DARTMOUTH

Institute for Security Technology and Society at Dartmouth

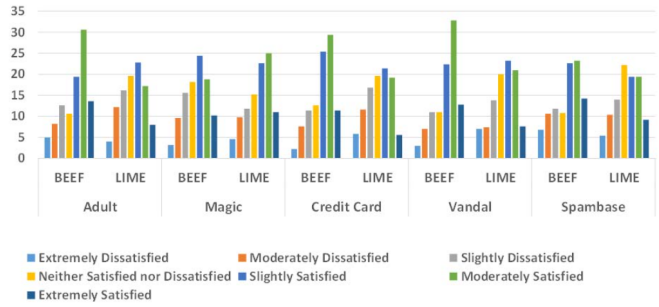
Explainable AI

Understanding is the basis for acceptance, trust and adoption of recommendations. We accept and adopt advice from real estate brokers on a house to buy because of trust in the real estate agent. A similar sense of acceptance, trust, and adoption is critical as ML techniques increasingly make high stakes decisions. Today, even ML experts have difficulty communicating why a classification or regression model made a certain prediction to domain experts (e.g. bankers, lawyers).

To help people better understand the insights of predictions made by ML models, we built explainers to extract the most important features behind model predictions.

BEEF (Balanced English Explanations of Forecasts)

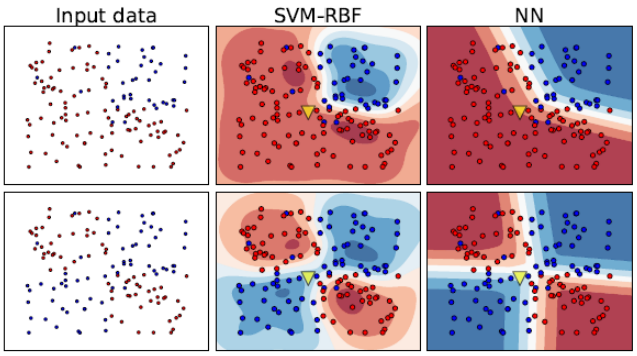
The BEEF is proposed to address the problem of automatically extracting balanced explanations from predictions generated by any classifier, which include not only why the prediction might be correct but also why it could be wrong. BEEF can generate such explanations in natural language. After showing that the problem of generating explanations is NP-complete, we focus on the development of a heuristic algorithm, empirically showing that it produces high-quality results both in terms of objective measures—with statistically significant effects shown for several parameter variations—and subjective evaluations based on a survey completed by 100 anonymous participants recruited via Amazon Mechanical Turk.



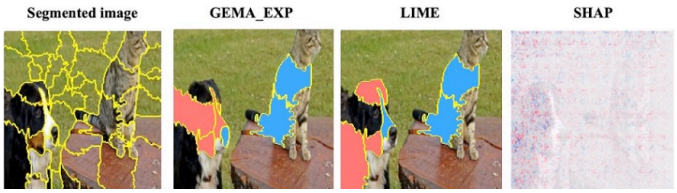
Percentage of positive, neutral, and negative responses obtained for BEEF and LIME (a comparison method) on the MTurk evaluation.

GEMA (Gradient based Explanation of Model predictions with anecdotes)

GEMA explain predictions by quantifying how features work together to generate a prediction. GEMA consists of a technical explanation generated by GEMA_EXP and anecdotes from GEMA_ANEC. GEMA_EXP uses an efficient algorithm based on approximate directional gradient to solve a challenging optimization problem under partial information. Technical explanations with feature combinations are then automatically generated to explain model predictions. Further, GEMA_ANEC supplements a technical explanation with a set of anecdotes. We show that GEMA outperforms past work on six diverse (relational, text, and image) datasets quantitatively and qualitatively. This work is currently under review.



Example datasets and classifiers. Each row contains the original dataset, the trained SVM-RBF classifier and Neural Network classifier. Class 1 (resp. 2) points are shown in blue (resp. red). The yellow triangle is a new data point which needs to be explained.



An images with the segmentation and prediction explanations.

Index	Review and selected important words with explainers
1	**GEMA**: my first visit to this fuel pizza was very disappointing .
	LIME: my first visit to this fuel pizza was very disappointing.
	SHAP: my first visit to this fuel pizza was very disappointing .
2	**GEMA**: the line was a bit long , but it moved fairly quickly .
	LIME: the line was a bit long , but it moved fairly quickly.
	SHAP: the line was a bit long , but it moved fairly quickly.
3	**GEMA**: i also ordered guacamole and was very disappointed .
4	**GEMA**: like the extra pico , guacamole , and rice.

Example of review sentiment prediction and GEMA explanation. Predictions on sentences 1 & 2 are respectively 0.0001 and 0.4769; on sentence 3 & 4 are respectively - 0.9916 and 0.0031

Additional Information

References

1. Sachin Grover, Chiara Pulice, Gerardo I. Simari, and V.S. Subrahmanian. BEEF: Balanced English Explanations of Forecasts. *IEEE Transactions on Computational Social Systems, (2019), 350-364, 6(2).*

PARTICIPANTS

Lead: V.S. Subrahmanian

Sachin Grover, Chiara Pulice, Gerardo I. Simari, Yanhai Xiong, Dongkai Chen.

Dartmouth
Security and Artificial
Intelligence Laboratory