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

Authors: Jianbo Chen et al., 2018
Presenter: TT Nguyen

SAIL@UNIST

June 3, 2018

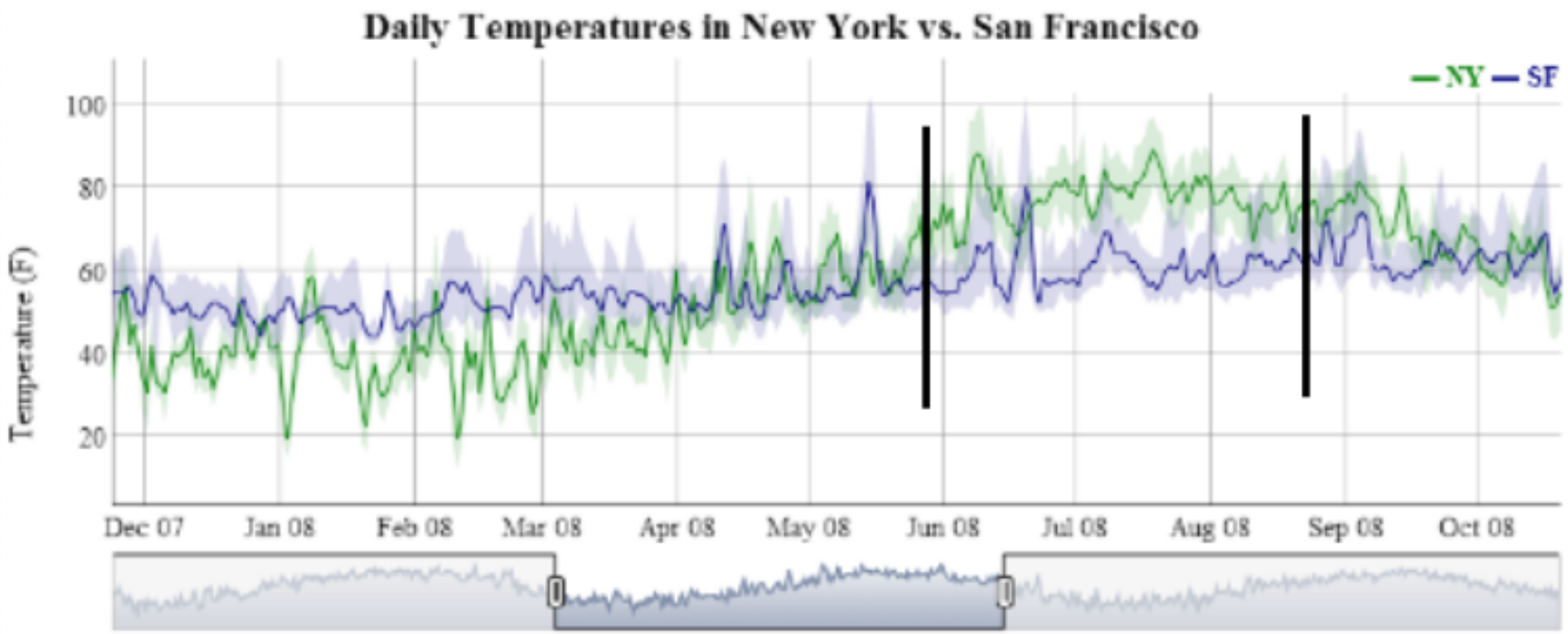
Contents

Contents

Motivation

- **Interpretability** of a machine learning model is important in many areas such as medicine, financial markets, criminal justice, and natural language understanding.
- In this scope, model **interpretability** means to tell which parts (constituents) of an *input variable* $\boldsymbol{X}$ is more important for the given model to predict a *response variable* $\boldsymbol{Y}$.

Example



Example 1: Time series classification

Example

Example 2: Sign Language Translation

The V&A Theator & Performance galleries opened in March 2009. ...
They hold the UK's biggest national collection of material about live performance.

Question: What collection does the V&A Theator & Performance galleries hold?

X

→

The V&A Theator & Performance galleries opened in March 2009. ...
They hold the UK's biggest national collection of material about live performance.

$y =$ "live performance"

Example 3: Reading comprehension

Problem Setting

- **Given**:
  - Input (random) variable: $X \in \mathbb{X}^d$
  - A given black-box predictive model $p_{model}(.|X)$ to be explained
  - Model response variable: $Y \sim p_{model}(.|X)$
- **Problem**: For each input instance $X = \boldsymbol{x}$, which subset of the input instance features is "**the most important**" for the model to make a prediction $p_{model}(.|\boldsymbol{x})$?

Note: the search space is 2^d!

