Learning to Explain: An Information-Theoretic Perspective on Model Interpretation

2020/09/29

楊明哲

論文情報

- 著者
  - Jianbo Chen (UC Berkeley, Ant Financialのインターン)
  - LeSong (Georgia Institute of Technology, Ant Financial)
  - Martin J. Wainwright (University of California, The Voleon Group)
  - Michael I. Jordan (University of California)
- 出典: ICML2018

概要

どんな論文?

- モデルが「判断(出力)の根拠となった特徴」を説明するような説明モデルを学習する方法を提案している.
- 従来の説明モデルと違って，一度説明モデルの学習を行えば説明の生成を毎回計算しなくていいので非常に高速.
- 選んだ理由:
  - 自分の研究に説明可能AIを取り入れたいと思ったため.
  - (最近ホットでもあるから.)

概要

研究背景

- ランダムフォレスト, カーネル法，ディープニューラルネットワークなど複雑なモデルが提案され，それらによって予測精度が高まっている.
- しかし，複雑なモデルによる出力結果を人間が解釈するのは困難にもなっている.
- 本論文では予測に重要な特徴量を見つけることで，モデルの解釈性を高めようとしている.

概要

既存研究について

1. 入力のベクトルに関して，正解の出力の勾配を求める．求めた勾配を使い，入力にマスクして説明とする．

2. 説明したいサンプルの周辺で局所的に簡単な識別モデルを作り，それを使って説明を行う．

   - LIME や DeepLIFT, kernel SHAPなどがある．

概要

既存研究との違い

- 局所的な説明モデルを全体的に学習し，入力の分布も考慮に入れられる.

- 局所的な説明モデルを追加しなくても良い.

	Training	Efficiency	Additive	Model-agnostic
Parzen (Baehrens et al., 2010)	Yes	High	Yes	Yes
Salient map (Simonyan et al., 2013)	No	High	Yes	No
LRP (Bach et al., 2015)	No	High	Yes	No
LIME (Ribeiro et al., 2016)	No	Low	Yes	Yes
Kernel SHAP (Lundberg & Lee, 2017)	No	Low	Yes	Yes
DeepLIFT (Shrikumar et al., 2017)	No	High	Yes	No
IG (Sundararajan et al., 2017)	No	Medium	Yes	No
L2X	Yes	High	No	Yes

概要

貢献

貢献は次のようになる

- 情報量をもとにして，インスタンスごとの特徴選択ができるフレームワークを提案した.
- モデルに依存せず効率のよい，特徴選択のアルゴリズムを提案した.
- 提案したアルゴリズムが人工データとリアルデータで効果があることを実験した.

フレームワーク

- 提案手法では，モデルに依存しないため回帰，分類モデルのどちらにも適応できる．
  （今回は分類モデルで考える）

- 前提として，モデルの出力を条件付き確率$\mathbb{P}_m(\cdot \mid x)$，説明変数$Y$,

  確率変数$X = x \in \mathbb{R}^d$とする．

- ここでは，相互情報量をインスタンスごとの特徴選択の指標として使う．

$$I(X;Y) = \mathbb{E}_{X,Y}\left[\log \frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right]$$

説明変数と選択された特徴量の相互情報量が最大となることが今回の目標．

フレームワーク

問題設定

- X: $X \sim \mathbb{P}_X(\cdot)$周辺分布から出てくる.
- Y: 条件付き分布$(Y \mid x) \sim \mathbb{P}_m(\cdot \mid x)$から出てくる.
- $\mathscr{O}_k = \{S \subset 2^d \mid |S| = k\}$: サイズが$k$であるべき集合．特徴選択に使う.
  - kはハイパーパラメータ
- $\mathscr{E}$: $\mathbb{R}^d$からべき集合にマッピングする説明モデル.
- S: $S = \mathscr{E}(x)$で選ばれた部分集合，これによって選択された特徴の確率変数x_sを得る.

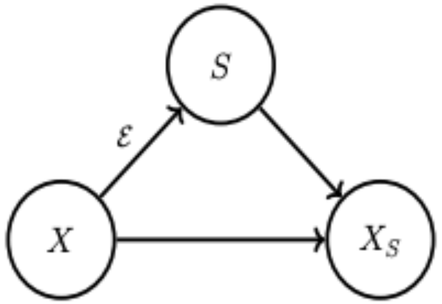

Figure 1. The graphical model of obtaining X_S from X.

フレームワーク

目的関数

- 新しい確率ベクトル$X_s \in \mathbf{R}^k$を用いると今回の目的は次式を最適化することになる.

$$\max_{\mathcal{E}} I(X_S; Y) \quad \text{subject to} \quad S \sim \mathcal{E}(X).$$

- 相互情報量を最大にするような，特徴の部分空間が識別モデルで重要な特徴であると説明することができる.

