Effect of Sampling on Robustness of LIME

Trustworthy AI Systems
Instructor: Dr. C. Anantaram

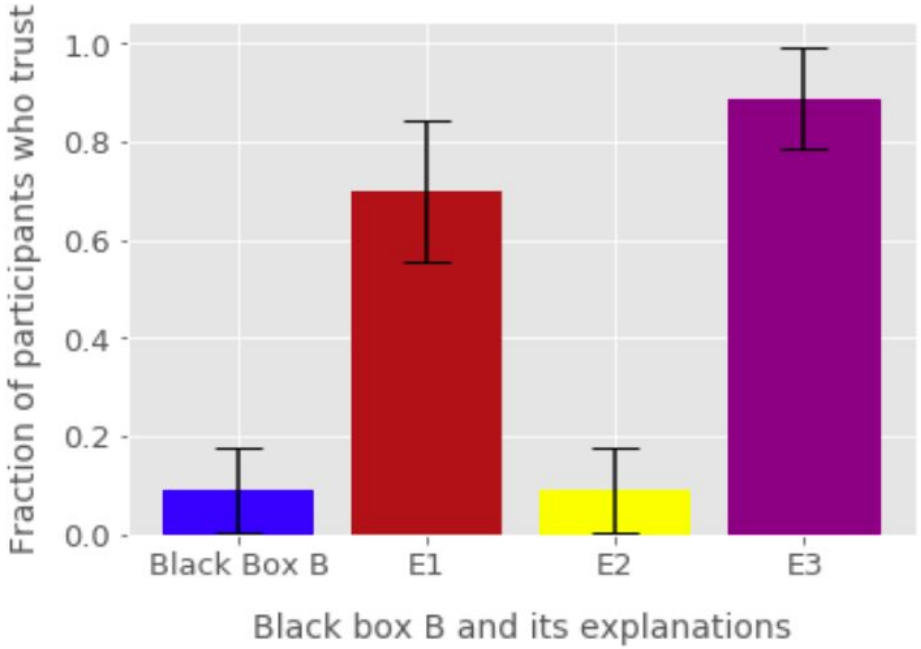
Sumedha Chugh - PhD21123
Sara Moin - PhD21035

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Outline

- Why do we need Robust Explanations
- How LIME works
- Literature Review
- Motivation
- Methodology
- Experiments Performed
- Results
- Conclusion & Future Scope
- References

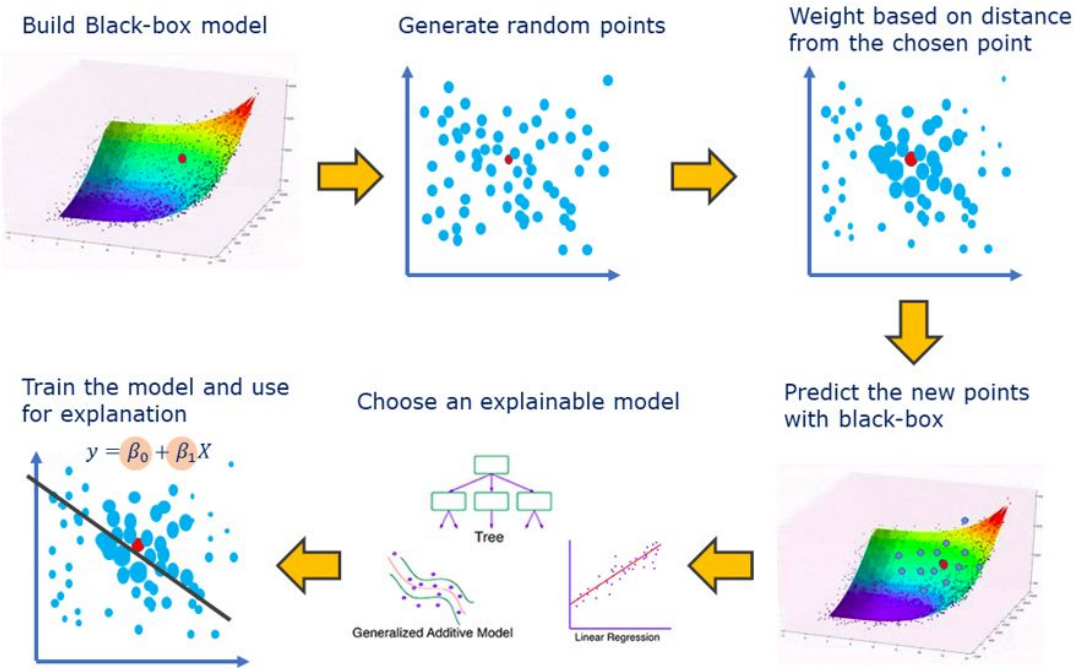
Why do we need Robust Explanations



Trustworthiness and Robustness [6]

