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

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan

Institute of Information Science

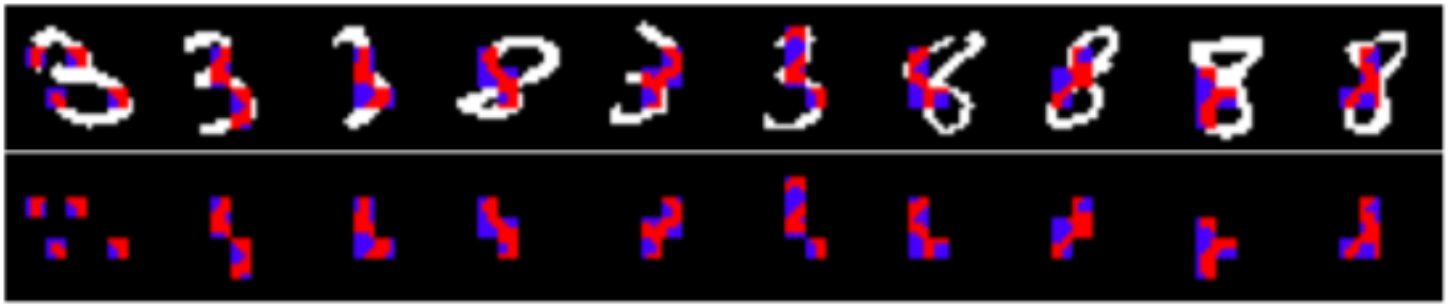
Academia Sinica

Taiwan

Aug 10, 2018

Motivation

Motivation : curious about which features are the key for the model to make its prediction on the instance



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.

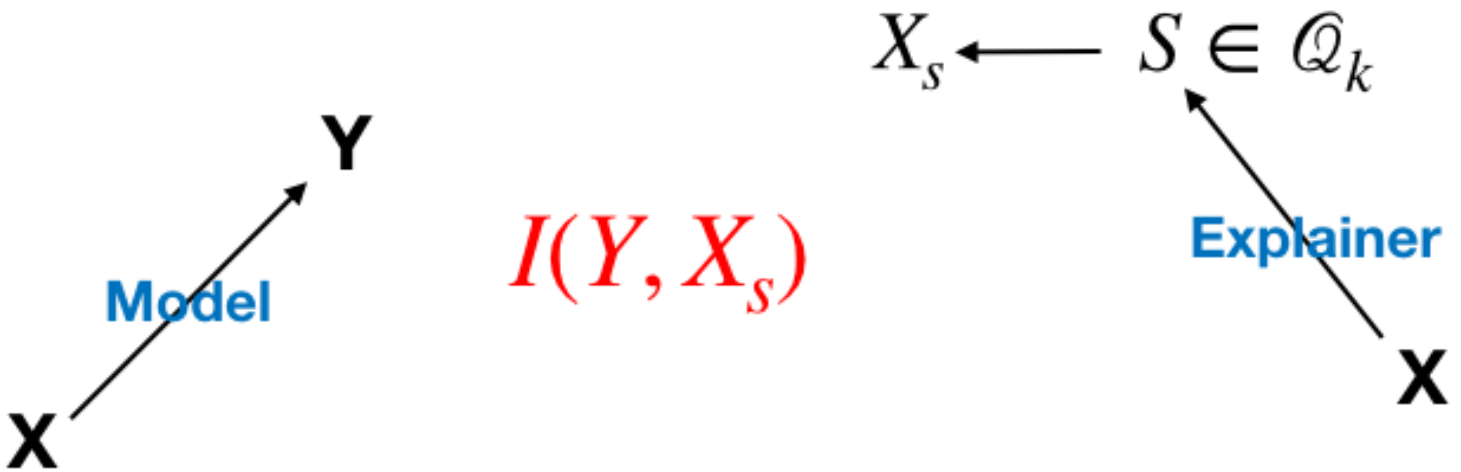
Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and 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?

Mutual information

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

where $p_{XY}(X,Y)$ is the joint probability density and $P_X(X), P_Y(Y)$ are the marginal probability density.

Intuitively, it corresponds to how much knowledge of one random vector reduces the uncertainty about the other.



Notations

- data : (X, Y)
- $(Y|x) \sim P_m(\cdot|x)$
- $\mathcal{Q}_k = \{S \subset 2^d : |S| = k\}$
- Given the chosen subset $S = \mathcal{E}(x)$, we use x_S to denote the sub-vector formed by the chosen features.

objective function :

$$max_{\mathcal{E}} I(X_S : Y) \text{ subject to } S \sim \mathcal{E}(X) \tag{1}$$

flow

It remains to solve (1) efficiently.

