”Why Should I Trust You?”: Explaining the Predictions of Any Classifier[1]

付莹

华南理工大学
工商管理学院

2020 年 10 月 18 日

Outline

1. Introduction

2. Local Interpretable Model-Agnostic Explanations

3. SubModular Pick For Explaining Models

4. Experiments
   - Simulated user experiments
   - Evaluation with human subjects

background[1]

A debate in 2017 NIPS: Interpretability is necessary in ML

1. 在一些情况下人们会需求可解释性：
   - 需要了解数据:什么特征会如何影响输出结果
   - 机器学习debugging
   - 可靠性：不懂的人关心可解释性是因为信任度。正如80年代时人们发现乘坐的女飞行员驾驶的飞机会下飞机。
2. 现实生活中，可解释性并没有那么重要。人人都在用复杂的东西，比如很少有人能够完全理解英特尔i7芯片的工作原理，但大家都很自然地用，而且很好用。
3. 假设你要做心脏手术，医生做有10%的死亡率，而手术机器人做只有1%的死亡率，医生出错时，医生可以理解自己犯了什么错，而机器不会。这种情况下，是选医生手术还是机器人?

[1] https://www.youtube.com/watch?v=2hW05ZfsUUo&abchannel=InterpretableMachineLearning

Overall and By-Instance Feature Importance[2]

Overall Feature Importance

关注的是哪些自变项 (Independent Variable, X) 对因变项 (Dependent Variable, Y) 的影响程度最高。以 tree-based 模型为例，以某个变量为节点，一刀切下去能够让样本分得最开，那么这个变量就是比较重要的。

By-Instance Feature Importance

针对某个想要分析的样本(分类任务中)，它被分成哪一个类，哪个变量对它造成影响最大。

根据患者的生理指标来判断用户是否会感冒

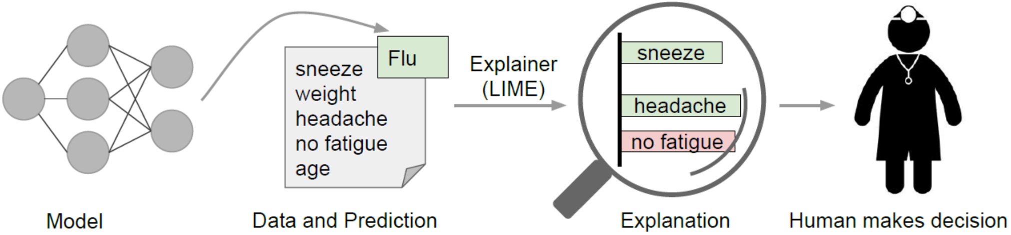
以Overall Feature Importance而言，可能 ‘发高烧’ 会是最重要的特征 (发高烧这个特征可以把整体患者分得最开); 但是对于其中某个病患(小明), 他并没有发烧，但是他有流鼻涕，所以也被预测成感冒。那么对于小明这个样本而言，将他预测成为感冒的原因，就不是发高烧，而是流鼻涕。

[2] https://medium.com/@kstseng/lime-local-interpretable-model-agnostic-explanation-%E6%8A%80%E8%A1%93%E4%BB%8B%E7%B4%B9-a67b6c34c3f8

Why we need explanation?

get humans to trust

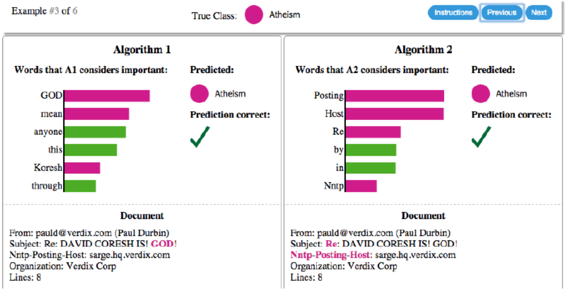
- 在用机器学习来帮助医生或律师等专业人士进行判断时，人们非常重视理由
- 根据患者的体重、头痛等特征训练一个患者是否患流感的模型。其中有一个患者被预测成患有流感，可解释的LIME模型则指出 “sneeze” 和 “headache” 是对结果(“Flu”)起支持作用的，而 “no fatigue” 是不起支持作用的。医生可以根据模型的结果以及原因解释去进行决策



