a range of values to determine the variance in predicted values by features.

3 Theoretical analysis

$X(s)$ is Markov if $P(X(s) = x|X(s_k) = x_k, X(s_{k-1}) = x_{k-1}, ..., X(s_1) = x_1) = P(X(s) = x|X(s_k) = x_k)$ for any selection of score points $s_1, ..., s_{k-1}, s_k$ such that $s_1 < ... < s_{k-1} < s_k$ and integers $x_1, ..., x_{k-1}, x_k$. The Markov property is score-homogenous when the transition probabilities only depend on the given score S [Paul and Baschnagel, 2013]. The definitions are derived from parallel survival analysis equations in [Aalen et al., 2008].

Let the model score S be a random variable with the inclusion function $I(t) = P(S > s)$. The inclusion function, the conditional probability that the response will occur with at least the score s given that the response has not received a lower score, is estimated by the product limit estimator using the multiplication rule [Kaplan and Meier, 1958] and results in the recall curve. Let $f(s)$ be the density of S. The standard definition of the hazard rate $\alpha(s)$ of S is the following with ds being infinitesimally small.

$$I(s) = P(S > s) = 1 - F(s) = \int_s^\infty f(s)ds \tag{1}$$

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Covariate	Coeff	exp(Coeff)
se(Coeff)	z	P value
MSE Ratio		
SMART 197 i	0.4998	1.6484
0.1634	3.059	0.00222
+8.1%		
SMART 242	1.3477	3.8486
0.5083	2.652	0.00801
+3.9%		

Table 1: CPH Explanation with Time Dependent Data

$$I(s) = \prod_{k=1}^{K} I(s_k|s_{k-1}) \tag{2}$$

$$\alpha(s) = \lim_{\Delta s \to 0} \frac{1}{\Delta s} P(s \leq S \leq s + \Delta s | S \geq s) = \frac{f(s)}{I(s)} \tag{3}$$

Explanations of the black box classification model can be found using the input variables as covariates in a Cox proportional hazards (CPH) regression model to explain the scores of in-class observations. The explanatory model is used to find the baseline hazard rate $\alpha_0(s)$. The effect of the covariates act multiplicatively on the baseline hazard.

$$\alpha(s|\mathbf{Z}) = \alpha_0(s) c(\beta^s Z) \tag{4}$$

$$\alpha(s|\mathbf{Z}) = \alpha_0(s) \exp(\beta^s Z) = \alpha_0(s) \exp(\sum_{k=1}^{p} \beta_k Z_k) \tag{5}$$

$$\frac{\alpha(s|\mathbf{Z})}{\alpha(s|\mathbf{Z}^*)} = \frac{\alpha_0(s) \exp(\sum_{k=1}^{p} \beta_k Z_k)}{\alpha_0(s) \exp(\sum_{k=1}^{p} \beta_k Z_k^*)} = \exp(\sum_{k=1}^{p} \beta_k (Z_k - Z_k^*)) \tag{6}$$

4 Experimental evaluation

The data chosen is time to failure data of Blackblaze hard drives. The network has three LSTM layers followed by three fully connected layers and a lookback window of 5 days of SMART statistics. The LSTM received 0.7571 accuracy, 0.9429 precision, and 0.5928 recall. Salient covariates are selected through forwards and backwards selection. The regression shows that a hard drive is 1.6484 times more likely to fail if it has a value greater than zero for SMART statistic 197 and 3.8486 times the normalized SMART 242 statistic times more likely to fail. Sensitivity analysis found the most variation in model performance by altering the values in SMART 184 End-to-End Error +23.1% MSE Ratio and SMART 7 Seek Error Rate +22.1% MSE Ratio.

5 Conclusion

The method describes a non-parametric counting process to define the cumulative probability of an in-class record occurring by a score segment through a Markov process state space model and formulates a new definition for the recall curve. Explanations are provided even when the features are time series and in order dependent models such as recurrent neural networks.

Acknowledgments

I would like to thank Blackblaze for providing the data for which without their contribution the analysis would not be possible.

References

O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar. Explaining deep classification of time-series data with learned prototypes. *arXiv preprint arXiv:1904.08935*, 2019.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

W. Paul and J. Baschnagel. *Stochastic processes*, volume 1. Springer, 2013.

H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush. S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363, 2018.

N. Tabatabaee, M. Ziyadi, and Y. Shafahi. Two-stage support vector classifier and recurrent neural network predictor for pavement performance modeling. *Journal of Infrastructure Systems*, 19(3): 266–274, 2012.