Smarter technology for all

Kernel-based LIME with Feature Dependency Sampling

Sheng Shi, Yangzhou Du, Wei Fan
Lenovo Research

Contents

The core characteristics of local explanation methods

An explainable model with good interpretability should be faithful to the original model, understandable to the observer, and graspable in a short time so that the end-user can make wise decisions.

Local explanation method learns a model from a set of data samples which is sampled around the instance being explained. The general framework of LIME:

$$J(\theta) = argmin L\big(f(x), g_\theta(\theta)\big) + \lambda\Omega(\theta)$$

Local fidelity:
the dissimilarity between the true label and predicted label is a measure of how unfaithful g(x) is approximating f(x)

Understandability:
the regularization term is a measure of complexity of explainable model

M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” (SIGKDD 2016)

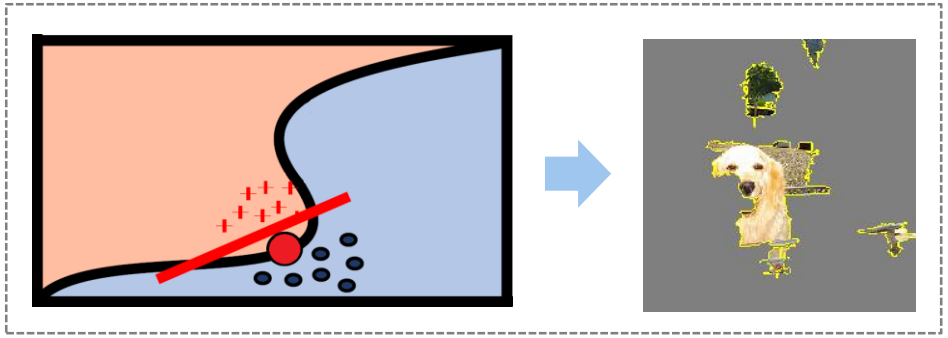
Two drawbacks in current existing local explanations

Drawback1: Perturbed samples are generated from a random uniform distribution ignoring the intrinsic correlation between features.

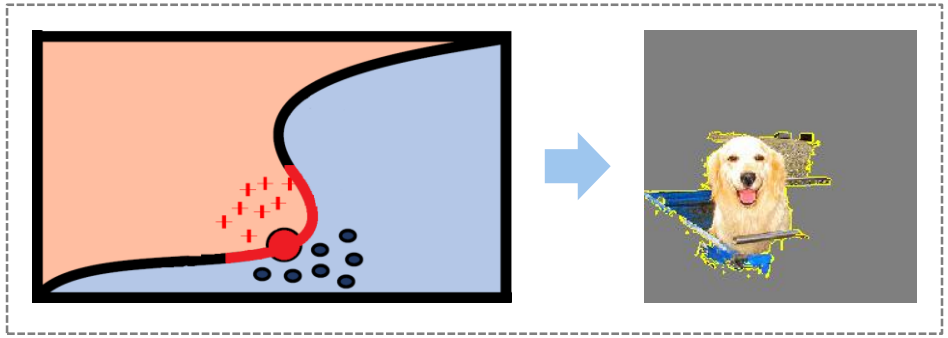
- The visual features of natural objects exhibit a strong correlation in the spatial neighborhood
- False information leads to poorly fitting of the local explanation model

Drawback2: Most existing methods assume the decision boundary is locally linear.

- This may produce serious errors as in most complex networks, the local decision boundary is non-linear.