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

1. Relax by maximizing a variational lower bound (2)

$$max_{\mathcal{E},Q} \mathbb{E}\left[logQ_S(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

2. Then, approximate Q by a single neural network g_α. (3)

$$\Rightarrow max_{\mathcal{E},\alpha} \mathbb{E}\left[\ln g_\alpha(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

3. Approximate sampling from categorical distribution by sampling from a continuous distribution : Gumbel-softmax trick. Therefore, the objective function becomes

$$max_{\theta,\alpha} \mathbb{E}_{X,Y,\eta}\left[\ln g_\alpha(V(\theta,\eta)\cdot X, Y)\right]$$

flow(1)

1. Relax by maximizing a variational lower bound (2)

$$max_{\mathcal{E},Q}\mathbb{E}\left[logQ_S(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

2. Then, approximate Q by a single neural network g_α. (3)

$$\Rightarrow max_{\mathcal{E},\alpha}\mathbb{E}\left[\ln g_\alpha(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

3. Approximate sampling from categorical distribution by sampling from a continuous distribution : Gumbel-softmax trick.
Therefore, the objective function becomes

$$max_{\theta,\alpha}\mathbb{E}_{X,Y,\eta}\left[\ln g_\alpha(V(\theta,\eta)\cdot X, Y)\right]$$

1. variational lower bound

A direct solution to (1) is not possible, so that we need to approach it by a variational approximation.

$$\begin{aligned} I(X_S : Y) =& H(Y) + < \ln P_m(Y|X_S) >_{P_m(X_S,Y)} \\ \geq& H(Y) + < \ln Q_S(Y|X_S) >_{Q_S(X_S,Y)} \end{aligned}$$

$Q_S(Y|X_S)$ is an arbitrary variational distribution. This bound becomes exact if $Q_S(Y|X_S) \equiv P_m(Y|X_S)$. Problem (1) can thus be relaxed to

$$max_{\mathcal{E},Q} \mathbb{E}\left[log Q_S(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X) \tag{2}$$

where Q is a collection of conditional distributions.

$$Q = \{Q_S(\cdot|x_S), S \in \mathcal{Q}_k\}$$

flow(2)

1. Relax by maximizing a variational lower bound (2)

$$max_{\mathcal{E},Q}\mathbb{E}\left[logQ_S(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

2. Then, approximate Q by a single neural network g_α. (3)

$$\Rightarrow max_{\mathcal{E},\alpha}\mathbb{E}\left[\ln g_\alpha(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

3. Approximate sampling from categorical distribution by sampling from a continuous distribution : Gumbel-softmax trick.
Therefore, the objective function becomes

$$max_{\theta,\alpha}\mathbb{E}_{X,Y,\eta}\left[\ln g_\alpha(V(\theta,\eta)\cdot X, Y)\right]$$

approximate Q

Recall that Q is a collection of conditional distributions.
Then we introduce a single neural network function $g_\alpha : \mathbb{R}^d \to \Delta_{c-1}$ for parametrizing Q, where Δ_{c-1} is a $(c-1)$-simplex for the class distribution

$$\Rightarrow max_{\mathcal{E},\alpha} \mathbb{E}\left[\ln g_\alpha(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X) \tag{3}$$

flow(3)

1. Relax by maximizing a variational lower bound (2)

$$max_{\mathcal{E},Q}\mathbb{E}\left[logQ_S(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

2. Then, approximate Q by a single neural network g_α. (3)

$$\Rightarrow max_{\mathcal{E},\alpha}\mathbb{E}\left[\ln g_\alpha(Y|X_S)\right] \text{ subject to } S \sim \mathcal{E}(X)$$

3. Approximate sampling from categorical distribution by sampling from a continuous distribution : Gumbel-softmax trick. Therefore, the objective function becomes

$$max_{\theta,\alpha}\mathbb{E}_{X,Y,\eta}\left[\ln g_\alpha(V(\theta,\eta)\cdot X, Y)\right]$$

approximate $\mathcal{E}$

$S \sim \mathcal{E}(X)$ is actually sampling from a categorical distribution, which is discrete.
$\Rightarrow$use Gumbel-Softmax distribution to approximate categorical distribution.
We define $w_\theta : \mathbb{R}^d \to \mathbb{R}^d$ that maps the input to a d-dimensional vector, with the ith entry of $w_\theta(X)$ representing the "importance score" of ith feature.
Sample a single feature out of d features independently for k times.

objective function

The objective function then becomes :

$$max_{\theta,\alpha}\mathbb{E}_{X,Y,\eta}\left[\ln g_\alpha(V(\theta,\eta)\cdot X, Y)\right]$$

where g_α denotes the neural network used to approximate the model conditional distribution, and the quantity θ is used to parametrize the explainer.

- In each update, we sample a mini-batch of unlabeled data with their class distributions from the model to be explained.
- During the explaining stage, the learned explainer maps each sample X to a weight vector $w_\theta(X)$ of dimension d, each entry representing the importance of the corresponding feature for the specific sample X.

Experiment-IMDB

- Post-hoc accuracy.
  We feed in the sample X_S to the model with unselected words masked by zero paddings. Then we compute the accuracy of using $P_m(y|X_S)$ to predict samples in the test data set labeled by $P_m(y|X)$
- Human accuracy.
  We ask humans on Amazon Mechanical Turk (AMT) to infer the sentiment of a review given the ten key words selected by explainer.

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