

Understanding Emotion Classification Through Shapley Values

Noah Lehmann

University of Applied Sciences Hof

Data Mining and Machine Learning, 23.12.2021

Table of Contents

Introduction

Solution

Shapley Values

Introduction to Emotion Analysis

Why?

- ▶ Recognize major changes in behaviour
- ▶ Pre-classify hate speech
- ▶ Detect depressive tendencies

How?

- ▶ Well-balanced data from multiple cultures
- ▶ Detect tendencies of word combinations
- ▶ Use on social media, text messengers and mail services

Text Classification

What data-set is used? → Text collection with five emotions (anger, fear, sadness, neutral, happy)

1. 5 emotions → further insights
higher complexity
2. Only detect sadness → specialized
less data required

Solution

Text Preparation

- ▶ Reduce Data-set to binary classification
- ▶ Sample all non-sad-classified rows
- ▶ Remove Stopwords
- ▶ Reduce word complexity
- ▶ Tokenize Texts

Primary Goal: Reduce complexity and feature set

Solution

Classification

Attempt and tweak multiple classifiers

- ▶ Naive Bayes
- ▶ Random Forest
- ▶ Neural Network

How do we get insights into the more complex models?

- ▶ Some have good explainability
- ▶ For others: **Shapley Values**

Shapley Values

Introduction

What is the primary objective?

- ▶ Explain any model

It is impossible to trust a machine learning model without understanding how and why it makes its decisions and whether these decisions are justified [6].

Shapley Values

Background

Game theory background:

- ▶ A set of players have contributed differently to achieving a prize
- ▶ How can the payout be calculated according to the contribution?

Real life example for application of Shapley Values:

- ▶ Sharing a taxi to different destinations
- ▶ How can the bill be distributed among the passengers?

Shapley Values

Application

Application in emotion analysis through texts

- How much did each word contribute to the models/ classifiers output?

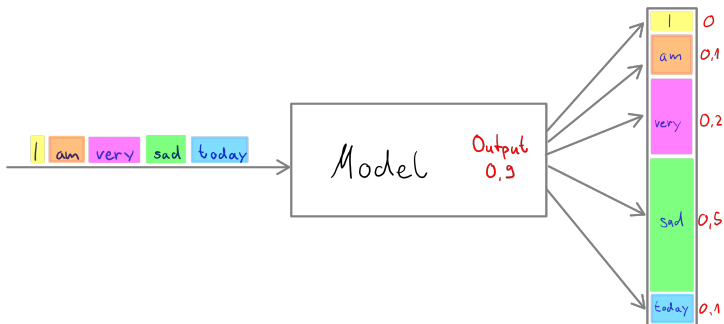


Figure: Blackbox feature contribution view

Shapley Values

Binary Classification

Comparison of weighed contribution of a feature-set to other inputs prediction results

► Example: Binary Classification



Figure: Feature contributions with reference to average model output

Shapley Values

Mathematical Definition

$N = \text{Set of Attributes}, n = |N|, v = \text{function}$

$$\Phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

Shapley Values

Example

Black Box Classifier for recognition of sentences as greeting

- ▶ Classifier model $\rightarrow v$, Example: "Hello there"
- ▶ $h = \text{"Hello"}, t = \text{"there"}$
- ▶ $v(h) = 0.5, v(t) = 0.25$
- ▶ $v(ht) = 1 \rightarrow$ (classified as greeting)
 $\{h, t\} \rightarrow \{h\}, \{h, t\}$ or $\{t\}, \{h, t\}$

$$v(\{t\}) * \frac{1}{n!} = \frac{1}{4} * \frac{1}{2}, (v(\{h, t\}) - v(\{h\})) * \frac{1}{n!} = (1 - \frac{1}{2}) * \frac{1}{2}$$

$$\rightarrow \Phi_t(v) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$v(\{h\}) * \frac{1}{n!} = \frac{1}{2} * \frac{1}{2}, (v(\{h, t\}) - v(\{t\})) * \frac{1}{n!} = (1 - \frac{1}{4}) * \frac{1}{2}$$

$$\rightarrow \Phi_h(v) = \frac{3}{8} + \frac{1}{4} = \frac{5}{8}$$

$$v(\{t\}) * \frac{1}{n!} = \frac{1}{4} * \frac{1}{2}, (v(\{h, t\}) - v(\{h\})) * \frac{1}{n!} = (1 - \frac{1}{2}) * \frac{1}{2}$$

$$\rightarrow \Phi_t(v) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$v(\{h\}) * \frac{1}{n!} = \frac{1}{2} * \frac{1}{2}, (v(\{h, t\}) - v(\{t\})) * \frac{1}{n!} = (1 - \frac{1}{4}) * \frac{1}{2}$$

$$\rightarrow \Phi_h(v) = \frac{3}{8} + \frac{1}{4} = \frac{5}{8}$$



Shapley Values

Observations

- ▶ v can be any function \rightarrow model-agnostic technique
- ▶ Feature contribution comparison through different models possible

Shapley Values

Limitations

- ▶ Selection of subsets of features necessary
 - ▶ Exponential scaling in number of features
 - ▶ Attempts to reduce complexity through approximations for known models
- ▶ Dependent of model reaction to unrealistic input
 - ▶ Example: longitude and latitude delivered as separate features in housing price estimation
 - Housing prices for lake sites could influence accuracy of approximated shapley values
 - ▶ Exploitable weakness

Bibliography I

- [1] Marko Cotra. Making Sense of Shapley Values. [Online; accessed 21.12.2021]. 2019. URL: <https://towardsdatascience.com/making-sense-of-shapley-values-dc67a8e4c5e8>.

- [2] Google Developers. Machine Learning Glossary - Google Developers. [Online; accessed 10.02.2022]. 2022. URL: https://developers.google.com/machine-learning/glossary/#recurrent_neural_network.

- [3] Scikit-learn Developers. Naive Bayes - scikit-learn 1.0.2 documentation. [Online; accessed 10.02.2022]. 2022. URL: https://scikit-learn.org/stable/modules/naive_bayes.html.

Bibliography II

- [4] Scikit-learn Developers.
sklearn.ensemble.RandomForestClassifier - scikit-learn 1.0.2 document
[Online; accessed 10.02.2022]. 2022. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

- [5] Cambridge Dictionary.
Game Theory, meaning in the Cambridge English Dictionary.
[Online; accessed 07.02.2022]. 2022. URL: <https://dictionary.cambridge.org/dictionary/english/game-theory>.

- [6] Divya Gopinath. Picking an explainability technique. [Online; accessed 29.11.2021]. 2021. URL: <https://towardsdatascience.com/picking-an-explainability-technique-48e807d687b9>.

Bibliography III

- [7] Divya Gopinath. The Shapley Values for ML Models. [Online; accessed 24.11.2021]. 2021. URL: <https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1>.
- [8] Matthew O. Jackson, Kevin Leyton-Brown, and Yoav Stoham. GTO-7-03: The Shapley Value. [Online; accessed 21.12.2021]. 2013. URL: <https://www.youtube.com/watch?v=qcLZMYPdpH4>.
- [9] Scott Lundberg. SHAP Kernel Explainer - SHAP latest documentation. [Online; accessed 10.02.2022]. 2022. URL: <https://shap-lrjball.readthedocs.io/en/latest/generated/shap.KernelExplainer.html>.