Deep learning based Feature Selection Study

Sanghyuck Lee

2021. 10. 05.

Outline

- Instanse-wise Feature Selection for Time Series Data
- 중견 연구 과제
- 갑상선 연구 과제
- 산학 협력 과제
- 보행자 데이터셋

Instanse-wise Feature Selection for Time Series Data

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

Authors: Sana Tonekaboni et al.

Conference: NeurIPS 2020

Main Algorithm

$$I(\mathbf{x}_{\mathrm{S}}, t) = \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \parallel \overset{\text{Prev}}{p(y|\mathbf{X}_{0:t-1})})}_{\text{T1: temporal distribution shift}} - \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \parallel \overset{\text{Selected Current}}{p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t})})}_{\text{T2: unexplained distribution shift}}$$

Algorithm 1 FIT

Input: f_θ**: Trained Black-box predictor model, time series** $\mathbf{X}_{0:T}$**, where** T **is the max time,** L**: Number of Monte-Carlo samples,** S**: A subset of features of interest.**

Output: Importance score matrix I

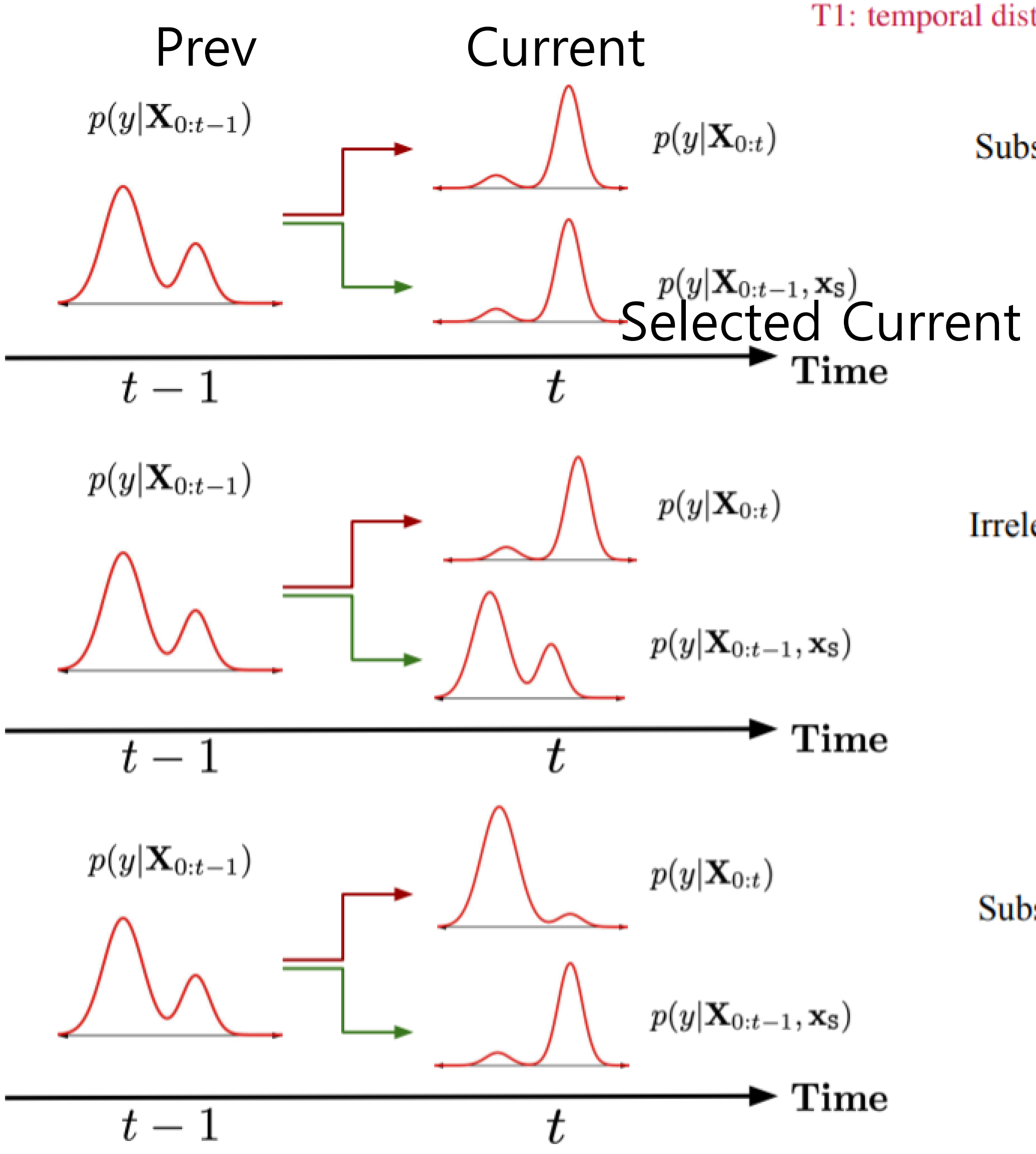
```
Train 𝒢 using X_{0:T}
for all t ∈ [T] do
    p(y|X_{0:t}) = f_θ(X_{0:t})
    p(y|X_{0:t−1}) = f_θ(X_{0:t−1})
    p(x_t|X_{0:t−1}) ≈ 𝒢(X_{0:t−1})
    for all l ∈ [L] do
        Sample x̂^{(l)}_{S^c,t} ∼ p(x_{S^c,t}|X_{0:t−1}, x_{S,t})
        p(ŷ^{(l)}) = f_θ(X_{0:t−1}, x_{S,t}, x̂^{(l)}_{S^c,t})
    p(y|X_{0:t−1}, x_{S,t}) ≈ (1/L) ∑_{l=1}^{L} p(ŷ^{(l)})
    I(x_S, t) = KL(p(y|X_{0:t}) ∥ p(y|X_{0:t−1})) − KL(p(y|X_{0:t}) ∥ p(y|X_{0:t−1}, x_{S,t}))
Return I(x_S, t)
```