提案手法

- 目的関数を直接解くことが困難であるから，変分近似を用いて解く．
- 相互情報量の下界を最大化することを目標になる．

$$
\begin{aligned}
I(X_S;Y) &= \mathbb{E}\Big[\log \frac{\mathbb{P}_m(X_S,Y)}{\mathbb{P}(X_S)\mathbb{P}_m(Y)}\Big] = \mathbb{E}\Big[\log \frac{\mathbb{P}_m(Y|X_S)}{\mathbb{P}_m(Y)}\Big] \\
&= \mathbb{E}\Big[\log \mathbb{P}_m(Y|X_S)\Big] + \text{Const.} \\
&= \mathbb{E}_X \mathbb{E}_{S|X} \mathbb{E}_{Y|X_S}\Big[\log \mathbb{P}_m(Y|X_S)\Big] + \text{Const.}
\end{aligned}
$$

提案手法

変分下界

- 一般的なモデルでは，条件付き分布の下界の期待値を計算することが困難であるため，次のような変分族を定義する．

- $$\mathcal{Q} := \left\{ \mathbb{Q} \mid \mathbb{Q} = \left\{ x_S \to \mathbb{Q}_S\left(Y \mid x_S\right), S \in \wp_k \right\} \right\}$$

- Jensenの不等式により，下界が次のように変更される．

- $$\begin{aligned} \mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y|X_S)] &\geq \int \mathbb{P}_m(Y|X_S) \log \mathbb{Q}_S(Y|X_S) \\ &= \mathbb{E}_{Y|X_S}[\log \mathbb{Q}_S(Y|X_S)], \end{aligned}$$

提案手法

最適化問題

- $\mathcal{Q}$を用いると，今回の目標が次のように変わる．
- $$\max_{\mathcal{E},\mathbb{Q}} \mathbb{E}\Big[\log \mathbb{Q}_S(Y \mid X_S)\Big] \qquad \text{such that } S \sim \mathcal{E}(X). \quad (4)$$
- 一般的な$\mathscr{E}, \mathcal{Q}$の場合，式(4)を解くのはいまだに困難である．そこで最適化可能な方法に変える必要がある．

提案手法

$\mathcal{Q}$をパラメータ化する

- $\mathcal{Q}$をパラメータ化するために，ニューラルネットワークを定義する．
- $g_\alpha : \mathbb{R}^d \times [c] \to [0,1],\ where [c] = 0,1,\ldots,c-1$
- 今回定義したg_αを用いて，$\mathbb{Q}_x := g_\alpha(\tilde{x}_s, Y)$とする．

提案手法

サブセットサンプリングの連続緩和

- 式(4)を直接推定するには，$_nC_k$この部分空間を合計する必要があり大変.
- これを解決するために，Gumbel-softmax trickを用いる.
  - カテゴリカル分布を連続になるように緩和する手法
- Gumbel-softmax trickを使って，重みづけられた部分空間のサンプリングの近似を行なっていく.

Gumbel-softmax trickの参考: http://peluigi.hatenablog.com/entry/2018/06/23/192435

提案手法

特徴をkだけ抽出する

- dこの特徴からkこの特徴をサンプリングする方法を次に示す.

1. dこの特徴ベクトルを独立にk回サンプリングを行う.

2. かぶった特徴は削除して残りを残す.

- こうするとせいぜいkだけ特徴が残る．これらを式にすると次のようになる.

$$C^j \sim \text{Concrete}(w_\theta(X)) \text{ i.i.d. for } j = 1, 2, \ldots, k,$$
$$V = (V_1, V_2, \ldots, V_d), \quad V_i = \max_j C_i^j.$$

提案手法

最終的な目的関数とその最適化

- 最終的に，最適化するべき目的のものは次のものである．
- $$\max_{\theta,\alpha} \mathbb{E}_{X,Y,\zeta}\Big[\log g_\alpha(V(\theta,\zeta)\odot X, Y)\Big], \qquad (5)$$
- $\mathbb{E}_{X,\zeta}$はθ, αに依存しないものであるから，訓練時では，同時に勾配降下を適用することができる．

提案手法

説明ステージ

- 学習された説明モデルによって，サンプルXから重みベクトルd次元の$w_\theta(X)$にマッピングされる．
- ここからスコア$p_\theta(X)$の大きい方を選択して，説明用の特徴とする．
- 各サンプルに対して，説明モデルを通すだけで，特徴を選択することができるので，既存研究よりも効率よく説明でき，かつモデルに依存しないでおこなえる．

実験

データセット

- 人工データと実際のデータの二つで実験を行う.

- 人工データでは，10次元ガウス分布から生成して，そこから4種類のデータセットを作成.