Why we need explanation?

use machine learning effectively

- 建立模型后，透过样本的可解释性，判断模型是否真的学到如人类所预期的事情。
- 建立一个文本分类模型，预测一份文件是“Christianity”(基督教)还是“Atheism”(无神论)。柱状图表示这份文件中对分类结果最重要的单词的重要性，也在文本中突出显示。颜色决定了这个词属于哪个类别(绿色代表“Christianity”，品红代表“Atheism”)



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Comments

- Algorithm1 和 Algorithm2 都预测当前文本是“Atheism”，然而 Algorithm2 判断的依据(最重要的单词)为一些功能性单词(“Posting”、“Host”等)，而这样的结果肯定不是我们想要模型学习的内容
- 上述的例子可能很极端，但它比想象中还要容易发生，而透过解释器，上述的惨剧，可以在模型真正使用前被发现

Outline

1 Introduction

2 Local Interpretable Model-Agnostic Explanations

3 SubModular Pick For Explaining Models

4 Experiments
- Simulated user experiments
- Evaluation with human subjects

Desired Characteristics for Explainers

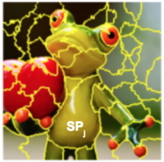
1. **Interpretable**
   - Provide qualitative understanding between the input variables and the response.
   - Interpretability must take into account the user's limitations.
2. **Local fidelity**
   - Local fidelity does not imply global fidelity
   - Must correspond to how the model behaves in the vicinity of the instance being predicted
3. **Model-agnostic**
   - An explainer should be able to explain any model
4. **Providing a global perspective**
   - Building upon the explanations for individual predictions, we select a few explanations to present to the user, such that they are representative of the model

LIME: Feature and Interpretable Data Representation

- **Feature Representation**: feeded into complex model(ANN, SVM and etc.), $x \in \mathbb{R}^d$
  - pixel from rgb channel in Image Recognition
  - word embedding of each word in Text Classification
- **Interpretable Data Representation**: a representation understandable to humans, $x' \in \{0, 1\}^d$
  - Image classification: A binary vector indicating the “presence” or “absence” of a super-pixel
  - Text sentiment: A binary vector indicating the “presence” or “absence” of a word

Original Image

$$[[r_1, r_2, \ldots, r_n], [g_1, g_2, \ldots, g_n], [b_1, b_2, \ldots, b_n]]$$



Interpretable Components

SP_1	SP_2	SP_3		SP_k
1	1	1	...	1

Intuition behind LIME

- Consider a particular model encoded as a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and a particular instance $x \in \mathbb{R}^d$ to explain. The goal is to explain the decison $f(x)$ that this model makes for one particular instance x
- f is too complicated, it is hopeless to try and fit an interpretable model globally, since the interpretable model will be too simple to capture all the complexity of f.
- A reasonable course of action is to consider a local point view, and to explain f in the neighborhood of some fixed instance.

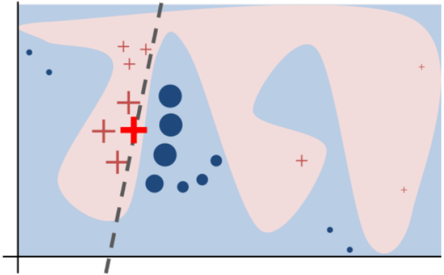


图: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets pre- dictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

LIME: Fidelity-Interpretability Trade-off

Notations

- G : Class of potentially interpretable models (e.g. Linear regression, Decision tree)
- g : Explanation (simple model, e.g. Linear regression). The domain of g is $\{0,1\}^{d'}$.
- $\Omega(g)$: measure of complexity of the explanation $g \in G$.
- $\pi_{\mathrm{x}(\mathrm{z})}$: proximity measure between an instance z to x, so as to define locality around x.
- $\mathcal{L}(f, g, \pi_{\mathrm{x}})$: measure of how faithful g is in approximating f in the locality defined by π_{x}

