Deep learning based Feature Selection Study

Sanghyuck Lee

2021. 09 07.

Outline

- What went wrong and when? Instance-wise feature importance for time-series black-box models
- 중견 연구 과제
- 갑상선 연구 과제
- 산학 협력 과제
- 11월 특강 자료 작성

Papers for Instance-Wise Feature Selection

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise variable selection using neural networks." *International Conference on Learning Representations*. 2018.

=> ICLR

Tonekaboni, Sana, et al. "What went wrong and when? Instance-wise feature importance for time-series black-box models." *Advances in Neural Information Processing Systems* 33 (2020).

=> NeurIPS

Panda, Pranoy, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. "Instance-wise Causal Feature Selection for Model Interpretation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

=> CVPR

Liyanage, Yasitha Warahena, and Daphney–Stavroula Zois. "Optimum Feature Ordering for Dynamic Instance–Wise Joint Feature Selection and Classification." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

=> ICASSP

Paper Review

Paper Title: **What went wrong and when? Instance-wise feature importance for time-series black-box models**

Authors: Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K. Duvenaud, Anna Goldenberg

Conference: NeurIPS 2020

Abstract of the Paper

Explanations of time series models are useful for high stakes applications like healthcare but have received little attention in machine learning literature. We propose FIT, a framework that **evaluates the importance of observations** for a **multivariate time-series black-box model** by **quantifying the shift in the predictive distribution over time**. FIT **defines** the **importance of an observation** based on **its contribution to the distributional shift** under a KL-divergence that contrasts the predictive distribution against a counterfactual where the rest of the features are unobserved. We also demonstrate the need to control for time-dependent distribution shifts. We compare with state-of-the-art baselines on simulated and real-world clinical data and demonstrate that our approach is superior in **identifying important time points and observations** throughout the time series.

Introduction of the Paper

Understanding what drives **machine learning models** to **output** a particular prediction can aid **reliable decision-making** for end-users and is critical in high stakes applications like healthcare. This problem has been particularly **overlooked** in the context of **time series datasets** compared to static settings. Time series domain is unique because the data's dynamic nature results in the features driving model prediction changing over time. Explaining time-series machine learning model prediction can be defined in many ways, with existing works **primarily** considering **two approaches**. The first class uses the notions of **instance-level feature importance** proposed in the literature for static supervised learning [36, 35, 7, 21]. These either compute the gradients of the output with respect to the input vector or perturb features to evaluate their impact on model output. None of the approaches of this type explicitly model the temporal dependencies that exist in time series data. The second class uses **attention models** explicitly designed for time series. A number of works show that **attention scores** can be interpreted as an importance score for different observations over time within the context of these models [8, 34, 14].

Introduction of the Paper

Instance <-> Time Point
Observation <-> Feature

Our contributions are as follows:

1. We pose the problem of **assigning importance to observations of a time series** as that of quantifying the contribution to the predictive distributional shift under a KL-divergence by contrasting the predictive distribution against a counterfactual where the remaining features are unobserved.
2. We use **generative models to learn the dynamics of the time series.** This allows us to model the distribution of measurements, and therefore **accurately approximate the counterfactual effect of subsets of observations over time.**
3. Unlike existing approaches, FIT allows us to evaluate **the aggregate importance of subsets of features over time**, which is critical in healthcare contexts where simultaneous changes in feature subsets drive model predictions [2].

Experiments

We evaluate our feature importance assignment method (FIT) on a number of **simulated data** (where ground-truth feature importance is available) and more **complex real-world clinical tasks**. We compare with multiple groups of baseline methods, described below:

1. Perturbation-based methods
2. Gradient-based methods
3. Attention-based method
4. Others: LIME, Gradient-SHAP

Experiments

4.1 Simulated Data

Evaluating the quality of explanations is challenging due to the **lack of a gold standard/ground truth for the explanations**. Additionally, explanations are reflective of model behavior; therefore, such evaluations are tightly linked to the **reliability of the model itself**. To account for that, we evaluate the functionality and performance of our baselines in a controlled simulated environment where the ground truth for feature importance is known. A description of all datasets used is presented below:

Experiments

4.1 Simulated Data

Results: We evaluate our method using the following metrics :

1. **Ground-truth test**: We evaluate the quality of feature importance assignment across baselines compared to ground-truth using AUROC and AUPRC.
2. **Performance Deterioration test**: Additionally, we perform a Performance Deterioration Test to measure the drop in the predictive performance of the black-box model when the **most important observations are omitted**. Note that all baseline methods provide an importance score for every sample at each time point. Assuming that an important observation is more informative, removing it from the data should worsen the overall predictive performance of the black-box model. Hence a larger performance drop indicates better importance assignment. Since the data is in the form of time series and **observations are correlated** in time, **we cannot eliminate the effect of an observation simply by removing that single measurement.** Therefore, we instead **carry-forward the previous measurement**.