- E1: Excludes both prohibited & Desired features
- E2: Includes both prohibited & Desired features
- E3: Includes prohibited & Excludes Desired features

How LIME works

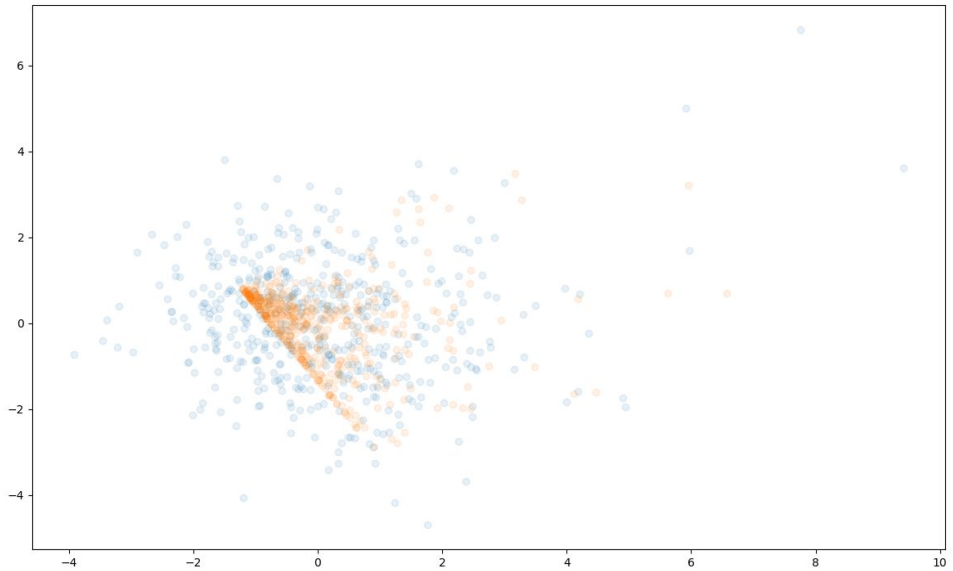


Working of LIME [13]

Literature Review

Work	Technique
[1]	Demonstrate that post hoc explanations techniques that rely on input perturbations, such as LIME and SHAP, are not reliable by crafting an adversarial attack
[2]	Makes LIME more robust by training a VAE network for generating sampling
[9]	Shows lack of stability and robustness in LIME & SHAP due to distribution shifts, generates robust and stable explanations of black box models based on adversarial training.
[10]	Introduces SmoothGrad which is a gradient based method, and a variant of LIME (perturbation based)
[11]	Introduces metrics to quantify robustness,demonstrate that current methods do not perform well according to these metrics, improves robustness by adversarial training
[12]	Shows sampling in these explanation methods by using data generators, prevents malicious manipulations.

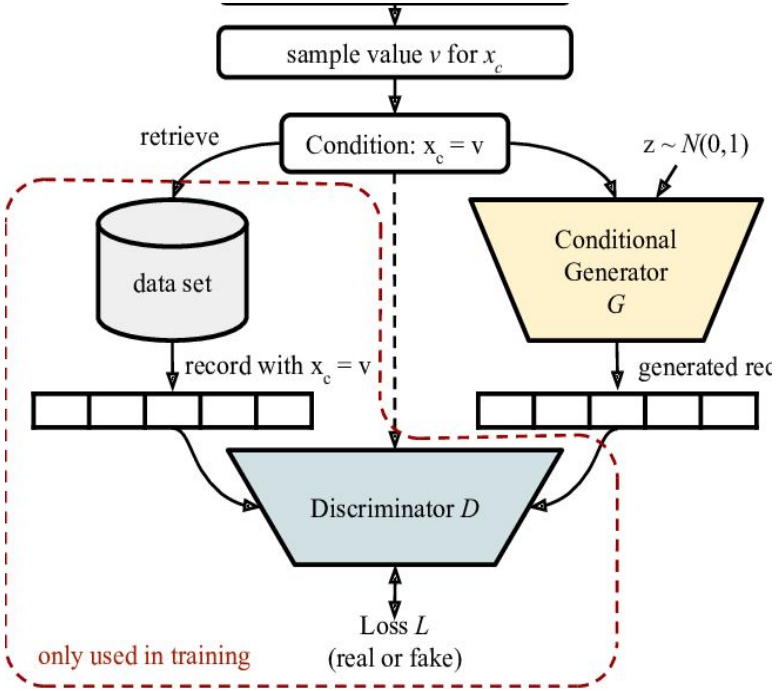
Motivation



PCA applied to the COMPAS dataset (blue) & its perturbations generated by random sampling (orange)