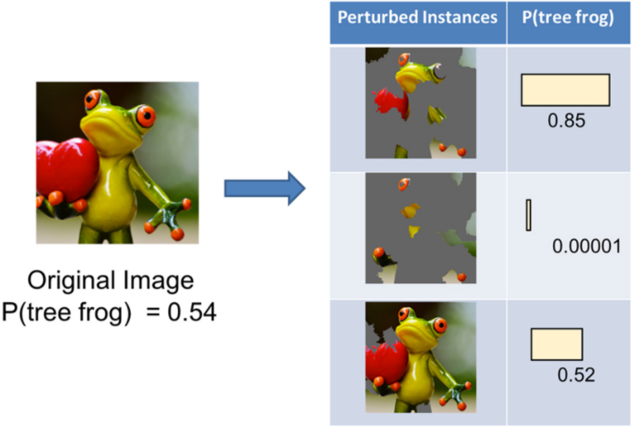
Ensure both interpretability and local fidelity

$$\xi(\mathrm{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\mathrm{x}}) + \Omega(g) \tag{1}$$

对样本x的解释就是要找到在x局部最相似，且复杂度尽量低的解释模型

LIME: Sampling for Local Exploration

- Sample instance around x' by drawing nonzero elements of x' to get a perturbed sample $\mathrm{z}' \in \{0,1\}^{d'}$ (which contains a fraction of nonzero elements of x').
- Recover the perturbed sample in the original representation $\mathrm{z} \in \mathbb{R}^d$ and obtain $f(\mathrm{z})$.



LIME: Sparse Linear Explanations

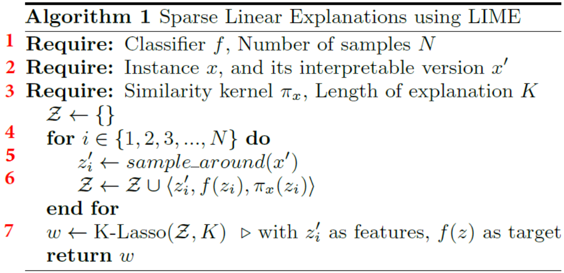
- Let G be the class of linear models, such that $g(\mathrm{z}') = w_g \mathrm{z}'$.
- $\pi_{\mathrm{x}(\mathrm{z})} = \exp(\frac{-D(\mathrm{x},\mathrm{z})^2}{\sigma^2})$, where D is the distance funcation(e.g. Cosine distance for text, L_2 distance for images) with width σ

$$\mathcal{L}\left(f, g, \pi_{\mathrm{x}}\right) = \sum_{\mathrm{z},\mathrm{z}' \in \mathcal{Z}} \pi_{\mathrm{x}(z)} \left(f(\mathrm{z}) - g\left(\mathrm{z}'\right)\right)^2 \tag{2}$$

- Approximate Ω by first selecting K features with Lasso(L_1 regularization) and then learning the weights via least squares, a procedure called K-LASSO.

LIME: Sparse Linear Explanations using LIME

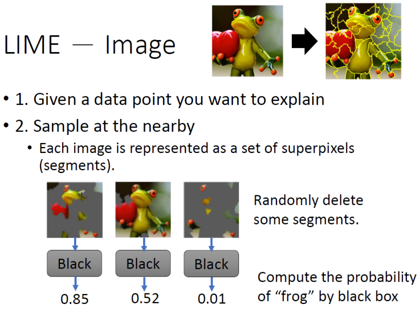
1. 一个好的分类器 f (复杂模型)
2. 选定需解释样本 x 及可解释维度 x′
3. 定义相似度指标以及解释的特征个数 K
4. 进行 N 次扰动
5. z_i' 表示从 x' 扰动而来
6. 还原 z' 到 z, 计算预测值 $f(z)$ 及相似度
7. 搜集 N 次扰动的样本后，利用LASSO Regression 求出对这个样本最有解释作用的变量对应的权重



