Classifier Explanation

Introduction to the Algorithms LIME and SP-LIME

Leonhard Applis

TH Nürnberg

21.1.2019

Table of Contents

Me: Hey Siri, order me a Pizza

Siri: *(After a short break that nearly drains your whole battery)* Ok, I'm calling your mother...

Me: Wait! Why would you do this!?

Siri: This is the 5th time you ordered Pizza this week.

What do we want from our model?

1. Why did failed predictions fail?
2. Why did correct predictions succeed?
3. Why is my model uncertain about a prediction?

special importance:
setting a model *live*, where it's not *prelabeled*

Interpretations must be ...

- *human-readable*
- reproducable (same input + same model → same output)
- **model agnostic**, meaning they can work with any (black-box) model

Difficulties:

- Models can be huge (millions of weights)
- Inputvectors can be huge (e.g. images)
- Some models are to complex by it's structure to be readable, (e.g. neural networks)

Example

Desired output of a "Atheism"-Classifier

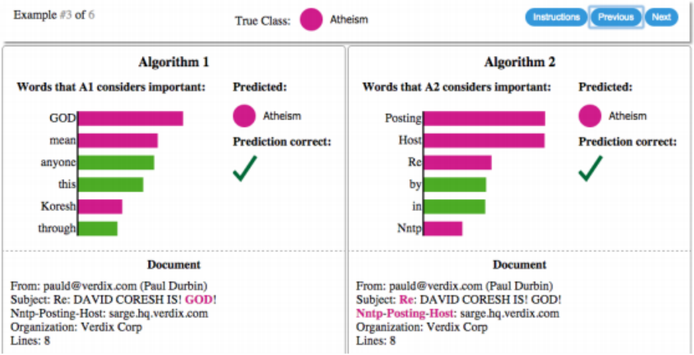

Figure: LIME-Text: predicting "Atheism" for given text

Both algorithms predict correct - yet Algorithm 2 has strange reasons.

Trusting a Model

trusting predictions $\neq$ trusting a model

What do we want?

1. get an *overview* of our Model
2. compare models in reasonable time
3. proove correctness & flaws of a model
4. improve our models

Prooving a Model

Several topics which benefit from machine learning, but need special care:

1. Terrorism-detection
2. Medical diagnosis & prescriptions
3. Fraud-detection

Noone will buy a model, if you can't prove that it's performing reasonable predictions.

Improving a Model

There are several issues, at which explanations can help you improve your models:

1. Filtering of Features
2. Find overfitted weighting of features
3. Find Links in Classification (Similiar Classes and Features)

Gaining insights from explanations can help you improve your model!

Table of Contents

What do we want:

- Human Readable Model Explanation
- For Every Classifier
- For Every Input

features $\neq$ human readable

To gain *readability*:

- show influence relative to each other, not as numbers
- only show most important features
- use *superpixels* instead of pixels

Classifier Explanati
Leonhard Applis
Intro
LIME
Example: Traffic Sign Recognition

Let:

1. G be any possible explanation model
2. g be our explanation Model
3. $\Omega(g)$ the complexity of our Model
   - Weights in a regressions model
   - Depth of an decisiontree
   - Number of trees in a random forest
4. $f : Features -> Class$ be the real classification
5. $\Pi_x(z)$ as proximity-measure from x to z
6. $\mathcal{L}(f, g, \Pi_x)$ measure of un-faithfullness of g compared to f given the proxmity Π_x

Wanted:

$$\xi(x) = argmin_{g \in G} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Read:

- We want for every input x
- an explanation(-model)
- where complexity of g and the failure of g are minimal
- given a set of possible explanations G

We do so by picking samples x' as subsets from an input x and **optimizing** our model g [1]

[1] We do not really check different models, we train one

Additional Requirements:

LASSO[2] - *Least Absolute Shrinkage and Selection Operator*

Machine Learning algorithm to select most important features relative to each other.

G are only *sparse linear regression models* (e.g. Decision Trees or simple logistic regression)

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

[2] Further Reading: https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Problem: The user can't check every prediction

Solution: SP-Lime presents a *smart* subset of predictions, which the user can check

1. Do LIME for more images/documents and different classes
2. Setup a table $feature \times class$
3. Present the user a selection of n-predictions, which show the strongest,distinct features

Note: The submodular-pick's basic idea would work without LIME

Table of Contents

given neural network :

- 6 layers, first 3 convolutional (complex structure)
- 43 Classes (complex output)
- 64x64 Images (complex input)
- trained with 8k images (rich data-input)
- tested with 2k images reaching 95% accuracy (good?)

