Model explainability and interpretability: A practical guide

Bianca Furtuna, @Fur_Bi

Applied Machine Learning Scientist

Microsoft

Model Interpretability

Transparency

Explanations

Causality

Trust

Easy to understand by humans

Interpretable decisions

Model Interpretability - Terms

Local vs. global explanations

Model-agnostic vs. Model-specific

Explanation coverage = region in input space where the explanation applies

Which of the following algorithms are interpretable?

K-nearest neighbour

Linear Regression

Neural Network

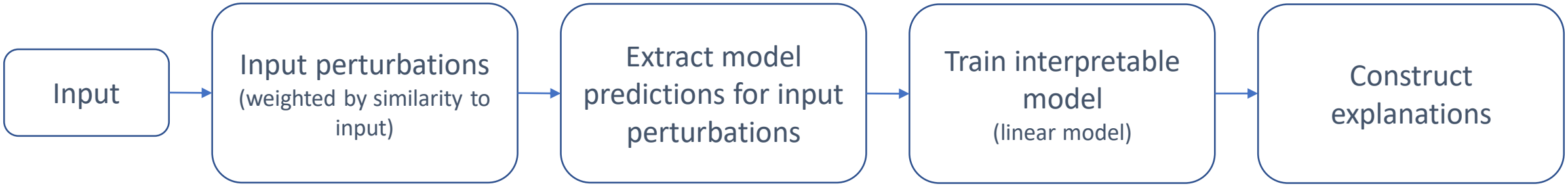
Decision tree classifier

Support Vector Machine

Focus of this talk