Algorithm 1 Sparse Linear Explanations using LIME

```
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its interpretable version $x'$
Require: Similarity kernel $\pi_x$, Length of explanation $K$
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ do
      $z_i' \leftarrow sample\_around(x')$
      $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z_i', f(z_i), \pi_x(z_i)\rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$  ▷ with $z_i'$ as features, $f(z)$ as target
  return $w$
```

LIME: An ExampleLIME: An Example[3]

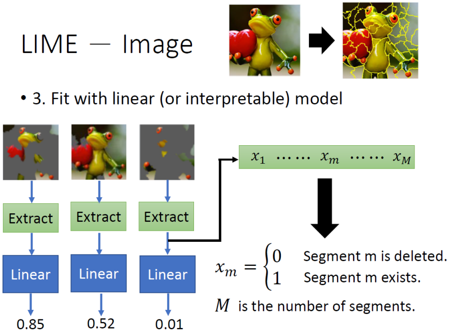


LIME — Image

- 1. Given a data point you want to explain
- 2. Sample at the nearby
  - Each image is represented as a set of superpixels (segments).

Randomly delete some segments.

Compute the probability of “frog” by black box



LIME — Image

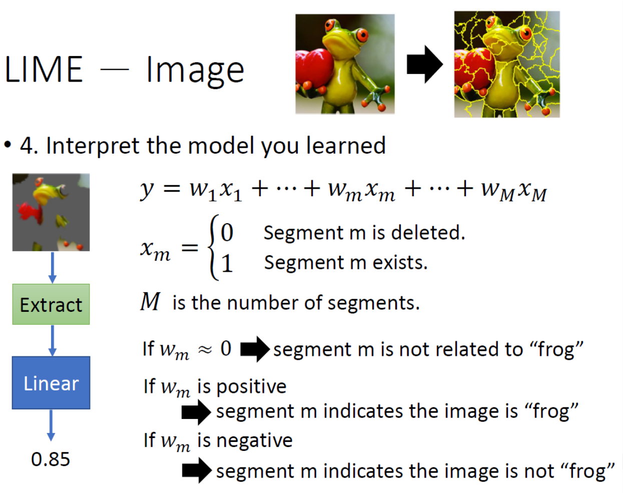
- 3. Fit with linear (or interpretable) model

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

[3] http://speech.ee.ntu.edu.tw/ tlkagk/courses/ML_2019/Lecture/XAI%20(v7).pdf

LIME: An Example[4]



LIME — Image

- 4. Interpret the model you learned

Extract

Linear

0.85

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

If $w_m \approx 0$ ➡ segment m is not related to “frog”

If w_m is positive
➡ segment m indicates the image is “frog”

If w_m is negative
➡ segment m indicates the image is not “frog”

[4] http://speech.ee.ntu.edu.tw/ tlkagk/courses/ML_2019/Lecture/XAI%20(v7).pdf

LIME的总结[5]

- LIME的主要思想是利用可解释性模型(如线性模型、决策树等）局部近似黑盒模型。通过对输入(待解释样本)进行轻微的扰动得到一系列扰动样本，黑盒模型输出这些扰动样本的预测值(探测黑盒模型的输出经过扰动后发生何种变化)。根据扰动样本与黑盒模型的相应预测训练一个可解释性模型，该模型通过扰动样本与输入样本的接近程度来加权。所学习的模型应该是机器学习模型局部预测的良好近似，但不一定是良好的全局近似。
- 一些疑问：
  1. 扰动样本是如何生成的: 对于文本和图像类数据是某个 word 或者 super-pixel 是 presence 还是 absence。但是对于结构化的表格数据呢?
  2. 应该选择多少个特征用于解释？(K越小解释这个模型越容易，较高的K可能会产生具有较高保真度的模型)。可以用向前、向后等等方法