- Distribution of perturbations generated are out of distribution from distribution of original data
- In perturbation based methods basic cause of poor robustness is random sampling
- Goal is to improve robustness through better sampling such that new perturbed data set generated has similar distribution as training data

Methodology



CT-GAN Model Architecture [14]

- Use Conditional Tabular GAN (CTGAN) model to generate more realistic synthetic data
- Captures heterogeneity of tabular data
- Contrary to image data, where pixel values normally follow a Gaussian-like distribution, continuous features in tabular follows multimodal distributions
- Uses mode-specific normalization by VGM (Variational Gaussian Mixture) model
- Issues:
  - Sparsity
  - Underrepresented features
- Introduces a Conditional Generator

Experiments Performed

Black Box Adversarial Attack [1]

- Goal:
  - Create an adversarial classifier that behaves like the original classifier (perhaps extremely discriminatory) on the input data points, but looks unbiased and fair on the perturbed instances, thus effectively fooling LIME

- Dataset: COMPAS, Size: 6172, Features: criminal history, demographics, risk score, jail and prison time, Positive Class: High Risk, Sensitive Feature: African-American

- Intuition
  - Real world data follows a distribution X_{dist} ,
  - Adversarial classifier that exhibits biased behavior on instances sampled from X_{dist} , remain unbiased on instances that do not come from X_{dist}
  - Feature importances output by LIME relies on perturbed instances (OOD), the resulting explanations will make the classifier designed by the adversary look innocuous

Experiments Performed

Black Box Adversarial Attack [1]

- Input:
  - The biased classifier f
  - An input dataset X that is sampled from the real world input data distribution X_{dist}
- Output:
  - Adversarial classifier e that behaves exactly like f when making predictions on instances sampled from X_{dist} , but will not reveal the underlying biases of f when probed with LIME
- Mathematically, $e(x) = f(x)$, if $x \in X_{dist}$
  $\psi(x)$, otherwise

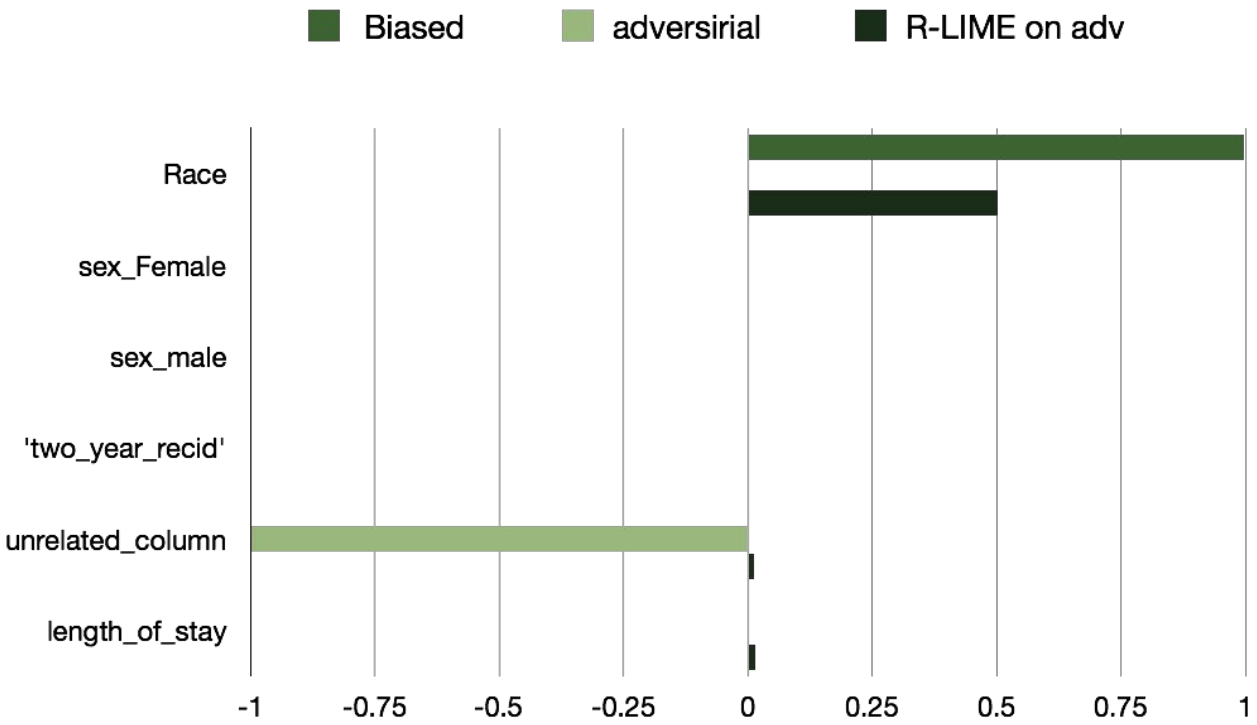
  where, e(x) is adversarial classifier, f(x) is biased model, ψ(x) is unbiased model