Notation	Description
$[K]$ for integer K	Set of indices $[K] = \{1, 2, \ldots, K\}$.
d, t	Index for feature d in $[D]$, time step t
$-d$	Set $\{1, 2, \cdots, D\} \setminus d$
$\mathrm{S} \subseteq \{1, 2, \cdots, D\}$	Subset of observations
S^c	Set $\{1, 2, \cdots, D\} \setminus \mathrm{S}$
Data	
$x_{i,t}$	Observation i at time t.
$\mathbf{x}_{\mathrm{S},t}$	Subset of observation S at time t.
$\mathbf{x}_t \in \mathbb{R}^d$	Vector $[x_{1,t}, x_{2,t}, \cdots, x_{d,t}]$
$\mathbf{X}_{0:t} \in \mathbb{R}^{d \times t}$	Matrix $[\mathbf{x}_0; \mathbf{x}_1; \cdots; \mathbf{x}_t]$
$p(y_t|\mathbf{X}_{0:t}) \triangleq f(\mathbf{X}_{0:t})$	Outcome of the model f, at time t

Table 4: Notation used in the paper.

Main Algorithm

$$I(\mathbf{x}_S, t) = \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \parallel \overset{\text{Prev}}{p(y|\mathbf{X}_{0:t-1})})}_{\text{T1: temporal distribution shift}} - \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \parallel \overset{\text{Selected Current}}{p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{S,t})})}_{\text{T2: unexplained distribution shift}}$$



Subset explains predictive temporal shift completely:
T1 > 0, T2 $= 0$

$$I(\mathbf{x}_S, t) > 0$$

Irrelevant subset, does not explain temporal shift:
T1=T2

$$I(\mathbf{x}_S, t) = 0$$

Subset does not explain temporal shift and changes model prediction:
T1>0, T2 $>$ T1

$$I(\mathbf{x}_S, t) < 0$$

Main Algorithm

$$I(\mathbf{x}_{\mathrm{S}}, t) = \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \,\|\, \overset{\text{Prev}}{p(y|\mathbf{X}_{0:t-1})})}_{\text{T1: temporal distribution shift}} - \underbrace{\mathrm{KL}(\overset{\text{Current}}{p(y|\mathbf{X}_{0:t})} \,\|\, \overset{\text{Selected Current}}{p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t})})}_{\text{T2: unexplained distribution shift}}$$

Algorithm 1 FIT

Input: f_θ**: Trained Black-box predictor model, time series** $\mathbf{X}_{0:T}$**, where** T **is the max time,** L**: Number of Monte-Carlo samples,** S**: A subset of features of interest.**

Output: Importance score matrix I

Train $\mathcal{G}$ using $\mathbf{X}_{0:T}$
for all $t \in [T]$ **do**
    $p(y|\mathbf{X}_{0:t}) = f_\theta(\mathbf{X}_{0:t})$
    $p(y|\mathbf{X}_{0:t-1}) = f_\theta(\mathbf{X}_{0:t-1})$
    $p(\mathbf{x}_t|\mathbf{X}_{0:t-1}) \approx \mathcal{G}(\mathbf{X}_{0:t-1})$
    for all $l \in [\mathrm{L}]$ **do**
        Sample $\hat{\mathbf{x}}^{(l)}_{\mathrm{S}^c,t} \sim p(\mathbf{x}_{\mathrm{S}^c,t}|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t})$
        $p(\hat{y}^{(l)}) = f_\theta(\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t}, \hat{\mathbf{x}}^{(l)}_{\mathrm{S}^c,t})$
    $p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t}) \approx \frac{1}{L}\sum_{l=1}^{L} p(\hat{y}^{(l)})$
    $I(\mathbf{x}_{\mathrm{S}}, t) = \mathrm{KL}(p(y|\mathbf{X}_{0:t}) \,\|\, p(y|\mathbf{X}_{0:t-1})) - \mathrm{KL}(p(y|X_{0:t}) \,\|\, p(y|X_{0:t-1}, \mathbf{x}_{\mathrm{S},t}))$
Return $I(\mathbf{x}_{\mathrm{S}}, t)$