Contents

Proposed Method

- An explainer,

$$\mathcal{E}_k : X \in \mathbb{X}^d \rightarrow X_S \in \mathbb{X}^k,$$

  assigns a probability to each subset $X_S = \boldsymbol{x}_S$ of an input instance $X = \boldsymbol{x}_S$
- An explainer is **optimal** if it assigns the **highest probability mass** to the subset $X_S = \boldsymbol{x}_S$ that is **the most informative** for the model response variable Y.
- The problem is then reduced to:

$$\boxed{\mathcal{E}_k^* = \max_{\mathcal{E}_k} I(X_S, Y)}$$

- $I(X_S, Y)$: measures the amount of information X_S contains about Y

Proposed Methods

- Design a *parametric* explainer $p(X_S|X)$ and its efficient sampling scheme $\boldsymbol{x}_S \sim p(X_S|X)$
- *Variational* approximation to the intractable mutual information $I(X_S, Y)$

Mutual information is intractable

$$I(X_S, Y) = \int p(\boldsymbol{x}_S, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}_S, \boldsymbol{y})}{p(\boldsymbol{x}_S)p(\boldsymbol{y})} d\boldsymbol{x}_S d\boldsymbol{y} \tag{1}$$

$$= \mathbb{E}_{Y|X_S}\mathbb{E}_{X_S}[\log \boxed{p(Y|X_S)}] - \mathbb{E}_Y[\log \boxed{p(Y)}] \tag{2}$$

intractable decoder ↓ $p(Y|X_S)$

constant ↓ $p(Y)$

where

$$p(\boldsymbol{y}) = \int p_{model}(\boldsymbol{y}|\boldsymbol{x}) p_{data}(\boldsymbol{x}) d\boldsymbol{x} \tag{3}$$

is constant (w.r.t $\mathcal{E}$, or $p(X_S|X)$, and

$$p(\boldsymbol{y}|\boldsymbol{x}_S) = \int p_{model}(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{x}_S) d\boldsymbol{x} \tag{4}$$

is intractable.

Variational Mutual Information

It follows from Jensen's inequality that

$$\mathbb{E}_{X_S}\mathbb{E}_{Y|X_S}[\log \boxed{p(Y|X_S)}] \geq \mathbb{E}_{X_S}\mathbb{E}_{Y|X_S}[\log \boxed{q(Y|X_S)}] \tag{5}$$

$$= \mathbb{E}_{X}\mathbb{E}_{X_S|X}\mathbb{E}_{Y|X}[\log \boxed{q(Y|X_S)}] \tag{6}$$

for any distribution $q(Y|X_S)$. In practice, $q(Y|X_S)$ is usually a neural network.

Design of a parametric explainer $p(X_S|X)$

- Note that $X_S \in \mathbb{X}^k, X \in \mathbb{X}^d$, thus there are $\binom{d}{k}$ k-element subsets of d features of X!
- Assigning probability mass to such a large space is impractically impossible $\rightarrow$ requires an efficient approximation

Gumbel-softmax relaxation of discrete subset sampling

Algorithm 1 Gumbel-softmax relaxation of discrete subset sampling

$w_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$

$\boldsymbol{p} := (p_1, ..., p_d) = softmax(w_\theta(X)) \in \mathbb{R}^d$ //Parametric explainer

//Sampling X_S from X

for $j =$ from 1 to k :

$\boldsymbol{\epsilon} := (\epsilon_i)_{i=1}^d$ where $\epsilon_i \sim Gumbel(0, 1), \forall 1 \leq i \leq d$

$\boldsymbol{C}^j := softmax(\frac{\log \boldsymbol{p} + \boldsymbol{\epsilon}}{\tau}) \in \mathbb{R}^d$

$\boldsymbol{V} := (\max_{1 \leq j \leq k} C_i^j)_{i=1}^d \in \mathbb{R}^d$ //Choose the most probable one among k values

$s = \boldsymbol{V} \odot X$

Contents

Experiment: Explaining CNN and Hierarchical LSTM for Sentiment Analysis

- The Large Movie Review Dataset (IMDB) is a dataset of movie reviews
- $50,000$ labeled movie reviews, with a split of $25,000$ for training and $25,000$ for testing
- Average document length: 231 words, and 10.7 sentences

Metrics

- **Post-hoc accuracy**: Compute the accuracy of $p_{model}(Y|_S)$ in the test set labeled by $p_{model}(Y|X)$
- **Variational mutual information (VMI)**:

$$VMI = \mathbb{E}_X \left[\sum_{y=1}^{c} p_{model}(y|X) \log \frac{p_{mode}(y|_S)}{\mathbb{E}_X p_{model}(y|X)} \right]$$

- **Human accuracy**: Humans infer the sentiment only based on the k selected keywords. The sentiment result is then aligned with the labels provided by the model of interest.

Explaining CNN for sentiment analysis

Each word as a feature and select the 10 most informative words:

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.

Figure: The most 10 informative words selected by L2X for CNN

Explaining CNN for sentiment analysis

	Taylor	Saliency	SHAP	LIME	L2X
Post-hoc accuracy	0.8444	0.839	0.7024	0.6756	**0.868**
VMI	0.2041	0.1226	0.0013	0.0032	**0.222**
Human accuracy	0.766	0.696	0.592	0.56	**0.804**

Explaining LSTM for sentiment analysis

Each sentence as a feature and select the most informative sentence:

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 ?

Figure: The most informative sentence selected by L2X for LSTM

Explaining LSTM for sentiment analysis

	Taylor	Saliency	SHAP	LIME	L2X
Post-hoc accuracy	0.818	0.621	0.659	0.736	**0.849**
VMI	0.1022	0.0984	0.0120	0.1465	**0.287**
Human accuracy	0.738	0.552	0.638	0.608	**0.774**

Contents

Conclusion

- A framework for feature selection via mutual information
- Variational approximation to mutual information
- Gumbel-softmax relaxation of discrete subset sampling
- Empirical results