[5] https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html

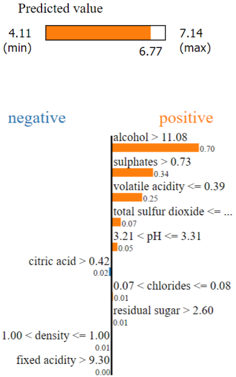
用LIME实现的例子(Tabular Data)

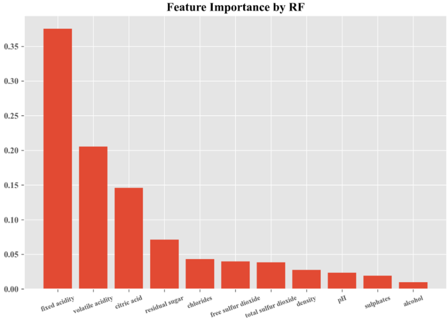
数据描述

- 对红葡萄酒质量评分(0-10)进行预测。数据来源于 kaggle
- 共1599条数据。每条数据输入的特征有11个：fixed acidity(固定酸度), volatile acidity(挥发性酸度), citric acid(柠檬酸), residual sugar(残糖), chlorides(氯化物), free sulfur dioxide(自由二氧化硫), total sulfur dioxide(总二氧化硫), density, pH, sulphates(硫酸盐) and alcohol
- 用随机森林构建回归模型，80%的数据进行训练，最大深度6，树的棵树为10。
- Python 有 LIME 的包，pip install LIME即可，选用 “LimeTabularExplainer”。解释特征跟模型训练特征一样。其余参数默认。

用LIME实现的例子(Tabular Data)

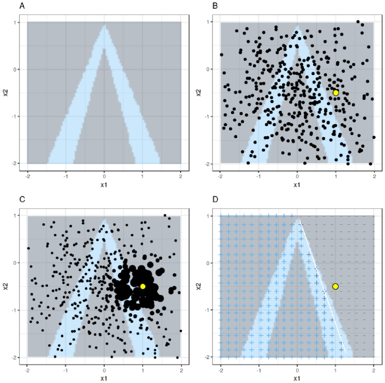
- 筛选一条实际评分为 8 (评分最高)的样本数据，用 LIME 对它进行解释
- 左：LIME, 为By-Instance Importance; 右: 随机森林重要性输出, 为Overall Feature Importance
- 对比这两幅图，可看到个体解释在一定程度上与总体重要性类似(“alcohol” 、 “sulphates” 等是对质量差起支持作用)





如何生成结构化数据的扰动样本

- If True, sample continuous features in perturbed samples from a normal centered at the instance being explained. Default, the normal is centered on the mean of the feature data(源码).
- 例子：LIME首先在全局进行采样，然后对于所有采样点，选出 x 的邻域，然后利用 x 的邻域范围拟合可解释性模型 → 可能导致一些不大可能出现的样本点来解释模型[2]

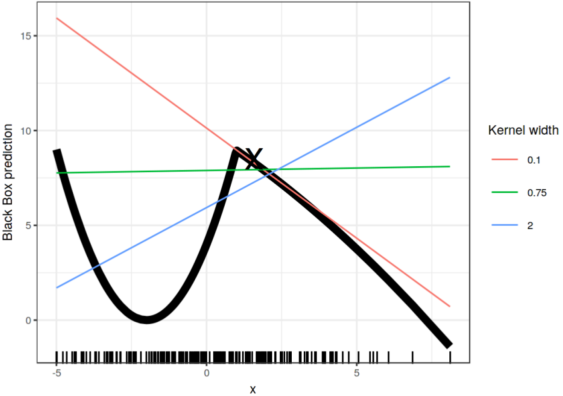


LIME algorithm for tabular data.
A) 训练随机森林模型得到的决策界面(深色是1，浅色是0)
B) 黄色点是我们感兴趣的输入样本点(待解释)，黑色点是通过正态分布随机采样的点
C) 给离黄色点近的黑色点较高的权重
D) 白色线表示对黄色点局部的拟合的可解释模型