Notation	**Description**
$[K]$ for integer K	Set of indices $[K] = \{1, 2, \ldots, K\}$.
d, t	Index for feature d in $[D]$, time step t
$-d$	Set $\{1, 2, \cdots, D\} \setminus d$
$\mathrm{S} \subseteq \{1, 2, \cdots, D\}$	Subset of observations
S^c	Set $\{1, 2, \cdots, D\} \setminus \mathrm{S}$
Data	
$x_{i,t}$	Observation i at time t.
$\mathbf{x}_{\mathrm{S},t}$	Subset of observation S at time t.
$\mathbf{x}_t \in \mathbb{R}^d$	Vector $[x_{1,t}, x_{2,t}, \cdots, x_{d,t}]$
$\mathbf{X}_{0:t} \in \mathbb{R}^{d \times t}$	Matrix $[\mathbf{x}_0; \mathbf{x}_1; \cdots; \mathbf{x}_t]$
$p(y_t|\mathbf{X}_{0:t}) \triangleq f(\mathbf{X}_{0:t})$	Outcome of the model f, at time t

Table 4: Notation used in the paper.

Main Algorithm

$$I(\mathbf{x}_S, t) = \underbrace{\overbrace{\mathrm{KL}(p(y|\mathbf{X}_{0:t})}^{\text{Current}} \parallel \overbrace{p(y|\mathbf{X}_{0:t-1}))}^{\text{Prev}}}_{\text{T1: temporal distribution shift}} - \underbrace{\overbrace{\mathrm{KL}(p(y|\mathbf{X}_{0:t})}^{\text{Current}} \parallel \overbrace{p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{S,t}))}^{\text{Selected Current}}}_{\text{T2: unexplained distribution shift}}$$

$p(y|\mathbf{X}_{0:t}) = f_\theta(\mathbf{X}_{0:t})$
$p(y|\mathbf{X}_{0:t-1}) = f_\theta(\mathbf{X}_{0:t-1})$
$p(\mathbf{x}_t|\mathbf{X}_{0:t-1}) \approx \mathcal{G}(\mathbf{X}_{0:t-1})$
for all $l \in [\mathrm{L}]$ **do**
  Sample $\hat{\mathbf{x}}^{(l)}_{S^c,t} \sim p(\mathbf{x}_{S^c,t}|\mathbf{X}_{0:t-1}, \mathbf{x}_{S,t})$
  $p(\hat{y}^{(l)}) = f_\theta(\mathbf{X}_{0:t-1}, \mathbf{x}_{S,t}, \hat{\mathbf{x}}^{(l)}_{S^c,t})$
$p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{S,t}) \approx \frac{1}{L}\sum_{l=1}^{L} p(\hat{y}^{(l)})$
$I(\mathbf{x}_S, t) = \mathrm{KL}(p(y|\mathbf{X}_{0:t}) \parallel p(y|\mathbf{X}_{0:t-1})) - \mathrm{KL}(p(y|X_{0:t}) \parallel \overbrace{p(y|X_{0:t-1}, \mathbf{x}_{S,t}))}^{\text{Selected Current}}$
Return $I(\mathbf{x}_S, t)$

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

Monte Carlo integration referenced by Murphy, Kevin P.

Estimate x_t by time series generator

↓

Estimate $p(x_{S^c,t}|X_{0:t-1}, X_{S,t})$

↓

Sampling $\hat{x}^{(l)}_{S^c,t'}$ L times

	S	S_c	y
0	Use observations	Use observations	$p(\hat{y}^{(l)})$
⋮			
$t-1$			
t		Use Estimations	

f_θ

Mean

↓

$p(y|X_{0:t-1}, X_{S,t})$
Selected Current

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Main Algorithm

Algorithm 1 FIT
Input: f_θ: Trained Black-box predictor model, time series $\mathbf{X}_{0:T}$, where T is the max time, L: Number of Monte-Carlo samples, S: A subset of features of interest.
Output: Importance score matrix I