Experiments

4.2 Clinical Data

Explaining models based on feature and time importance is critical for clinical settings. Therefore, having verified performance quantitatively on simulated data, we test our methods on a more complex clinical data set, called MIMIC III, that consists of de-identified EHRs for ~ 40, 000 ICU patients at the Beth Israel Deaconess Medical Center, Boston, MA [17]. We use time series measurements such as vital signals and lab results for 2 evaluation tasks, with 2 distinct black-box models respectively.

1. **MIMIC Mortality Prediction**: An RNN-based prediction model that uses static patient information (age, gender, ethnicity), 8 vital and 20 lab measurements to predict mortality in the next 48 hours.
2. **MIMIC Intervention Prediction Model**: This model predicts a set of non-invasive and invasive ventilation and vasopressors using the same set of features as above. Further details of both black-box models and data can be found in Appendix B.5.

Experiments

Results: Due to the lack of ground-truth explanation for real data, we use a number of **performance deterioration** tests to evaluate the quality of explanations generated by different baseline methods. Specifically, we use the following performance deterioration tests: i) **95-percentile test:** drop observations with an importance score in the top 95-percentile of the score distribution according to each baseline. ii) **k-drop test**: remove the top k most important observations for each time series sample. Note that for k-drop test, we have selected patients who undergo significant change in health state – i.e. patients with highest change in risk of mortality (> 0.85) or a mean likelihood of intervention status (> 0.25) in the 48-hours of ICU stay. **This ensures that the top observations that are omitted are significant scores.**

Experiments Summary

Compare FS methods for $X -> \hat{Y}$ of any timeseries black-box model,

1) Simulation experiment

Measure the performance of each feature selection methods using a virtually generated GT

2) Real world data experiment

Measure the performance of each feature selection methods that deteriorated by removing the selected features.

In term of Accuracy and Time

- 사전 학습 모형을 입력으로, 각 X에 대한 중요도 점수 할당을 진행하며, **학습이나 예측에 대한 정확도 향상이나 시간 단축이 목적이 있다기 보다**는, **이미 학습 완료된 블랙박스 모형에 대한 설명력을 증진**시키고자 함.
  (구체적으로 논문에서는, 기존의 신뢰할 수 있는 RNN 모형을 선정하여 비교 실험을 진행하였고, 가장 특징을 잘 선택하는 방식으로 제안 방식을 확인함.)
- **특징 선택의 관점에서의 정확도**를 측정함. 그 방식으로는, **특징 선택이 정확할수록, 해당 모형을 더 잘 설명**한다고 가정한 후, 특징 선택의 정확도를 높임.
  (구체적으로 논문에서는, Ground-Truth 특징 중요도를 생성해서, 특징 선택의 정확도를 측정하거나, 해당 특징을 이전 시간 포인트 특징으로 대체 후, 정확도가 얼마나 떨어졌는지를 측정함.)
- 시간 관점에서의 설명은 찾지 못했는데, 마찬가지로, 새로운 모형 학습을 위한 프레임워크는 아니기 때문에, 학습 시간 및 예측 시간에서는 강한 이점을 내세우기 어려울 것으로 보임.

중견 연구 제안서 작성

- 다음 주 보고 때, 초안 회신 예정

갑상선 연구

- 중앙대병원측 세그멘테이션 검토 완료 후, Activity 실험 계획

산학 협력 프로젝트

- 특허 출원: 명세서 초안 수신 완료, 검토 예정

11월 특강 자료 작성

- 주제: '사회적 약자와 보편적 복지 향상을 위해 활용될 수 있는 AI 기술 조사'

선별주의는 **자산조사**에 의해 복지기본선 이하의 소외계층, 즉 국민기초생활보장제도상 선정대상을 주된 복지정책의 대상으로 하자는 관점

보편주의는 사회적 시민권을 지반으로 **자산조사에 관계**없이 누구에게도 차별을 제도화하지 않으며 모든 국민을 대상으로 사회복지정책 지원을 하자는 관점 (손병덕, 2020)

1. 사회적 약자를 위해 활용될 수 있는 AI 기술:
    1) AI 로봇, AI 스피커: 독거노인 등 취약 계층을 위한
    2) 교육 AI: 외국인 및 소년 소녀 가장을 위한
    3) 장애인 생활 보조 AI
2. 보편적 복지 향상을 위해 활용될 수 있는 AI 기술:
    1) 아동학대 방지를 위한 AI 시스템
    2) AI 기반 고령 친화 스마트 시티

초안 보고 예정일: 9/21 화

References

손병덕, "사회복지정책론", 학지사, 2020.

Liyanage, Yasitha Warahena, and Daphney–Stavroula Zois. "Optimum Feature Ordering for Dynamic Instance–Wise Joint Feature Selection and Classification." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

Panda, Pranoy, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. "Instance-wise Causal Feature Selection for Model Interpretation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Tonekaboni, Sana, et al. "What went wrong and when? Instance-wise feature importance for time-series black-box models." *Advances in Neural Information Processing Systems* 33 (2020).

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise variable selection using neural networks." *International Conference on Learning Representations*. 2018.