- 2-dimensional XOR as binary classification. The input vector X is generated from a 10-dimensional standard Gaussian. The response variable Y is generated from $P(Y = 1|X) \propto \exp\{X_1 X_2\}$.
- Orange Skin. The input vector X is generated from a 10-dimensional standard Gaussian. The response variable Y is generated from $P(Y = 1|X) \propto \exp\{\sum_{i=1}^{4} X_i^2 - 4\}$.
- Nonlinear additive model. Generate X from a 10-dimensional standard Gaussian. The response variable Y is generated from $P(Y = 1|X) \propto \exp\{-100 \sin(2X_1) + 2|X_2| + X_3 + \exp\{-X_4\}\}$.
- Switch feature. Generate X_1 from a mixture of two Gaussians centered at ± 3 respectively with equal probability. If X_1 is generated from the Gaussian centered at 3, the $2-5$th dimensions are used to generate Y like the orange skin model. Otherwise, the $6-9th$ dimensions are used to generate Y from the nonlinear additive model.

-

実験

実験結果

- 本研究の手法(L2X)と既存手法(DeepLIFT, SHAP, LIME …)を比較する.

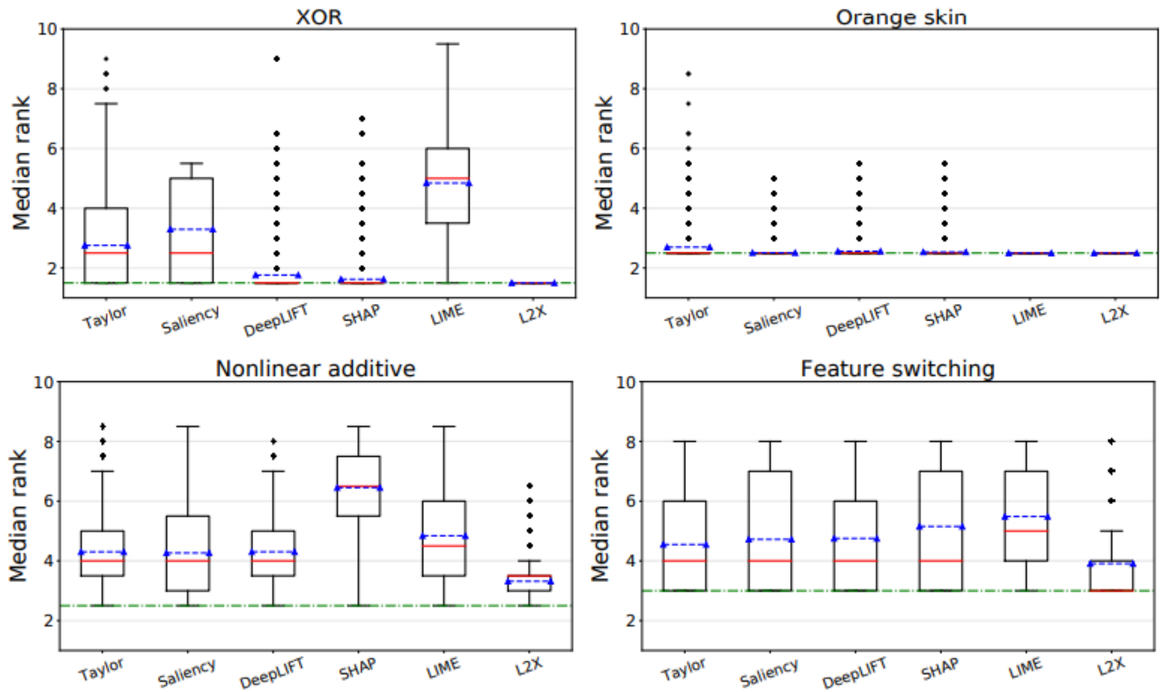

Figure 3. The box plots for the median ranks of the influential features by each sample, over 10, 000 samples for each data set. The red line and the dotted blue line on each box is the median and the mean respectively. Lower median ranks are better. The dotted green lines indicate the optimal median rank.

実験

実験結果

- 説明時の実行時間の結果を以下に示す.