LIME



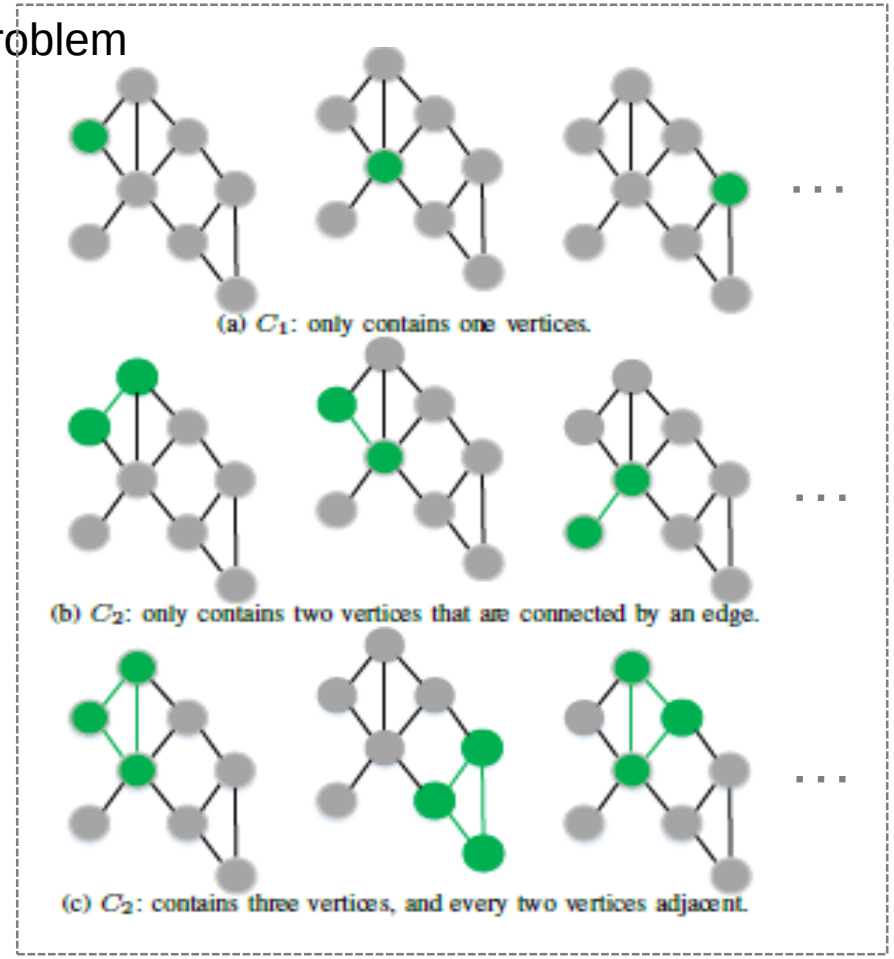
KLFDS

Feature Sampling with Intrinsic Dependency

KLFDS: Kernel-based LIME with Feature Dependency Sampling

Design an unique local sampling process which incorporates the feature clustering method to handle the feature dependency problems.

- Convert the super-pixel image into an undirected graph
- Perturbed sampling operation is formalized as clique set construction problem



Nonlinear Boundary of Local Decision

Experiments show a simple linear approximation will significantly degrade the explanation fidelity.

We adopt SVR with kernel function to approximate nonlinear boundary.

- In approximation processing, when data are not distributed linearly in the current feature space, we use kernel function to project data points into higher dimensional feature space and find the optimal hyperplane

$$err(f(x_i), (g(k(x_i), w)) =$$

$$\begin{cases} 0, & \|f(x_i) - g(k(x_i), w)\| \leqslant \varepsilon \\ \|f(x_i) - g(k(x_i), w)\| - \varepsilon, & \|f(x_i) - g(k(x_i), w)\| > \varepsilon \end{cases}$$

$$\min_{w, \xi_i, \hat{\xi}_i} \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) + \lambda \|w\|^2$$

$$s.t:$$

$$f(x_i) - g(k(x_i), w) \geqslant \varepsilon + \xi_i;$$
$$f(x_i) - g(k(x_i), w) \leqslant \varepsilon + \hat{\xi}_i;$$
$$\xi_i \geqslant 0; \hat{\xi}_i \geqslant 0, i = 1, 2, ..., N.$$

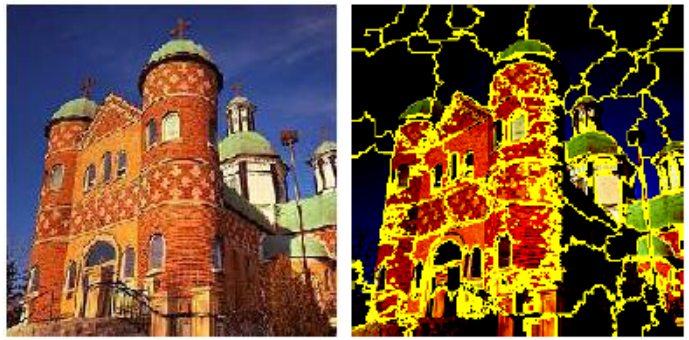
Algorithm 2 Kernel-based LIME with Feature Dependency Sampling (KLFDS)

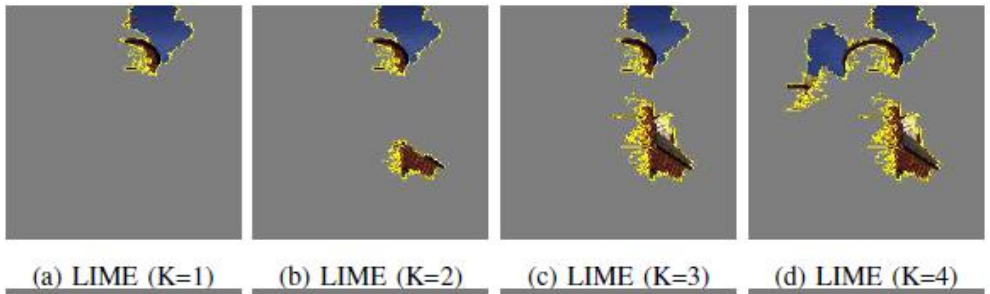
Require: Classifier f, Instance x,

1: get interpretable presentation of x' (e.g. superpixel image for image and bag of word for text)
2: get $f(x')$ by classifier f
3: incorporate the feature clustering method into sampling process to activate a subset of features
4: initial $Z \leftarrow \{\}$
5: **for** $z' \in C$ **do**
6: get z by recovering z'
7: $Z \leftarrow Z \cup (z'_i, f(z_i), \pi_x(z_i))$
8: **end for**
9: use kernel function to project data points into higher dimensional feature space: $g(x, w) = \sum_{i=1}^{N} w_i k(x - x').$;
10: use the support vector regression to search for a hyperplane
11: **return** feature coefficient

