Explanation Models

An Area Unfortunately Dominated by Computer Scientists

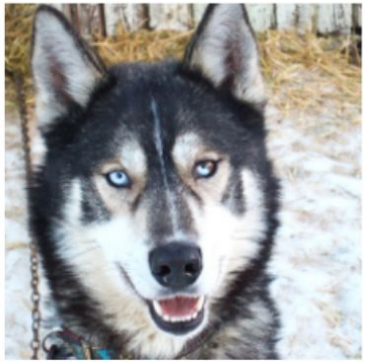
What is an Explanation Model?

- Model that explains a different model

→ ask a **better** question

What is the Goal of your Explanation Model?

- Understanding
- Trust
- Feature Selection
- Actionable Advice



(a) Husky classified as wolf

(b) Explanation

How Do We Differentiate Explanation Models?

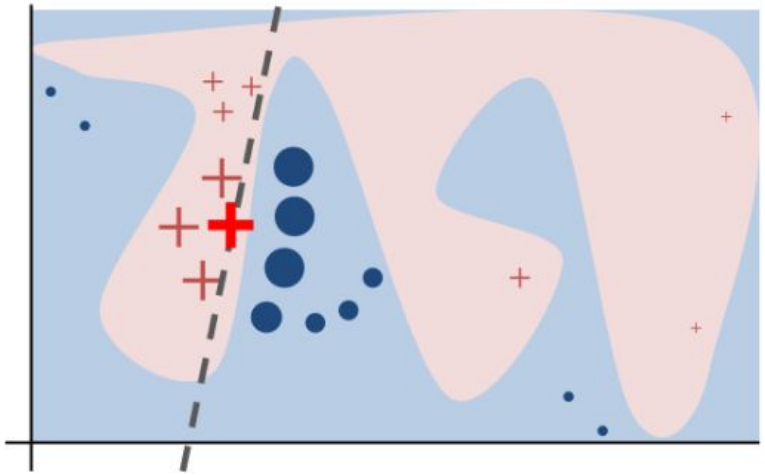
- Intrinsic or Post-Hoc?
- Model-Specific or Model-Agnostic?
- Local or Global?

https://gilberttanner.com/blog/introduction-to-machine-learning-model-interpretation

Local Model-Agnostic Explanation Models

- Tangent Approximation
- LIME
- SHAP
- Neighbourhood SHAP

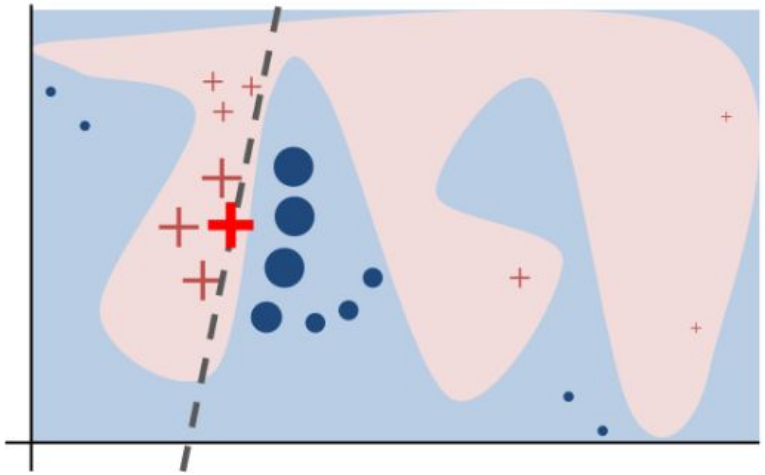
Tangent Line Approximation



Ribeiro et al. (2017)

- black box classification model ***f***: pink and blue areas
- instance being explained: bold red cross
- instances sampled locally and weighted by their proximity: red crosses, and blue circles
- locally faithful explanation ***g***: dashed line

Tangent Line Approximation



Ribeiro et al. (2017)

- black box classification model ***f***: pink and blue areas
- instance being explained: bold red cross
- instances sampled locally and weighted by their proximity: red crosses, and blue circles
- locally faithful explanation ***g***: dashed line

explanation “must correspond to how the model behaves in the vicinity of the instance being predicted”

Tangent Line Approximation or LIME in its original definition (Ribeiro, 2017)

