“Why Should I Trust You?” Explaining the Predictions of Any Classifier

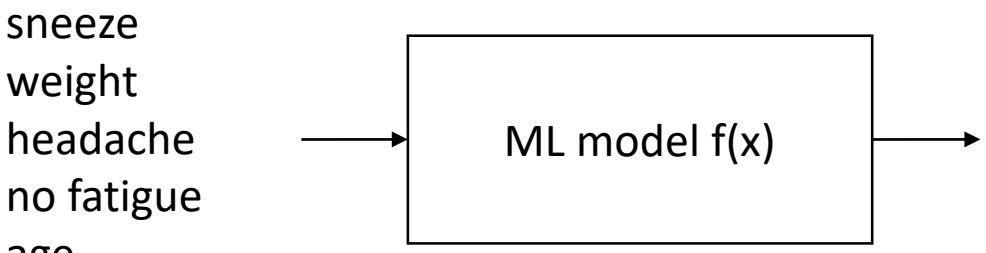
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

KDD 2016

Presented by Jinkyu Koo

Sep. 19, 2018

Introduction (1/2)

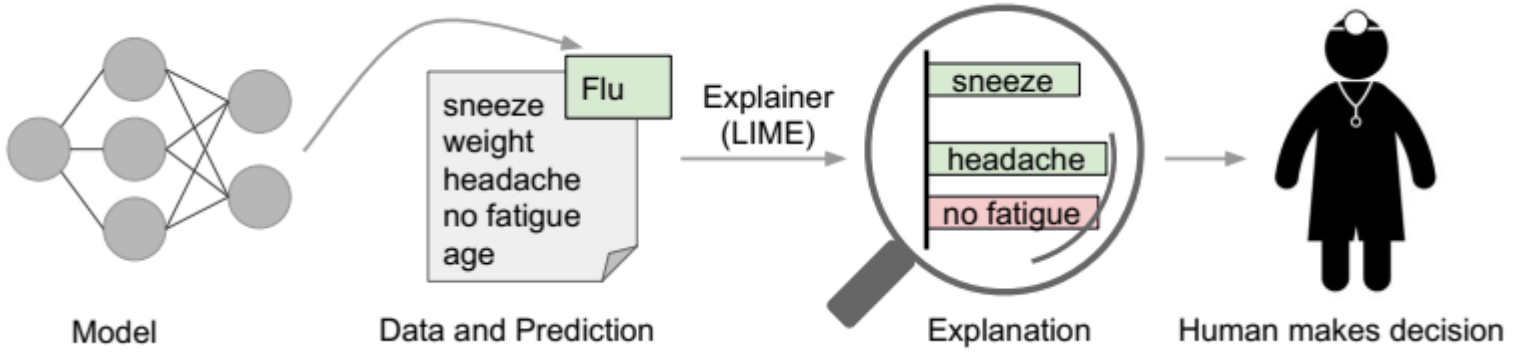


Flu! But why?

- Want to explain why we get this decision.
- If this is because of weight and age, the prediction is still true, but the reason does not make sense.

- Currently, models are evaluated using accuracy metrics on an available validation dataset.
- However, the evaluation metric may not be indicative of the product's goal.

Introduction (2/2)