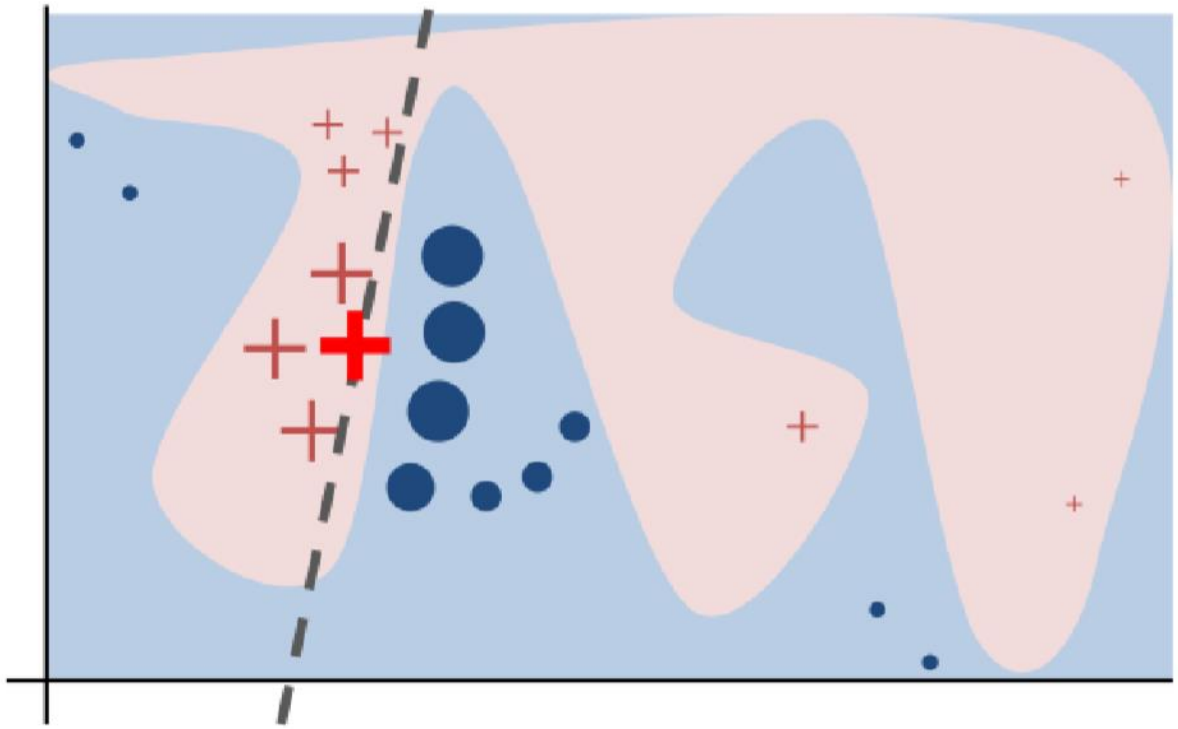
LIME – quick intro

"identify an **interpretable** model over the interpretable representation that is **locally faithful** to the classifier"



Source: Ribeiro, M. T., Singh, S., and Guestrin, C. "Why should i trust you?: Explaining the predictions of any classifier". In SIGKDD, pp. 1135–1144. ACM, 2016. https://arxiv.org/abs/1602.04938

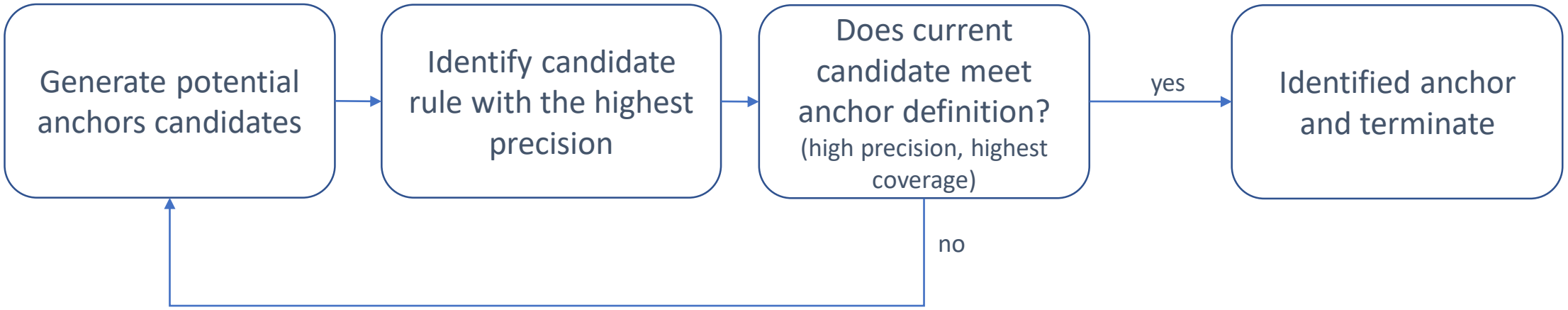
LIME – quick intro



Source: Ribeiro, M. T., Singh, S., and Guestrin, C. "Why should i trust you?: Explaining the predictions of any classifier". In SIGKDD, pp. 1135–1144. ACM, 2016. https://arxiv.org/abs/1602.04938