The NN was trained with Tensorflow and is shipped with your notebook.

```python
from lime import lime_image
from skimage.segmentation import mark_boundaries

#Setup the Explainer
explainer = lime_image.LimeImageExplainer()

#Explain the predictions
explanation = explainer.explain_instance(
    image, model.predict, top_labels=43, hide_color=0,
    num_samples=1000)

#Show the mask for a class
temp, mask = explanation.get_image_and_mask(
    10, positive_only=True, num_features=5, hide_rest=False)

plt.imshow(mark_boundaries(temp / 2 + 0.5, mask))
```

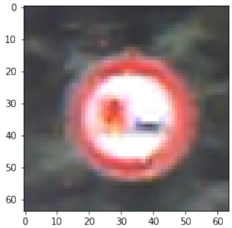

Figure: No Overtaking - Sample Image from Test-data

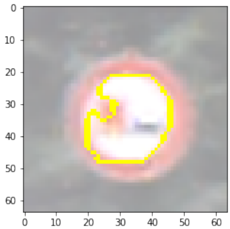

Figure: Prediction - showing the 5 Superpixels for *no overtaking*

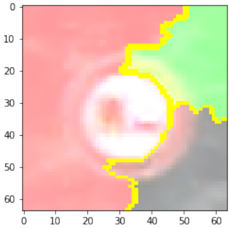

Figure: Prediction - showing the 4 most important Superpixels for *right of way crossing*

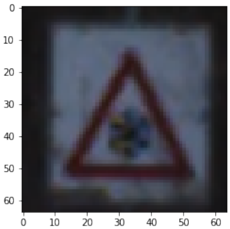

Figure: Frost - Sample Image from *Training*-data

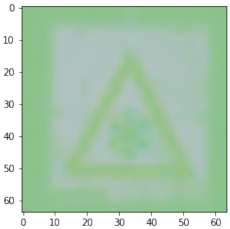

Figure: Prediction for *Frost*, a sign for overfitting

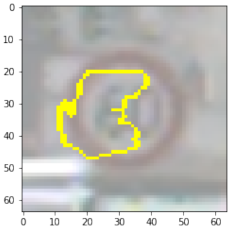

Figure: Prediction: 60 - only 6 is circled

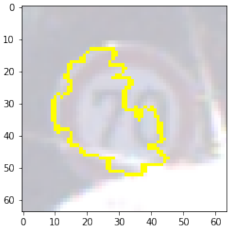

Figure: Prediction: 70 - only 7 is circled

The model seems to recognize numbers!

Let's have some fun!

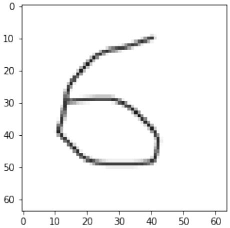

Figure: Only number 6, no Street Sign

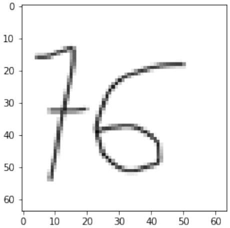

Figure: Number 76 - what will be predicted?

Let's have some fun!

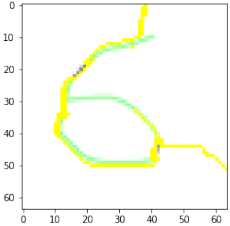

Figure: Number 6 - 78% *Speed Limit 60*

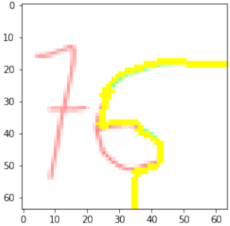

Figure: Number 76 - 99.9% *No Overtaking*

- Accuracy $\neq$ Quality
- most of the predictions *look good*
- trainings-data is heavily overfitted
- everything that is not a streetsign causes trouble
- there can (still) be much more *hidden* problems

Questions?

Referendums-Questions

- give an example of unreadable features and it's human-readable LIME-Interpretation
- name some measures to improve your model after using the explanations
- what is the main difference between LIME and ANCHOR