Experiments

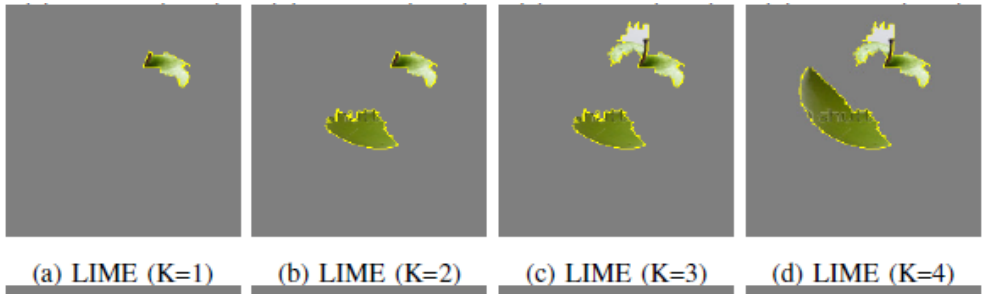
We perform various experiments to explain the Google's pre-trained Inception neural network on Imagenet database. We compare the experimental results in terms of understandability and fidelity.





(a) LIME (K=1) (b) LIME (K=2) (c) LIME (K=3) (d) LIME (K=4)

(a) KLFDS (K=1) (b) KLFDS (K=2) (c) KLFDS (K=3) (d) KLFDS (K=4)



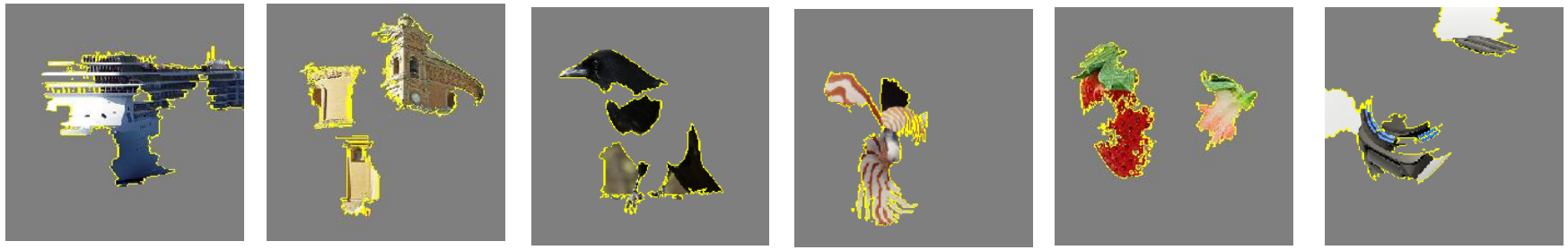
(a) LIME (K=1) (b) LIME (K=2) (c) LIME (K=3) (d) LIME (K=4)

(a) KLFDS (K=1) (b) KLFDS (K=2) (c) KLFDS (K=3) (d) KLFDS (K=4)

Experiments

We perform various experiments to explain the Google's pre-trained Inception neural network on Imagenet database. We compare the experimental results in terms of understandability and fidelity.

Original images



The super-pixels explanations by LIME

The super-pixels explanations by KLFDS

Experiments

We perform various experiments to explain the Google's pre-trained Inception neural network on Imagenet database. We compare the experimental results in terms of understandability and fidelity.

Error									
	castle	yawl	Granny-smith	church	magpie	strawberry	linear	Digital-watch	nautilus
LIME	0.2211	0.2053	0.3085	0.2248	0.2655	0.5994	0.2753	0.0674	0.3846
KLFDS	**0.0012**	**0.001**	**0.0012**	**0.0005**	**0.0010**	**0.0013**	**0.0012**	**0.0013**	**0.0006**

R-square									
	castle	Yawl	Granny-smith	church	magpie	strawberry	linear	Digital-watch	nautilus
LIME	0.3219	0.4662	0.5769	0.4644	0.3602	0.5299	0.6341	0.5834	0.3949
KLFDS	**0.896**	**0.9803**	**0.8118**	**0.5890**	**0.7955**	**0.8282**	**0.8414**	**0.9980**	**0.8872**

Compared with LIME in term of interpretability and fidelity, KLFDS has better performance in explaining classification

Conclusion and Future work

We design and develop a novel, high-fidelity local explanation method to address the two challenges in current existing explanations. By simultaneously preserving feature dependency and local non-linearity, our method produces high-fidelity and high-interpretability explanations.

There are some avenues of future work that we would like to explore. This paper only describes the modified perturbed sampling method for image classification. We will apply the similar idea to text processing and structural data analytics. Besides, we will improve other post hoc explanations techniques that rely on input perturbations such as SHAP and propose a general optimization scheme.

thanks.

Smarter technology for all

Lenovo