1: Train $\mathcal{G}$ using $\mathbf{X}_{0:T}$
2: **for all** $t \in [T]$ **do**
3: $\quad p(y|\mathbf{X}_{0:t}) = f_\theta(\mathbf{X}_{0:t})$
4: $\quad p(y|\mathbf{X}_{0:t-1}) = f_\theta(\mathbf{X}_{0:t-1})$
5: $\quad p(\mathbf{x}_t|\mathbf{X}_{0:t-1}) \approx \mathcal{G}(\mathbf{X}_{0:t-1})$
6: $\quad$ **for all** $l \in [\mathrm{L}]$ **do**
7: $\quad\quad$ Sample $\hat{\mathbf{x}}^{(l)}_{\mathrm{S}^c,t} \sim p(\mathbf{x}_{\mathrm{S}^c,t}|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t})$
8: $\quad\quad p(\hat{y}^{(l)}) = f_\theta(\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t}, \hat{\mathbf{x}}^{(l)}_{\mathrm{S}^c,t})$
9: $\quad p(y|\mathbf{X}_{0:t-1}, \mathbf{x}_{\mathrm{S},t}) \approx \frac{1}{L}\sum_{l=1}^{L} p(\hat{y}^{(l)})$
10: $\quad I(\mathbf{x}_{\mathrm{S}}, t) = \mathrm{KL}(p(y|\mathbf{X}_{0:t}) \,\|\, p(y|\mathbf{X}_{0:t-1})) - \mathrm{KL}(p(y|X_{0:t}) \,\|\, p(y|X_{0:t-1}, \mathbf{x}_{\mathrm{S},t}))$
11: Return $I(\mathbf{x}_{\mathrm{S}}, t)$

	S	S_c
0	$I_{0,S}$	
$\vdots$	$\vdots$	
$t-1$	$I_{t-1,S}$	
t	$I_{t,S}$	

1. For evaluating instance-level feature importance when |S| = 1. This task is **comparable to what most feature importance algorithms do**.
2. In cases where there are **pre-defined set** of features, for instance a set of blood works in the clinic, FIT can be used to **find the aggregate importance** of those features together.
3. For **finding the minimal set** of observations to acquire in a time series setting, to approximate the predictive distribution well.

Comment

Instance wise feature selection method for time series black box model

=> By using **Time series Generator, Monte Carlo, Kullback–Leibler Divergence**

=> Measure importance of **each time step** and **each feature or feature subset**

=> But, because it estimates conditional distribution **using black box model** for Kullback–Leibler divergence**,** it is **difficult** to use **without information about the relation between X and Y**.

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

중견 연구 제안서 작성

- 심미성 제안서 기반으로 재작성 예정

갑상선 연구

- 이번 주 목요일 미팅 예정

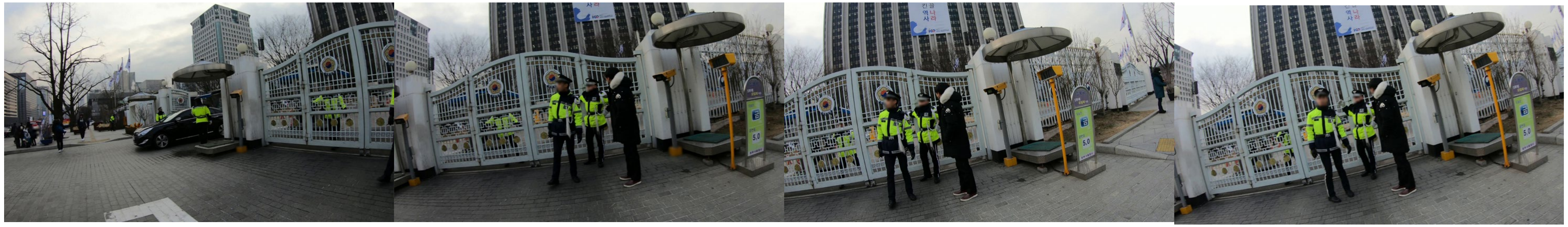
- 데이터셋 코멘트 연락 수신 대기중

산학 협력 프로젝트

- 결과보고서 작성 완료, 제출 계획 중

보행자 데이터셋

<광주과학기술원 도로상 보행자, 차, 도로표지, 안전표지, 신호기, 정류장 등 이미지 데이터셋>
- 포함 정보: 촬영날짜, 환경, 객체분류번호, Bounding Box 좌표(x, y, w, h).
- 완전한 자료는 신청 필요.



<광주과학기술원_도로상 경찰관 및 안전요원, 보행자 이미지>
- 포함 정보: 촬영날짜, 환경, 객체분류번호, Bounding Box 좌표(x, y, w, h).
- 완전한 자료는 신청 필요.

보행자 데이터셋

<ETH 데이터셋>

- 주석정보 포함되어 있음.
- 다운로드 가능. 이 밖에도 다운로드 가능한 데이터셋이 있는 것으로 보임.

References

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.