如何确定邻域范围？

如何生成结构化数据的扰动样本：领域范围的确定[2]

- kernel: similarity kernel that takes euclidean distances and kernel width as input and outputs weights in (0,1). If None, defaults to an exponential kernel.
- kernel_width: kernel width for the exponential kernel. If None, defaults to sqrt (number of columns) * 0.75



对于 $x = 1.6$ 这个样本点，不同的邻域范围 $\{0.1, 0.75, 2\}$ 对应的可解释性模型是完全不同的，甚至相悖

LIME:解释的不稳定性[3]

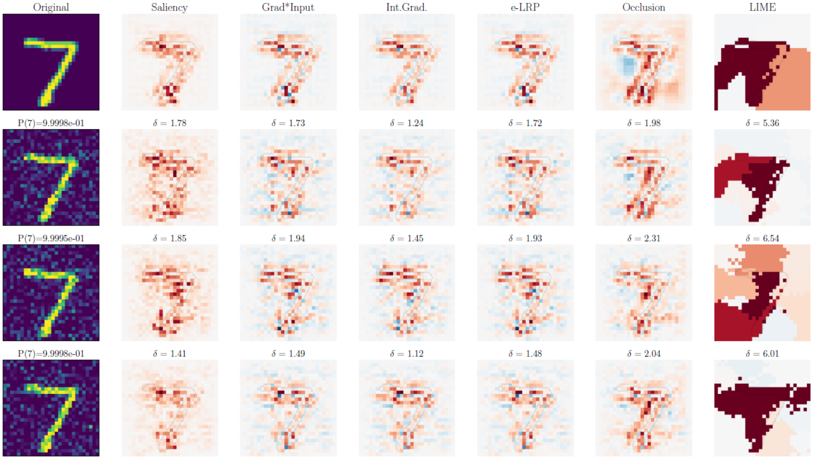


图: Explanations of a CNN model prediction's on a example MNIST digit (top row) and three versions with Gaussian noise added to it. The perturbed input digits are labeled with the probability assigned to the predicted class by the classifier.δ is the ratio $\|f(x) - f(x')\|_2 / \|x - x'\|_2$ for the pertured x'

Outline

1. Introduction

2. Local Interpretable Model-Agnostic Explanations

3. SubModular Pick For Explaining Models

4. Experiments
   - Simulated user experiments
   - Evaluation with human subjects

Get a global understanding through explain a set of individual instances

Motivation

- Explanations of a single prediction provides understanding into reliability of classifier, it's not sufficient to assess trust in the model as a whole.
- Budget B denotes the number of explanations users are willing to inspect to understand a model.
- **pick step**: Given a set of instances $\mathcal{X}$, select a diverse, representative set with B instances to inspect.
- The picked instances should cover the important components.

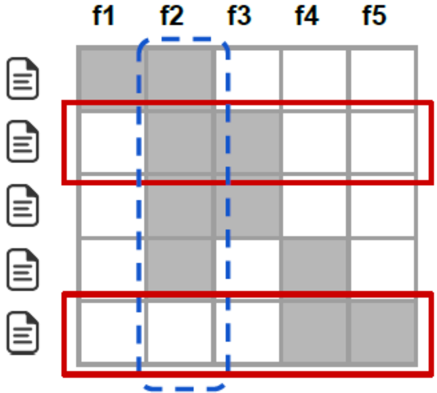


图: Toy example W.Rows represent instances (documents) and columns represent features(words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Get a global understanding through explain a set of individual instances

Notations

- Given a set of instance $\mathcal{X}(|\mathcal{X}| = n)$, an $n \times d'$ explanation matrix W represents the local importance of the interpretable components for each instance.
- For each components(column) j in W, I_j denotes the global importance of that component in the explanation space.
- Intuitively, we want I such that features that explain many different instances have higher importance scores. (i.e. $I_2 > I_1$ as figure shows)

