Explaining Models with LIME

Alicia Wirth and Saba Mateen

Agenda

1. Basics
2. Image Data
    1. Use Case
    2. Implementation
3. Tabular Data
    1. Use Case
    2. Implementation
4. Comparison
5. Conclusion

Basics

- Objective: Comparison InterpretML and AIX360 using **LIME**

  - Output is very fast
  - Can be applied to any machine learning model
  - Works for tabular, text and image data

Image Data

Use Case

- Classification of images according to seven members of a korean band
- Obtained pre-trained model

Image data set size: ~2.000
Classification Method: ResNet

- Pickle file containing results

Image Data

Use Case

Jungkook

⇒ Input: Image

⇒ Output: Predicted member the image portrays

Image Data

Implementation

- Input image
- Load pre-trained model
- Transform image to Pytorch tensor
- Explaining classification using LimeImageExplainer()
    - Apply mask to visualize classification decisions
    - Show mask and probability for each member

Image Data

Implementation

```
from aix360.algorithms.lime import LimeImageExplainer
from skimage.c
```


```
explainer = Li
explanation =
```


```
temp, mask = explanation.get_image_and_mask(explanation.top_labels[0], positive_only=True,
num_features=10, hide_rest=False)
plt.imshow(label2rgb(mask,temp, bg_label = 0), interpolation = 'nearest')
```

Tabular Data

Use Case

- Classification of images in form of tabular data according to seven members of a korean band
  - using result file acquired from previous use case as base

Tabular data set size: ~2.000
Classification Method: Random Forest + PCA

```
0,20090.jpg,2,"[554, 445, 757, 975]",0.9999998807907104,"(690, 789)","(1034, 775)","(809, 982)","(726, 1156)","(1037, 1141)"
1,20098.jpg,2,"[772, 49, 819, 1182]",0.9998756647109985,"(876, 501)","(1236, 493)","(976, 709)","(901, 963)","(1124, 970)"
3,20108.jpg,2,"[971, 88, 668, 846]",0.9368537068367004,"(1207, 422)","(1426, 436)","(1276, 504)","(1224, 722)","(1390, 723)"
4,20111.jpg,2,"[996, 50, 576, 741]",0.9998229146003723,"(1079, 376)","(1326, 347)","(1165, 511)","(1135, 670)","(1313, 652)"
5,20117.jpg,2,"[1049, 151, 498, 649]",0.9997857213020325,"(1110, 410)","(1321, 413)","(1124, 531)","(1092, 692)","(1273, 699)"
```

Tabular Data

Use Case

eye_l	eye_r	nose	mouth_l	mouth_r
"(690, 789)"	"(1034, 775)"	"(809, 982)"	"(726, 1156)"	"(1037, 1141)"

Hoseok

⇒ Input: Coordinates of facial features

⇒ Output: Predicted member corresponding to features

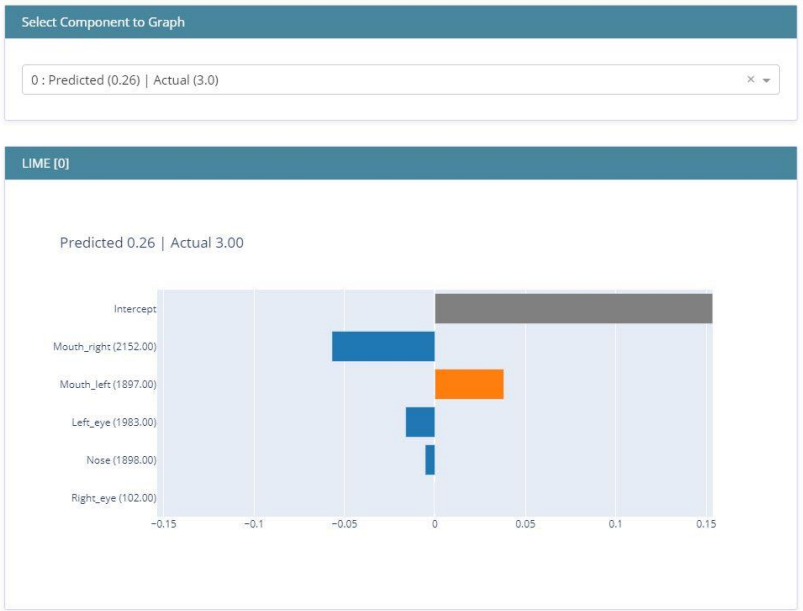
Tabular Data

Implementation

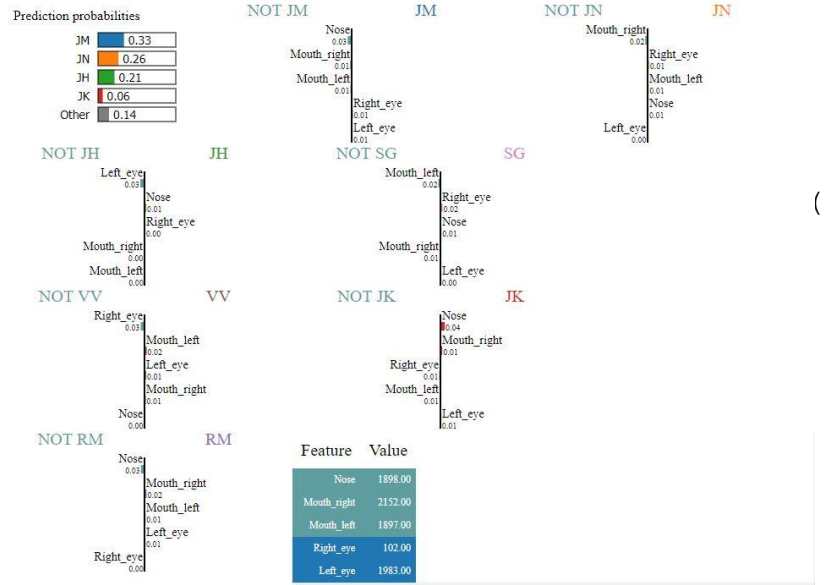
- Transform pickle to csv and modify file
- Input encoded coordinates of facial features
- Train model with training data
- Use LIME’s explanation methods for tabular data

Tabular Data
Implementation

InterpretML



AIX360




```
explanation.show_in_notebook(show_all=False)
```

Tabular Data
Comparison

InterpretML	AIX360
Always selects the probability of the second class to display	Seems to select the correct probability
Unclear what exactly is displayed	Helpful captions
Can only be applied for tabular data	Can be applied to image, text and tabular data
Documentation not good	Lots of resources

Conclusion

- AIX360 offers better fundamentals and functionalities to work with

  ⇒ easier for beginners

- InterpretML’s LIME still seems to be in development

  ⇒ lacks options to analyze models for different types of data

Sources

- Source Code: https://github.com/SabaMt/LIME-AIX-InterpretML.git

- https://www.onclick360.com/interpretable-machine-learning-with-lime-eli5-shap-interpret-ml/
- https://www.kaggle.com/choiseokhyeon/bts-crop2
- https://de.wikipedia.org/wiki/Datei:%E2%80%98LG_Q7_BTS_%EC%97%90%EB%94%94%EC%85%98%E2%80%99_%EC%98%88%EC%95%BD_%ED%8C%90%EB%A7%A4_%EC%8B%9C%EC%9E%91_(42773472410)_(cropped).jpg
- https://commons.wikimedia.org/wiki/File:Jeon_Jung-kook_at_Golden_Disk_Awards,_5_January_2019_07.jpg