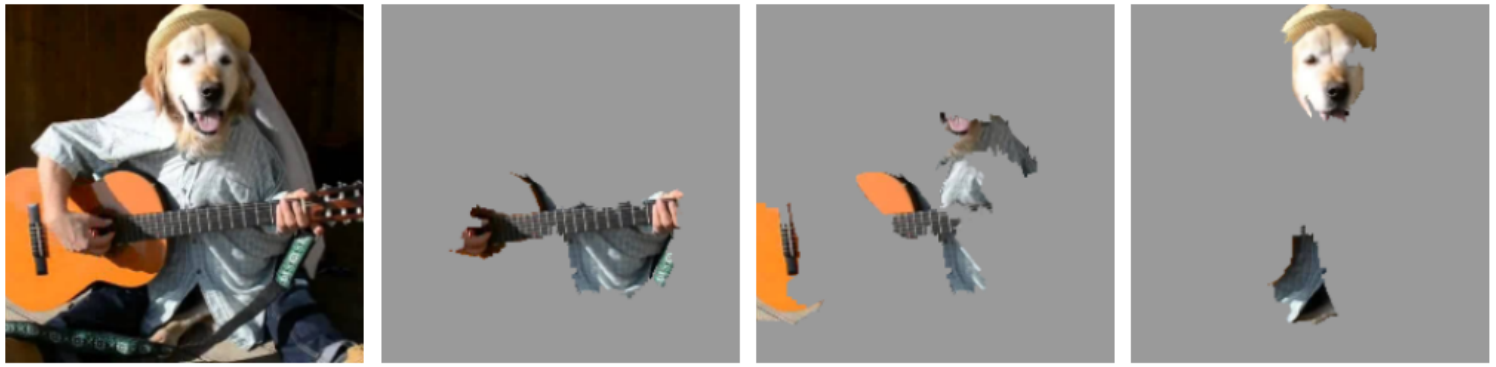


Explanation
= key features

- In this case, an explanation is a small list of symptoms with relative weights --- symptoms that either contribute to the prediction (in green) or are evidence against it (in red).
- Humans usually have prior knowledge about the application domain, which they can use to accept (trust) or reject a prediction if they understand the reasoning behind it.

Two things of Local Interpretable Model-agnostic Explanations (**LIME**):
- Local explanation: for each image, say because of these patches, this image is a cat.
- Global explanation: to make a decision, the model considers these patches.

Local explanation



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Superpixels explanations for the top 3 predicted classes.

- Figure 4b in particular provides insight as to why acoustic guitar was predicted to be electric: due to the fretboard. This kind of explanation enhances trust in the classifier (even if the top predicted class is wrong).

Sparse Linear Explanations

Model-agnostic: assume we do not know what f(z) looks like.

Super pixels or patches or features

$z \in \{0,1\}^{d'}$

1. Sample perturbed images z' that contain a fraction of non-zero elements of z.
2. Estimate f(z) by g(z') like $g(z') = w \cdot z' = w_1 z'_1 + w_2 z'_2 + \cdots$
3. Minimize the difference:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$

Weight: the closer the higher

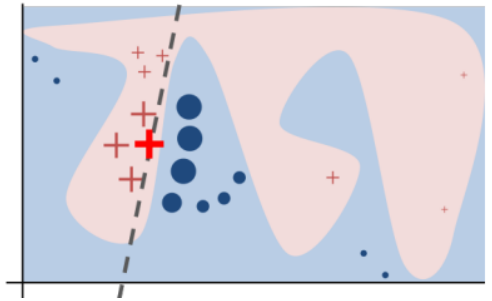
4. Select domiant features using Lasso.

$+\sum |w_i|$

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 π_x, Length of explanation K

```
Z ← {}
for i ∈ {1, 2, 3, ..., N} do
    z'_i ← sample_around(x')
    Z ← Z ∪ ⟨z'_i, f(z_i), π_x(z_i)⟩
end for
w ← K-Lasso(Z, K)  ▷ with z'_i as features, f(z) as target
return w
```



The dashed line is the learned explanation that is locally (but not globally) faithful.

Global explanation (submodular pick)

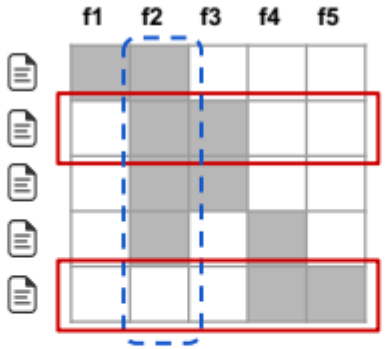
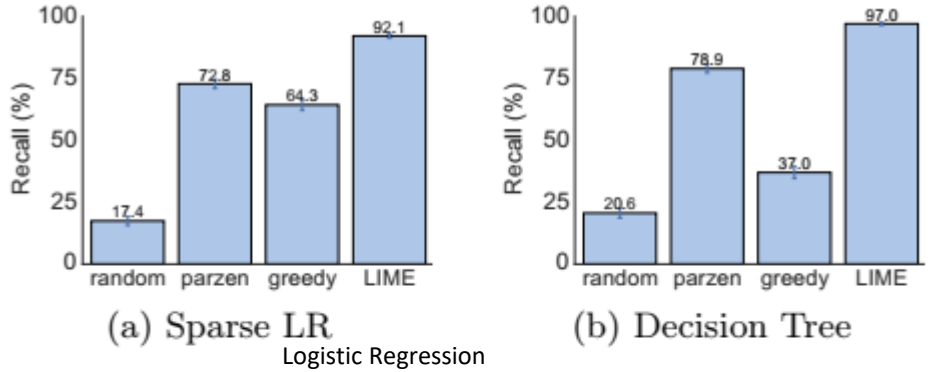


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Want to answer what features matter for the model to make a decision.

- Select images that can cover as many features as possible.
- Equivalently, select the features that can explain as many images as possible.