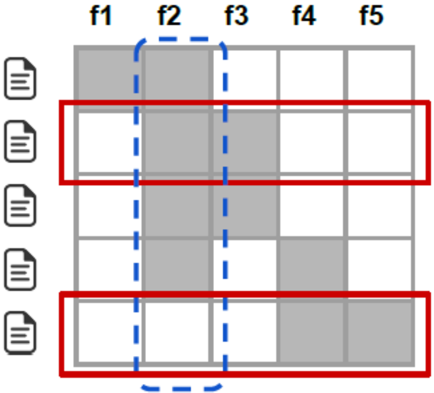


图: Toy example W.Rows represent instances (documents) and columns represent features(words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

How to pick B instances covering the important components

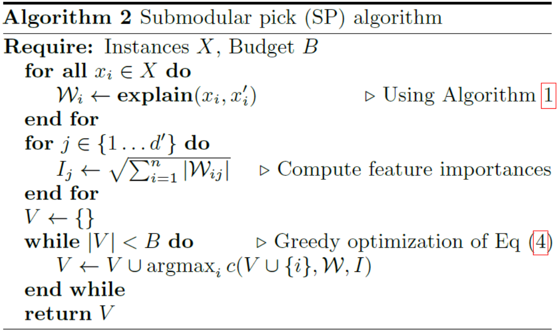
coverage: total importance of the features that appear in at least one instance in a set V.

$$c(V, \mathrm{W}, I) = \sum_{j=1}^{d'} 1_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j \qquad (3)$$

Pick problem:

$$Pick(\mathrm{W}, I) = \arg \max_{V, |V| \leq B} c(V, \mathrm{W}, I) \qquad (4)$$

A greedy algorithm adding the instance with highest $c(V \cup \{i\}, \mathrm{W}, I) - c(V, \mathrm{W}, I)$ to the solution, iteratively.



Algorithm 2 Submodular pick (SP) algorithm

```
Require: Instances $X$, Budget $B$
  for all $x_i \in X$ do
    $\mathcal{W}_i \leftarrow$ explain$(x_i, x'_i)$          ▷ Using Algorithm 1
  end for
  for $j \in \{1 \ldots d'\}$ do
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$          ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ do          ▷ Greedy optimization of Eq (4)
    $V \leftarrow V \cup \mathrm{argmax}_i\, c(V \cup \{i\}, \mathcal{W}, I)$
  end while
  return $V$
```

Outline

1. Introduction

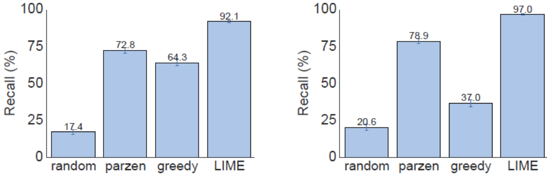
2. Local Interpretable Model-Agnostic Explanations

3. SubModular Pick For Explaining Models

4. Experiments
   - Simulated user experiments
   - Evaluation with human subjects

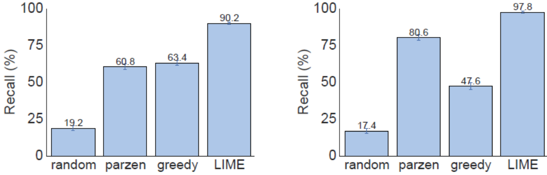
Simulated user experiments1: Are explanations faithful to the model

- Use LR and decision trees to train two sentiment analysis datasets(*books* and *DVDs*) with maximum number of features is 10.
- Generate explanations and compute the fraction of these important features that are ecovered by the explanations.
- Recall averaged over all the test instances is reported.



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for tw interpretable classifiers on the books dataset.



(a) Sparse LR (b) Decision Tree

Simulated user experiments2: Should I trust this prediction

- Randomly select 25% of features to be "untrustworthy"(users can identify)
- define trustworthiness:
  - labeling test set predictions from a black box classifier as "untrustworthy" if the prediction changes when untrustworthy features are removed from the instance, and "trustworthy" otherwise.