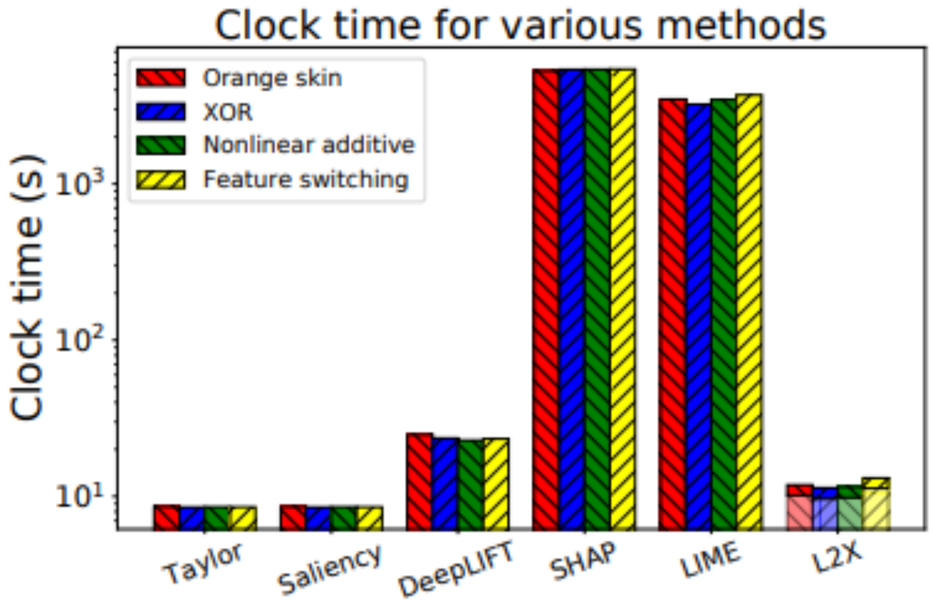

Figure 2. The clock time (in log scale) of explaining 10,000 samples for each method. The training time of L2X is shown in translucent bars.

実験結果

テキストデータでのハイライト(映画レビューのデータセット)

Truth	Model	Key words
positive	positive	Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.
negative	negative	Sorry to go against the flow but I thought this film was unrealistic, boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the crisis she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value. I wish I would have skipped it.
negative	positive	This movie is chilling reminder of Bollywood being just a parasite of Hollywood. Bollywood also tends to feed on past blockbusters for furthering its industry. Vidhu Vinod Chopra made this movie with the reasoning that a cocktail mix of deewar and on the waterfront will bring home an oscar. It turned out to be rookie mistake. Even the idea of the title is inspired from the Elia Kazan classic. In the original, Brando is shown as raising doves as symbolism of peace. Bollywood must move out of Hollywoods shadow if it needs to be taken seriously.
positive	negative	When a small town is threatened by a child killer, a lady police officer goes after him by pretending to be his friend. As she becomes more and more emotionally involved with the murderer her psyche begins to take a beating causing her to lose focus on the job of catching the criminal. Not a film of high voltage excitement, but solid police work and a good depiction of the faulty mind of a psychotic loser.

Table 2. True labels and labels predicted by the model are in the first two columns. Key words picked by L2X are highlighted in yellow.

Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.
negative	negative	You get 5 writers together, have each write a different story with a different genre, and then you try to make one movie out of it. Its action, its adventure, its sci-fi, its western, its a mess. Sorry, but this movie absolutely stinks. 4.5 is giving it an awefully high rating. That said, its movies like this that make me think I could write movies, and I can barely write.
negative	positive	This movie is not the same as the 1954 version with Judy garland and James mason, and that is a shame because the 1954 version is, in my opinion, much better. I am not denying Barbra Streisand's talent at all. She is a good actress and brilliant singer. I am not acquainted with Kris Kristofferson's other work and therefore I can't pass judgment on it. However, this movie leaves much to be desired. It is paced slowly, it has gratuitous nudity and foul language, and can be very difficult to sit through. However, I am not a big fan of rock music, so its only natural that I would like the judy garland version better. See the 1976 film with Barbra and Kris, and judge for yourself.
positive	negative	The first time you see the second renaissance it may look boring. Look at it at least twice and definitely watch part 2. it will change your view of the matrix. Are the human people the ones who started the war? Is ai a bad thing?

Table 3. True labels and labels from the model are shown in the first two columns. Key sentences picked by L2X highlighted in yellow

実験結果

クラウドワーカに評価してもらう

- 映画レビューの文に対して，クラウドワーカに10個の感情に拘るキーワードを選択してもらう.

- 複数の文章を複数のワーカに評価してもらい，その平均をHuman accuracyとする.

	IMDB-Word	IMDB-Sent	MNIST
Post-hoc accuracy	0.90.8	0.849	0.958
Human accuracy	0.844	0.774	NA

Table 4. Post-hoc accuracy and human accuracy of L2X on three models: a word-based CNN model on IMDB, a hierarchical LSTM model on IMDB, and a CNN model on MNIST.

結論

- 相互情報量を基にして，インスタンスごとの特徴選択をするフレームワークを提案した.
- L2Xにより，はじめてリアルタイムでブラックボックスなモデルを説明することが可能になった.

感想

- 数式や手法をきちんと理解できていないため，もっと勉強する必要を感じた.
- ベイズ統計勉強しないと