Master Thesis - Coursework

Title: Exploration of applicability of explainable artificial intelligence techniques for sentiment analysis applied for English language

by Rohan Kumar Rathore

Advisor: Dr. Anton Kolonin

Scientific workshop "Big Data Analytics"
Novosibirsk State University, Russia

May 14, 2020

Table of Content

Introduction

- Artificial intelligence : Artificial agents achieving goals smartly
- Machine learning : Algorithmic models responsible for smartness
- Explainable artificial intelligence : Techniques to explain the models

Abstract highlights

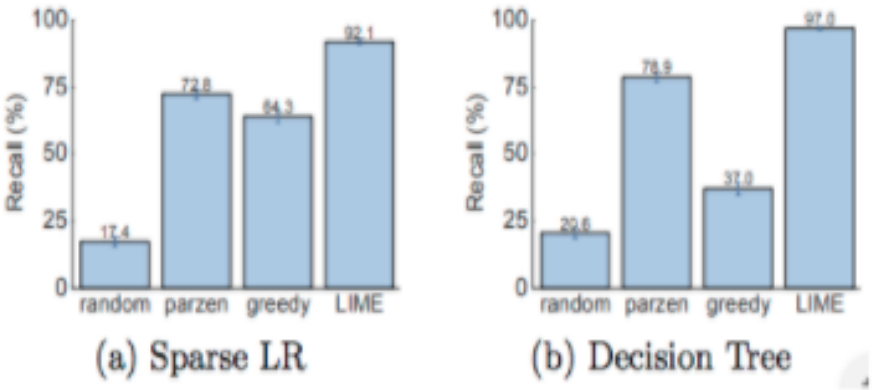
Explainable artificial intelligence (XAI) techniques for sentiment analysis

- Sentiment analysis model development
  - Data: IMDB movie reviews
  - Model: multisentiment, context based, neutral-mixed capable, unbiased
- Study of proven XAI techniques
  - Explaining with surrogates - LIME, SmoothGrad
  - Explaining with local perturbations - SA, PDA
  - Propagation-based approaches - LRP
  - Meta-explanations - SpRAy
- Exploration of proven XAI techniques
  - Local interpretable model-agnostic explanations (LIME)
  - Layer-wise relevance propagation (LRP)
- Possibility of exploring unproven XAI technique
  - Conjunctive normal form/ Disjunctive normal form for NN
  - Bayesian Neural Network

Exploration of LIME for sentiment analysis - I

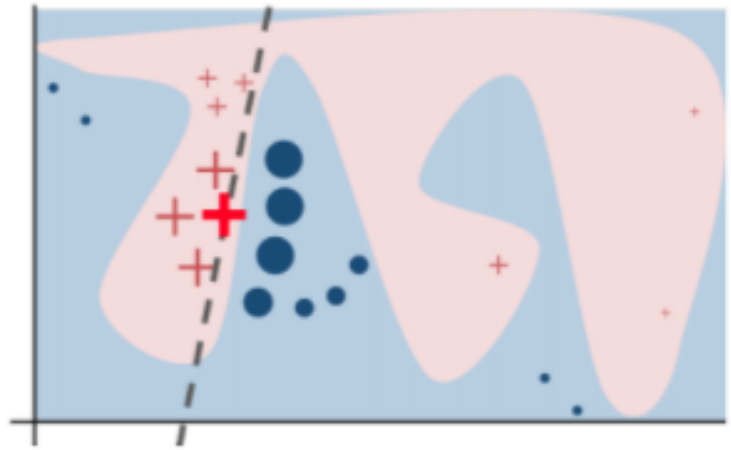
Local interpretable model-agnostic explanations

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



(a) Sparse LR (b) Decision Tree

Recall on truly important features



Toy example to present intuition for LIME.

Exploration of LIME for sentiment analysis - II

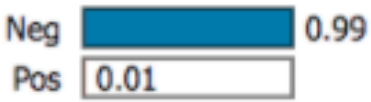
Local interpretable model-agnostic explanations

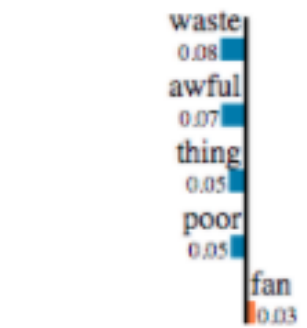
Text with highlighted words

Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write to much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcedly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book.lbr /llbr /lFans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.lbr /llbr /l2/10

```
SVM model prediction : Neg sentiment
True value : Neg sentiment
```