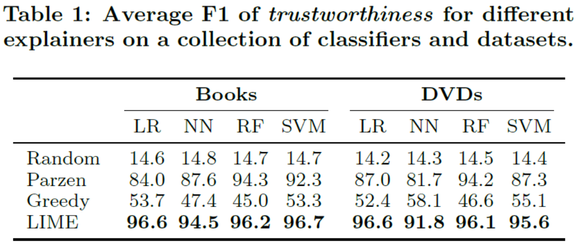
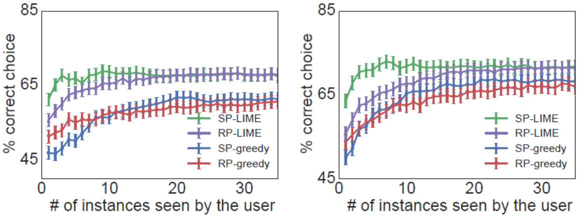


Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

Simulated user experiments3: Can I trust this model

- Add 10 artifically “noisy” features. Train two classfiers with random forest with 30 trees until their validation accuracy is within 0.1% and test accuracy differ at least 5%.
- Evaluate whether a user can identify the better classier based on the explanations of *B* instances from the validation set.
- Simulated human marks the set of artical features that appear in the *B* explanations as untrustworthy
- Select the classier with fewer untrustworthy predictions, and compare this choice to the classier with higher held-out test set accuracy.



(a) Books dataset (b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

Evaluation with human subjects1: Can users select the best classifier?

- Text sentiment classification with *religion dataset*.
  - SVM trained on the original 20 newsgroups dataset
  - SVM trained on a "cleaned" dataset with removing not generalized features.
- Two ways of explanations(greedy and LIME) are given to 100 users (Amazon Mechanical Turk-by no means machine learning experts). Users are needed to select which classfier is better.

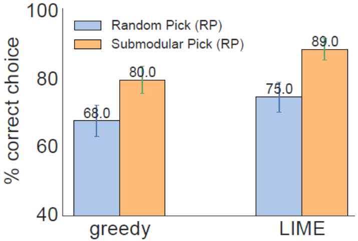


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Evaluation with human subjects: Can non-experts improve a classifier?

- 20 newsgroups data for sentiment classification. 10 subjects from Amazon Mechanical Turk are asked to identify which words from the explanations should be removed.
- Number of Interpretable features is 10 and 10 instances with explanations given to each subjects(greedy or LIME).
- Interaction goes on 3 rounds. In each round, each subject mark a word for deletion and 10 different classifiers is trained(for each subject).
- Explanations for each classifier are then passed to a set of 5 users in a new round.

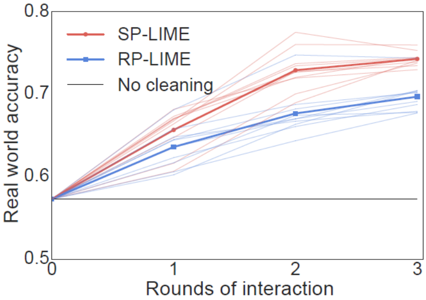
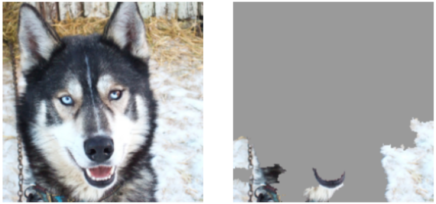


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

Evaluation with human subjects3: Do explanations lead to insights?

- Image classification with all pictures of wolves had snow in the background, while pictures of huskies did not.
- 10 pictures used for test with one wolf is not in a snowy background(the prediction is a Husky) and one husky is (the prediction is a Wolf), other 8 examples are classified correctly
- Before and after given explanation by LIME, 30 subjects are asked whether they trust or not the model.



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Reference I

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, “”’ why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[2] C. Molnar, *Interpretable Machine Learning*. Lulu. com, 2020.

[3] D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” *arXiv preprint arXiv:1806.08049*, 2018.