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

black box

neighbourhood kernel around x

explanation model

fit *g* to *f* in a small neighbourhood around *x*

ensure *g* is simple

Problems with Tangent Line Approximations

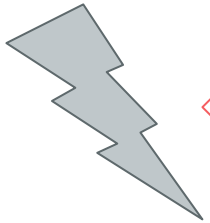
- If a linear fit is good enough, why not just have a locally linear black box? (Rudin, 2019)
- Consider the black box

$$f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$$

Problems with Tangent Line Approximations

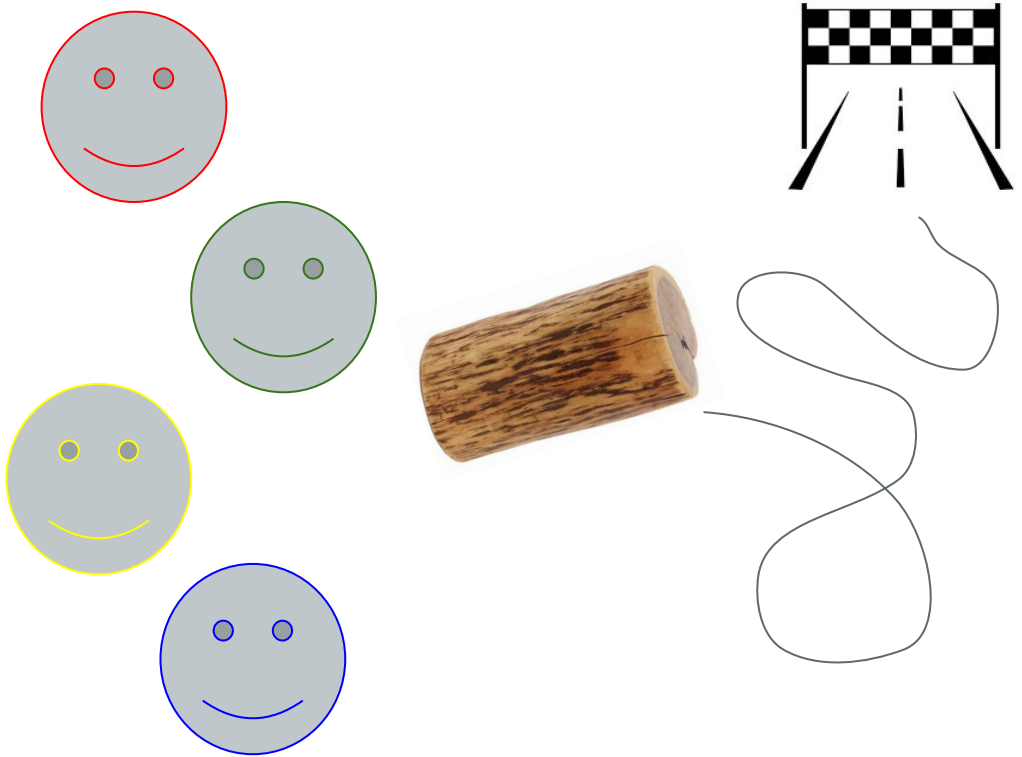
- If a linear fit is good enough, why not just have a locally linear black box? (Rudin, 2019)
- Consider the black box

$$f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$$



for x1=-0.001, the feature attribution of Feature-2 is **negative**

Shapley Values in Game Theory



You have

- four players playing a game
- a value function that summarises the value of the game, i.e. the distance the wood piece travelled

You want to know

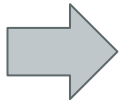
- what each player could contribute to the game

Shapley Values in Game Theory

Approach 1:

Assign an order to the players (i.e. Player 1 goes first, then Player 2, ...), and let each player carry the chunk of wood as far as they can, and attribute each player with the distance they carried the chunk of wood, i.e.

Attribution of Player 2 = Value Function after Player 2 played - Value Function after Player 1 played



Unfair since Player 2 might carry the chunk of wood in the steepest section of the path!

Shapley Values in Game Theory

Approach 2:

Sample an order of the players (i.e. Player 2 goes first, then Player 5, ...) from a **uniform distribution**, and let each player carry the chunk of wood as far as they can, and attribute each player with the **expected** distance they carried the chunk of wood where the expectation is taken over the order of players, i.e.

Attribution of Player 2 =

Expected Value Function after Player 2 played - Expected Value Function before Player 2 played

Shapley Values in XAI

- Players are now features → typically no ordering

$$\phi_v(j, x) = \mathop{\mathbb{E}}_{S}\left[v(S \cup j, x) - v(S, x)\right]$$

Shapley value of feature *j* at test instance *x*

Expectation over all possible feature subsets *S*

Distribution of *S* satisfies

$$P(\{S \mid |S| = k\}) = P(\{S \mid |S| = l\})$$

Choice of Value Function

Imagine you only have access to the features in subset *S*, how are you going to evaluate your black box *f*?