Prediction probabilities





Exploration of LRP for sentiment analysis - I

Layer-wise relevance propagation

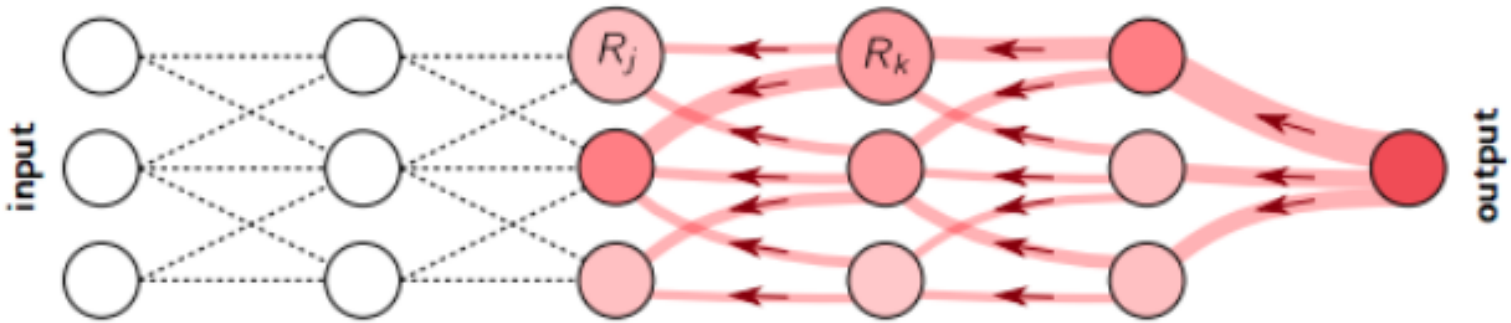


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

Basic Rule

$$R_j = \sum_k \frac{a_j \cdot \rho(w_{jk})}{\epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})} R_k,$$

Epsilon Rule

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k$$

Gamma Rule

Exploration of LRP for sentiment analysis - II

Layer-wise relevance propagation

- **Glove Embedding**
- **Simple Dense Keras network**
- **Accuracy - 91%**

```
True class: Negative review
Predicted class: Negative review ( 0.9515501 ) [0=True, 1=False]

Top 10 - Positive Contribute
[('ed', 0.23428376), ('running', 0.19387451), ('believe', 0.1913349), ('not', 0.18128063), ('plo
t', 0.16896152)]

Top 10 - Negative Contribute
[('him', -0.1360347), ('film', -0.14448667), ('who', -0.16167434), ('introduces', -0.17170446),
('ray', -0.19200623)]

Text:
 I was ed when couldn see this one when it was screening at the Philly Film Fest last year so when
saw that it was going to be on cable tonight put it on remind as soon as could So was it worth the
wait Well let backtrack tad as have yet to give you the plot Sean Crawley is young man who doesn k
now what his path in life is Enter Duke George Wendt who introduces him to his boss Ray Danny Bald
win One night Ray totally hammered asks Sean to off the guy that they had Sean following around An
d it goes on from there Which leads me back to the question posed Was it worth the wait Yes and no
the buildup was pretty good and George Wendt stole the movie for me He just took the ball and ran
with it But it nowhere near as violent as was led to believe and somewhere along the movies runnin
g time the ball is not only dropped but fumbled and taken in the other direction know where this p
oint happened exactly but can say without spoiling the film But needless to say it happened The en
ding doesn save the film either Poor Stuart Gordon nothing can be good like Re animator or Castle
Freak My Grade CWhere saw it Showtime ExtremeEye Candy Kari Wuhrer shows her ta tas in one fantasy
and then in the next more ta tas and it pans down and OH MY GOD MY EYES MY EYES
```

Next steps

- Improve the model to meet all the functional goal [in-progress]
- Explore the (CNF/ DNF) technique [in-progress]
- Explore Bayesian Neural Networks for sentiment analysis

References

- Samek W., Müller KR., Towards Explainable Artificial Intelligence, LNCS, vol 11700, Springer (2019)
- Marco TR., Sameer S., Carlos G., "Why Should I Trust You?" Explaining the Predictions of Any Classifier, International Conference on Knowledge Discovery and Data Mining 1135-1144 (2016)
- Montavon G., Binder A., Lapuschkin S., Samek W., Müller KR., Layer-Wise Relevance Propagation: An Overview, LNCS, vol 11700, Springer (2019)
- Anton K., Artificial Intellegence - state of affairs and perspective, Papers, aigents.com (2019)