Experiments Performed

Process flow

- f: discriminatory classifier: makes predictions purely based on sensitive attribute
- Generate x_p by adding random noise sampled from N (0, 1) to each feature value
- Train the OOD classifier
  - Dataset X: label False: indicating not OOD samples,
  - Dataset: Xp: label True: Indicating OOD samples unless they are already in X
  - Train a classifier (random forest with 100 tree estimators) on the combined dataset $X \cup Xp$ and their corresponding class labels
- Build the unbiased classifier ψ by constructing synthetic uncorrelated features that have zero correlation with sensitive attributes
- Use vanilla LIME and LIME with GAN (num_features = 1000) sampling for generating explanations
- Evaluate robustness by checking where race shows up in top 3 important features if higher importance then model is robust to attack o/w not

Results



- More Robust to adversarial attack
- Effective only for tabular data

Importance Scores of various features for LIME & Robust LIME on biased & Adversarial models

Results

- Explanation using LIME on biased f:
  - [('race', 0.9986343456272183), ('sex_Female', 2.0027981228438182e-05), ('unrelated_column', -1.791320273434123e-05)]

- Explanation using LIME on adversarial model:
  - [('unrelated_column', -0.9981443627773987), ('sex_Male', 0.0012015987966744917), ('two_year_recid', -0.0005130675716730023)]

- Explanation using LIME with GAN on adversarial model:
  - [('race', 0.5016209270590931), ('length_of_stay', 0.016096179525062 17), ('unrelated_column', 0.012876938185426817)]
  - Percentage: About 43% times race was in top 3 features

Conclusion & Future Scope

- Analysed structure of popular perturbation based methods
- Literature review to know the underlying cause for poor robustness of LIME
- Changed the sampling technique for LIME from random sampling to GAN based for tabular data
- Analysed this method along with LIME for robustness on a popular adversarial attack
- LIME with CTGAN generated samples is more robust to adversarial attacks than LIME with random sampling as it generates samples that are in-distribution

- Try other sampling methods such as data generators, VAEs etc. and compare robustness of these methods
- Extend this work to other perturbation based methods like SHAP

References

[1] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830

[2] Vreš, D. and Robnik Šikonja, M. (2020) Better sampling in explanation methods can prevent dieselgate-like deception Submitted to International Conference on Learning Representations https://openreview.net/forum?id=s0Chrsstpv2

[3] Aditya Saini and Ranjitha Prasad. 2022. Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 599–608. https://doi.org/10.1145/3514094.3534191

[4] Saito, S., Chua, E., Capel, N., & Hu, R. (2020). Improving LIME Robustness with Smarter Locality Sampling. ArXiv. /abs/2006.12302

References

[5] Visani, Giorgio, et al. "Statistical stability indices for LIME: obtaining reliable explanations for Machine Learning models." arXiv preprint arXiv:2001.11757 (2020)

[6] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[8] Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. ArXiv. /abs/1705.07874

References

[10] Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S. & Lakkaraju, H.. (2021). Towards the Unification and Robustness of Perturbation and Gradient Based Explanations. Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research 139:110-119 Available from https://proceedings.mlr.press/v139/agarwal21c.html.

[11] Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. ArXiv. /abs/1806.08049

[12] D. Vreš and M. Robnik-Šikonja, “Preventing deception with explanation methods using focused sampling,” Data Mining and Knowledge Discovery, Dec. 2022, doi: 10.1007/s10618-022-00900-w.

[13] https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe

[14] Borisov, Vadim & Leemann, Tobias & Seßler, Kathrin & Haug, Johannes & Pawelczyk, Martin & Kasneci, Gjergji. (2021). Deep Neural Networks and Tabular Data: A Survey.

Thank you !

Open to Feedback and Questions

saram@iiitd.ac.in
sumedhac@iiitd.ac.in