Just impute each missing feature with a sample from a reference distribution *r(X*|x)*

$$v(S, x) = \mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x)} [f(x_S, X^*_{\overline{S}})]$$

So far, *r* has always been chosen as any global distribution

Example

Consider a hiring algorithm

$$f\left(x_{strong},\ x_{male}\right) = x_{male}$$

with all features and the outcome being binary.

Now imagine we have

- P(male) = 0.5,
- P(strong) = 0.5,
- P(male|strong) = 0.8,
- P(male|not strong) = 0.2

We want to know why we would hire a **strong male x=(1, 1)**

Possible orderings are

- strong, male
- male, strong

Shapley value of Feature *strong* =

average{

v(x_strong, x_male) - v(x_male),

v(x_strong) - v(.)

}

How do we compute v(strong)?

conditional imputation

$$v(strong) = E_{x\_male|x\_strong=1}[f(x\_strong=1, x\_male)]$$

vs

$$v(strong) = E_{x\_male}[f(x\_strong=1, x\_male)]$$

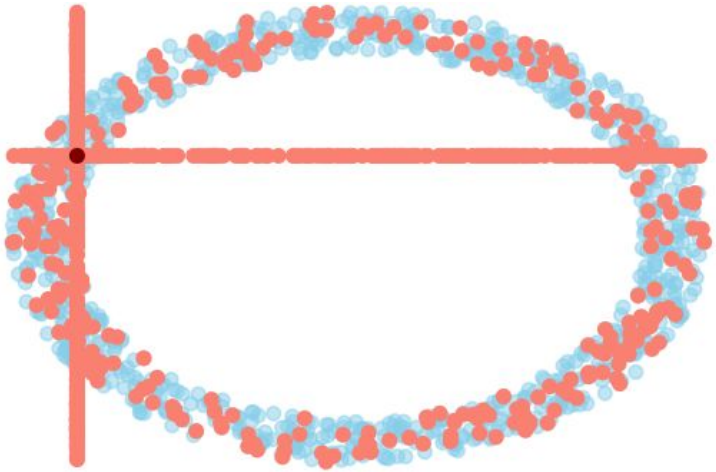
$$= E_x[f(x\_strong, x\_male)|do\ x\_strong=1]$$

