DARTMOUTH

Institute for Security Technology and Society at Dartmouth

Explainable AI

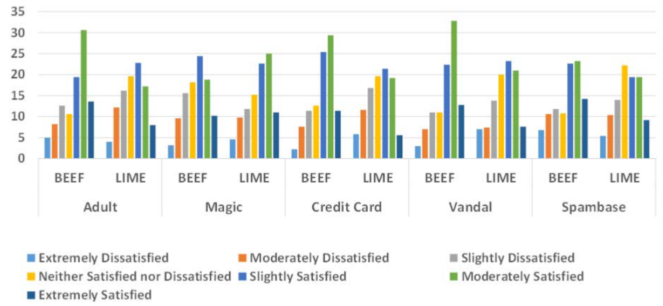
Understanding is the basis for acceptance, trust, and the adoption of recommendations. We accept and adopt advice from real estate brokers on which house to buy because we trust the real estate agent. A similar sense of trust is critical as ML techniques are increasingly used to make high stakes decisions. However, even ML experts have difficulty communicating with relevant stakeholders (e.g., government officials, bankers, lawyers) about why a classification or regression model made a specific prediction.

To help people better understand the insights of predictions made by ML models, we built a series of systems to extract the most critical features behind model predictions.

BEEF (Balanced English Explanations of Forecasts)

BEEF is an approach used to address the problem of automatically extracting *balanced* explanations from predictions generated by any classifier, which means the explanations include not only why the prediction might be correct but also why it could be wrong. BEEF can generate such explanations in natural language. We first show that the problem of generating explanations is NP-complete. Then, we focus on the development of a heuristic algorithm and empirically demonstrate its performance. Results show that BEEF produces high-quality explanations in terms of both 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.

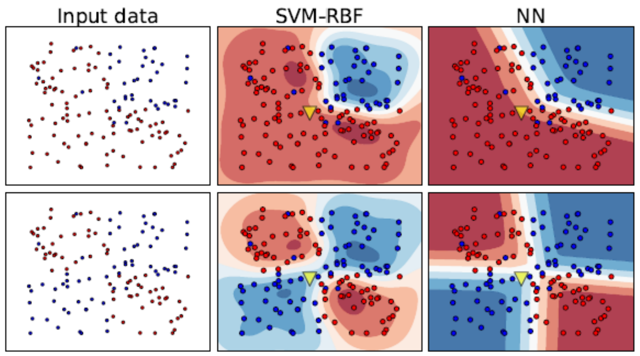
BEEF significantly improves the satisfaction of users comparing to LIME across all five datasets.



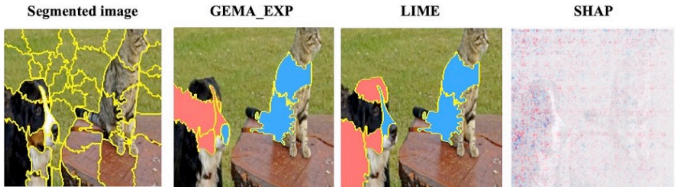
Percentage of positive, neutral, and negative responses that were obtained for BEEF and LIME (a comparison method) from our MTurk evaluation.

GEMA (Gradient-based Explanation of Model predictions with Anecdotes)

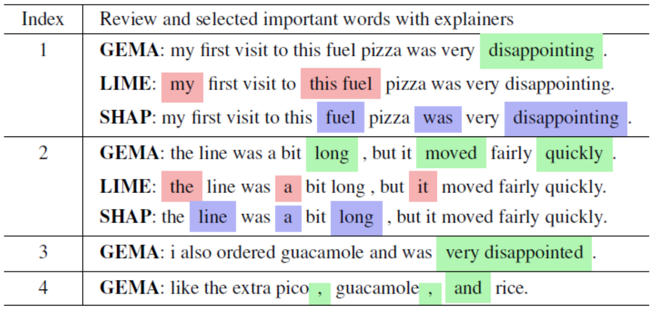
GEMA explain predictions by quantifying how features work together to generate a forecast. GEMA consists of a technical explanation computed by GEMA_EXP and anecdotes from GEMA_ANEC. GEMA_EXP uses an efficient algorithm based on the approximate directional gradient to solve a challenging optimization 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 datasets (including relational, text, and image datasets) quantitatively and qualitatively. This work is currently under review. More details will be available upon its acceptance.



Example datasets and classifiers. Each row contains the original dataset, a trained SVM-RBF classifier, and a 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

GEMA successfully figure out important combinations of features in image, text and relational datasets.

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