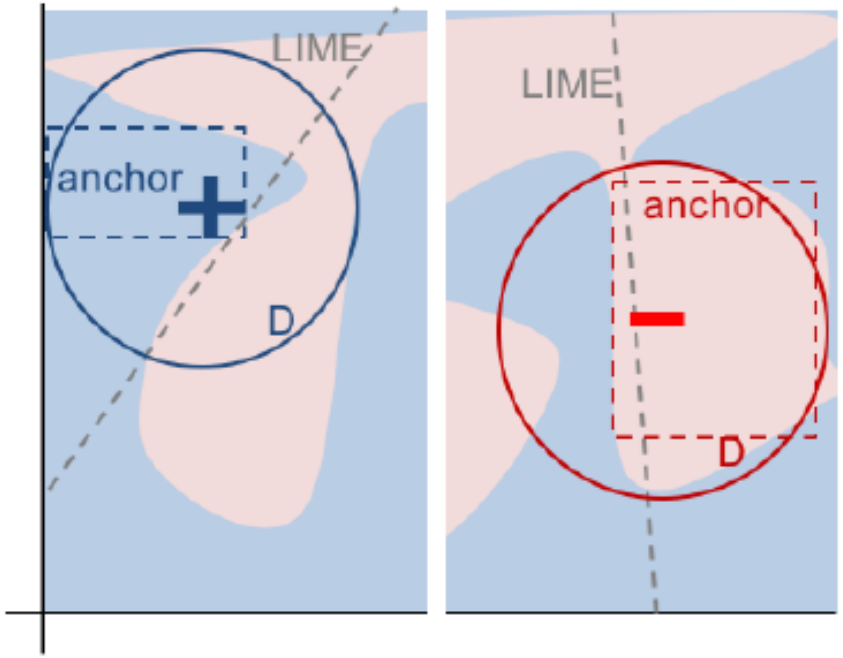
Anchors – quick intro

rule that “anchors” a prediction locally such that in instances that the anchors holds, the prediction is consistent



Source: Ribeiro, M. T., Singh, S., and Guestrin, C. “Anchors: High-Precision Model-Agnostic Explanations”. Proc. of 32nd Conference on Artificial Intelligence (AAAI), pages 1527–1535.https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

Anchors vs. LIME

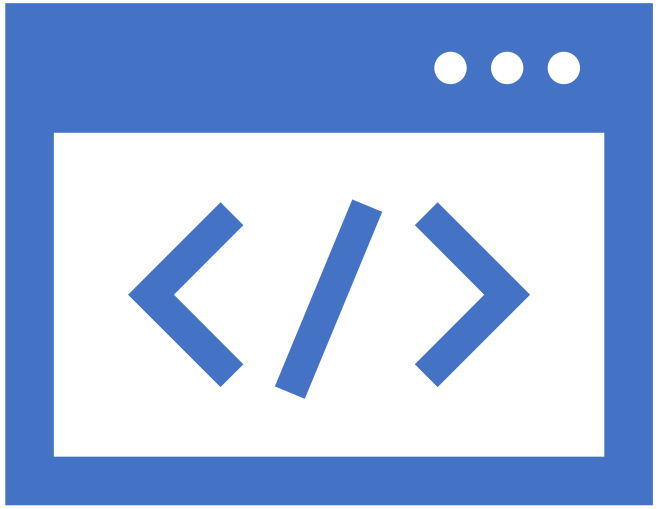


Source: Ribeiro, M. T., Singh, S., and Guestrin, C. “Anchors: High-Precision Model-Agnostic Explanations”. Proc. of 32nd Conference on Artificial Intelligence (AAAI), pages 1527–1535. https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

Practical uses

Data type	Perturbations	Interpretable representation
Tabular data	Fix some features and sample the rest of the row	Features in the dataset
Images	Turn Super-pixels on/off	Presence or absence of super-pixels
Text classification	Fix some tokens and replace the rest by random words	Presence/absence of tokens (words) in the sentence

Some actual examples …



Practical tips

Model debugging
and validation

Model understanding

Pros & Cons

- Easy to use with existing models
- Give an intuition about the model behaviour locally
- Help model understanding before deployment
- Good model validation tools in addition to existing methods

- LIME builds a linear model (the local boundary can be highly non-linear)
- LIME does not return any coverage information
- In practice they don’t work well for all examples
- No global understanding of what the model has learnt
- Need a perturbation distribution to be defined for every domain

Resources

- “The mythos of model interpretability”, Zachary Lipton : https://arxiv.org/abs/1606.03490
- SHAP : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
- LIME github: https://github.com/marcotcr/lime
- Anchors github: https://github.com/marcotcr/anchor
- Skater = open source unified framework for model interpretability: https://datascienceinc.github.io/Skater/index.html
- Interpretability resources list on github: https://github.com/lopusz/awesome-interpretable-machine-learning
- InterpretML is an open-source package for training interpretable models and explaining blackbox system: https://github.com/microsoft/interpret

Thank you!