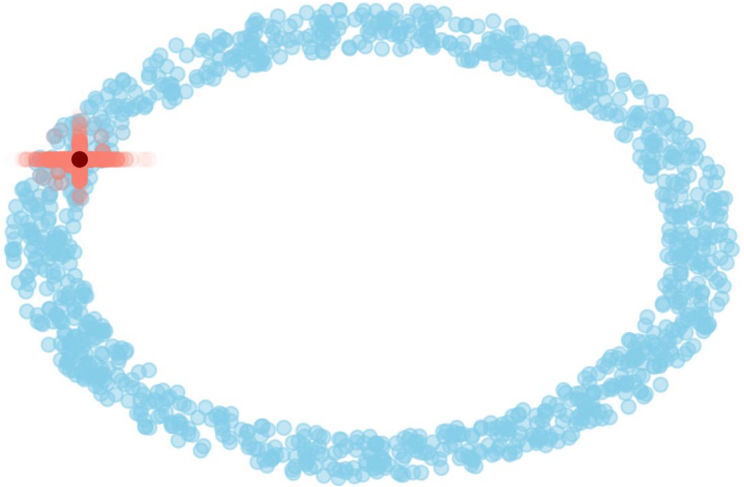
marginal imputation

On Locality of Local Explanation Models

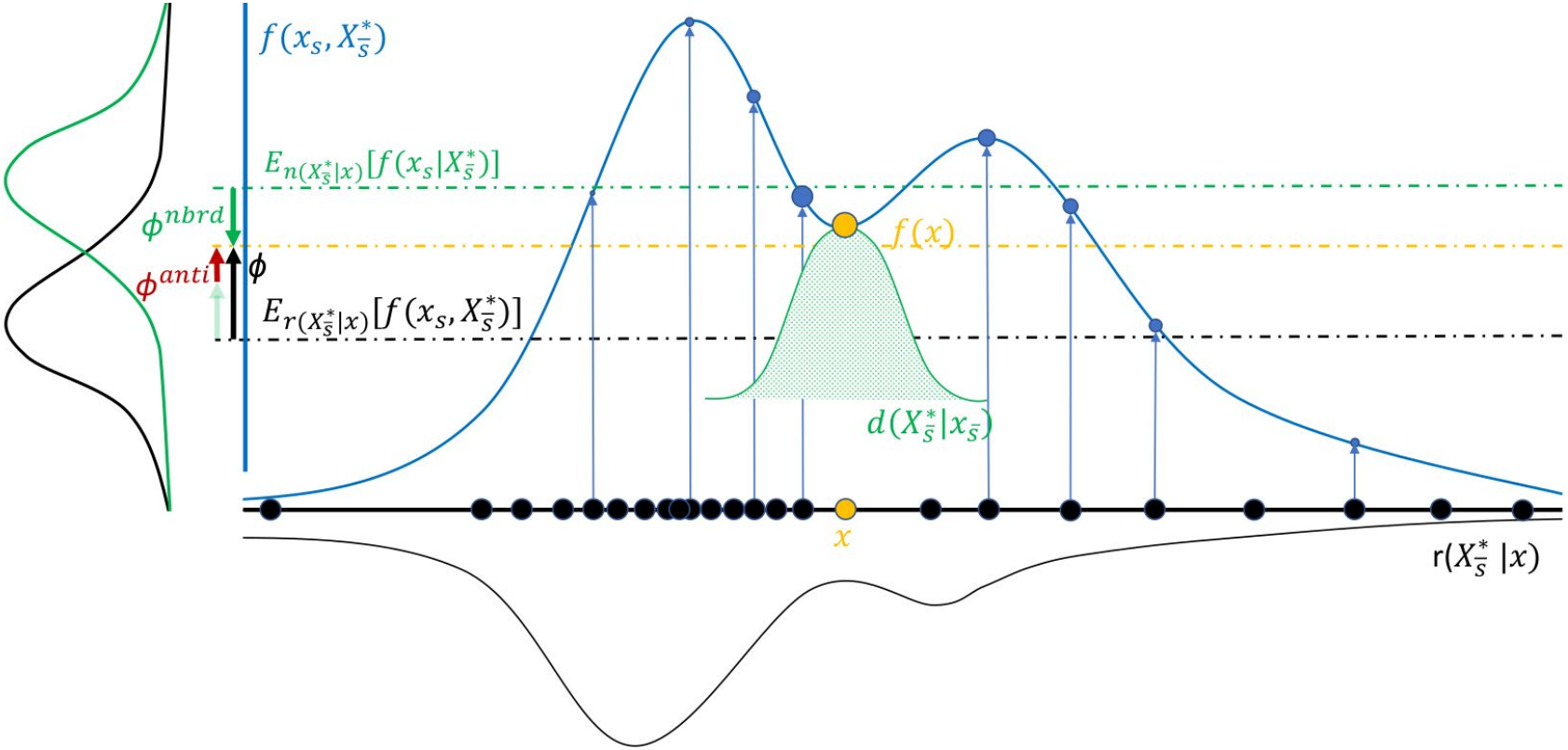
SG*, LTM* KDO, CCH

Neighbourhood SHAP evaluates on-manifold

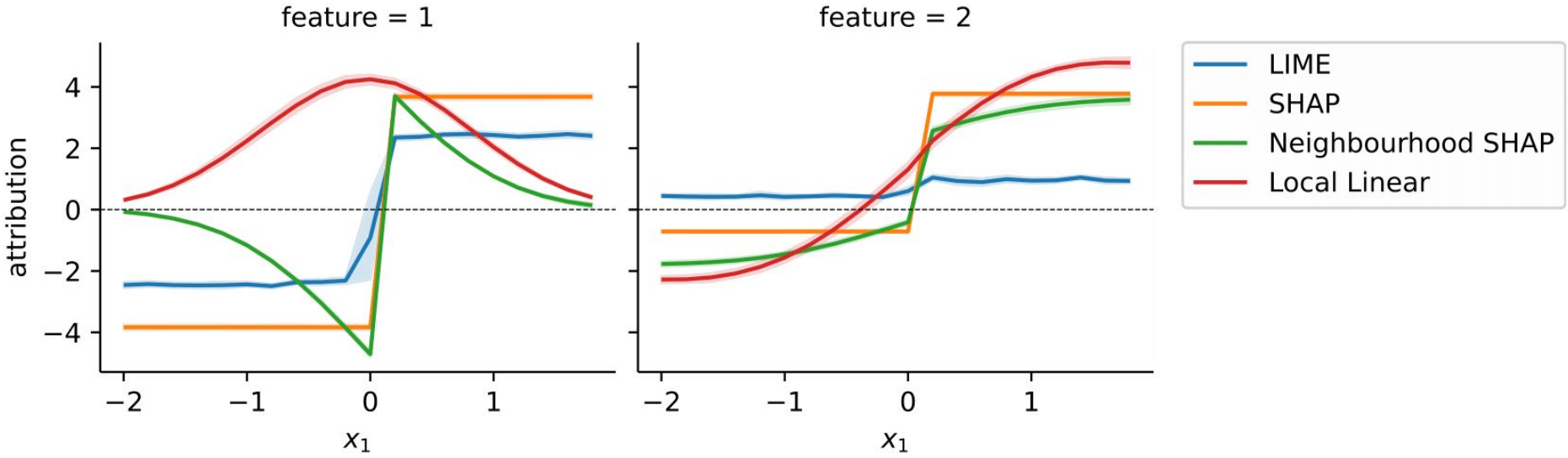




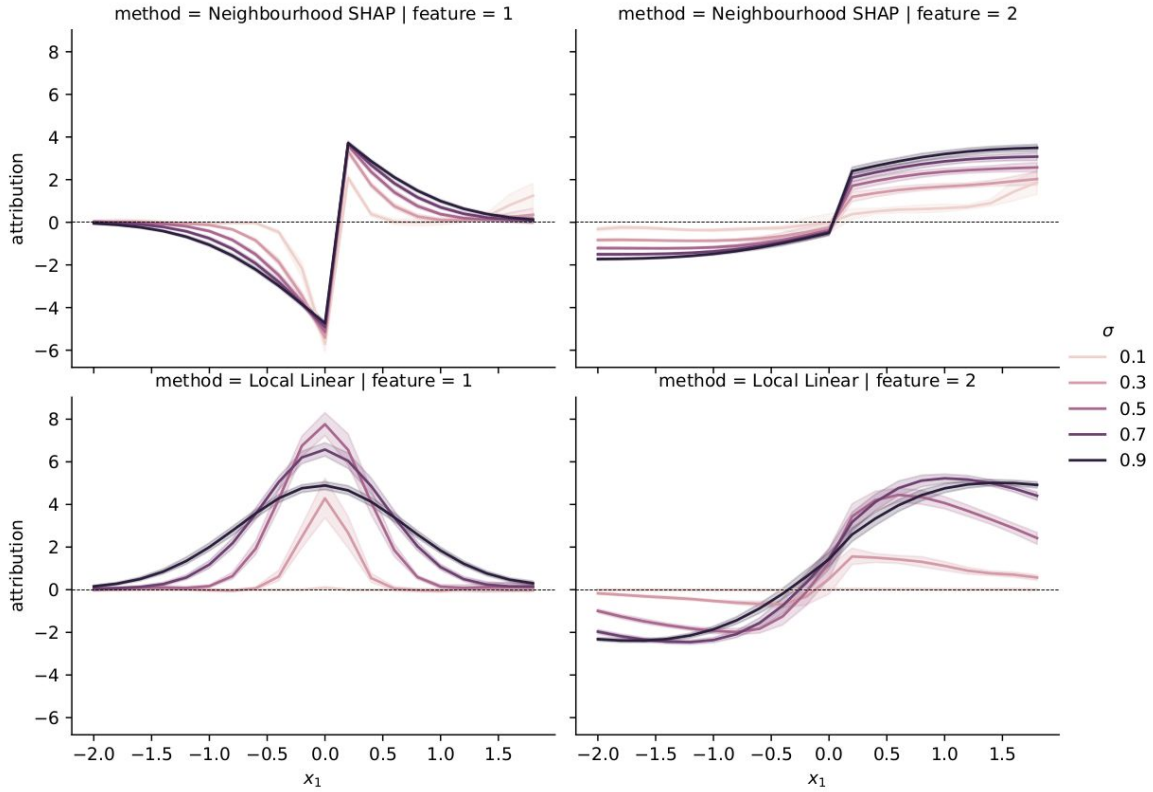
Neighbourhood SHAP samples locally



Neighbourhood SHAP just looks great (Part I)



Neighbourhood SHAP just looks great (Part II)



References

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1.5 (2019): 206-215.

Tanner, Gilbert. “Introduction to Machine Learning Model Interpretation”, Blog post at https://gilberttanner.com/blog/introduction-to-machine-learning-model-interpretation. 2019