Are explanations faithful to the model? (true positive rate)



Logistic Regression

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

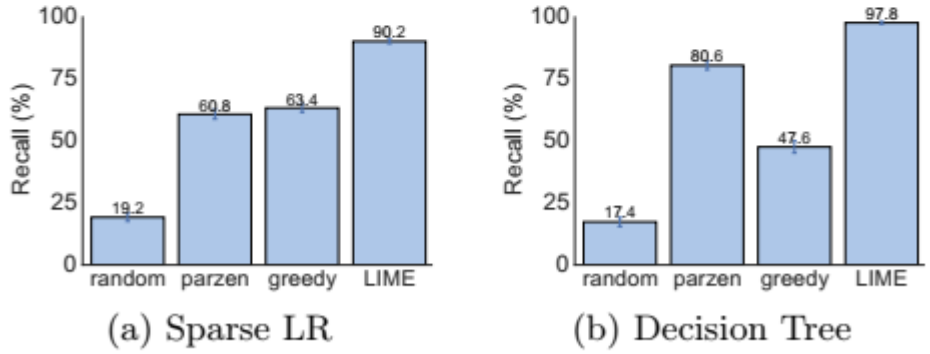


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

- **Parzen:** take the *K* features with the highest absolute gradients as explanations.
- **Greedy**: greedily remove features that contribute the most to the predicted class until the prediction changes.
- **Random:** randomly picks *K* features as an explanation.

- Train both classifiers such that the maximum number of features they use for any instance is 10, and thus we know the gold set of features that the are considered important by these models.
- For each prediction on the test set, we generate explanations and compute the fraction of these gold features that are recovered by the explanations.

Can I trust this model?

- Evaluate whether the explanations can be used for model selection, simulating the case where a human has to decide between two competing models with similar accuracy on validation data.
- Add 10% of random noisy features.
- Select the classifier that predicts relying on fewer noisy features.

This recreates the situation where the models use not only features that are informative in the real world, but also ones that introduce spurious correlations.

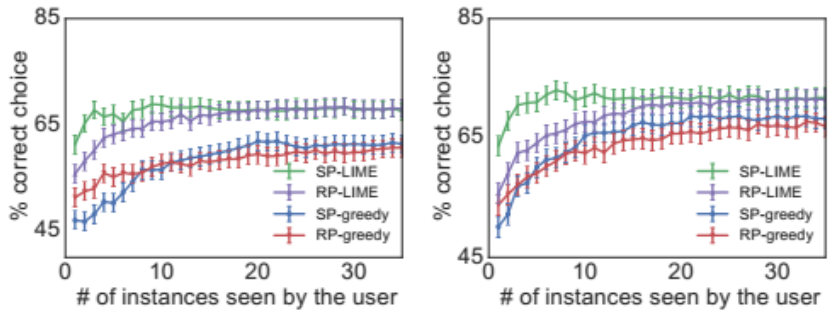
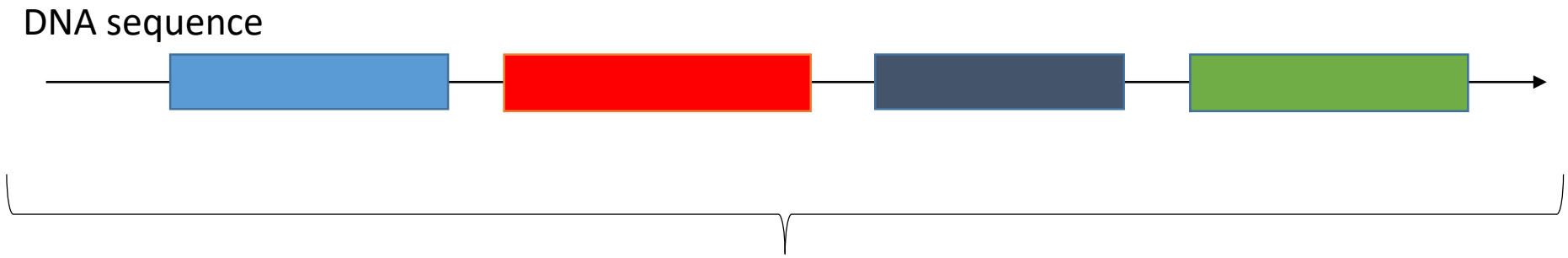


Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

SP: submodular pick for validation instances
RP: random pick for validation instances

Combining submodular pick with LIME outperforms all other methods.

Application



Say your clustering algorithm categorizes this DNA segment to a promoter,
a region of DNA that initiates transcription of a particular gene.
Then what part of this segment makes it be a promoter?

Using this paper's idea, we can find this chunk of sequence is key by which the segment can be a promoter.