Feature Instability Search for Multi-Way Explainability

Sean Kelly (skelly26@alumni.nd.edu) and Meng Jiang

University Of Notre Dame

R²HCAI: The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI

Motivation

Current explainability frameworks (LIME, SHAP):

1) Lack a reliable quantitative definition of explainability
2) Aren't evaluated on a true ground truth measure
3) Fail to account for multi-way feature interactions

Contributions

i. The concept of feature (in)stability as a measure of explainability of the output of a model

ii. A synthetic model with deterministic ground truth multi-way feature explainability to evaluate explainability frameworks

iii. An informed stability descent based search algorithm as an attempt to quantifying multi-way feature stability, or importance, for a given binary prediction

iv. A feature importance ranking evaluation loss function capable of comparing one-way feature explainability frameworks to more expressive frameworks (DFEST)

Feature Instability

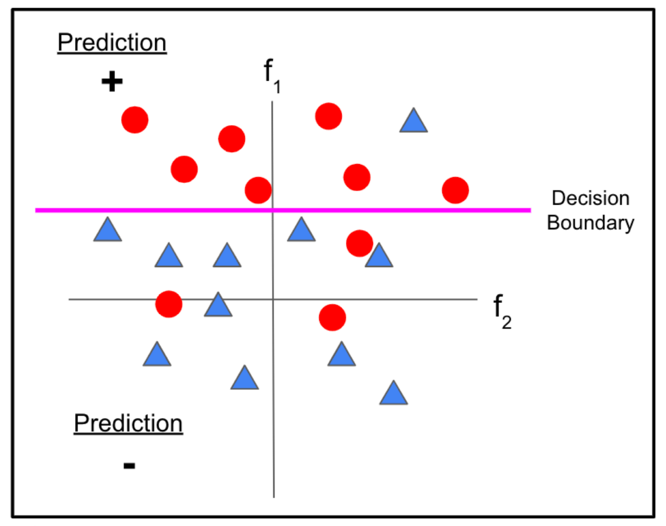


Figure 1a

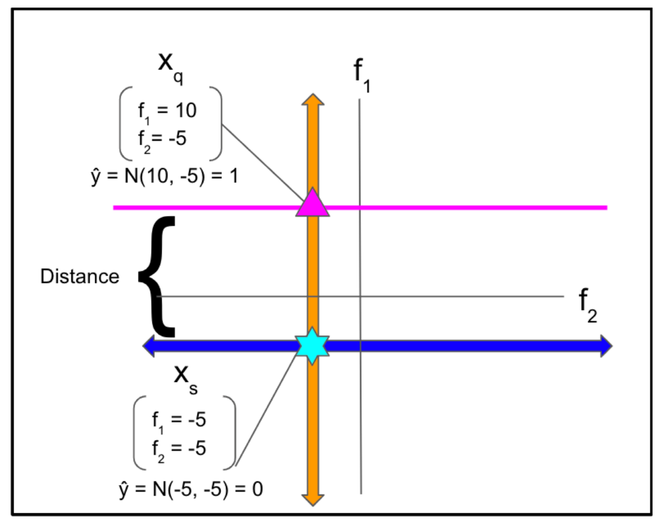


Figure 1b

$$\text{k-way Feature Instability} = \frac{1}{\text{Distance}} = \frac{1}{\triangle'(x_s, x_q)}$$

- Feature instability is a measure of post-hoc explainability that explains the feature interactions responsible for a given input source's prediction
- f_1 is unstable and f_2 is stable w.r.t. the model and any given input (x_q) to the model, representing a 1-way feature interaction, as only 1 feature is relevant

Synthetic Ground Truth Model

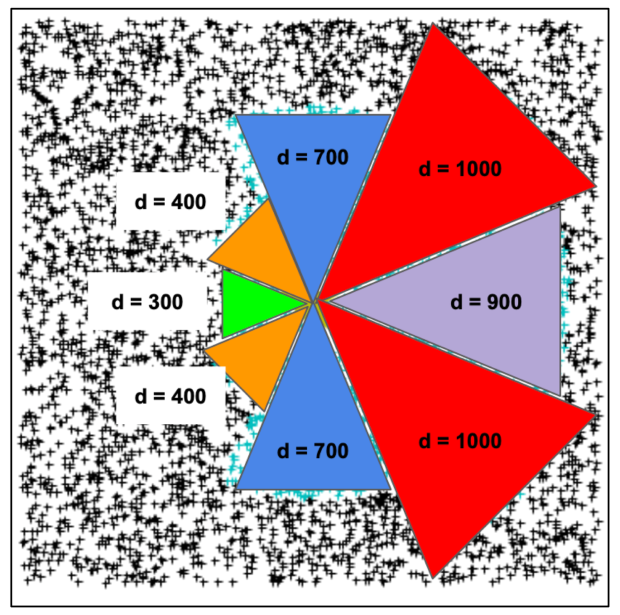


Figure 2b

Feature-Interaction Cluster Determination

Input $x_q = (f_1, \ldots, f_n)_q$
$cluster = [0_0, \ldots, 0_d]$
if $f_{max} > 0$ **then**
  $cluster[f_{max}] = 1$
else
  $cluster[f_{max}] = -1$
end if
for $f_i \in f$ **do**
  if $|\frac{f_{max}}{f_i}| \leq 2$ **then**
    if $f_i > 0$ **then**
      $cluster[f_i] = 1$
    else
      $cluster[f_i] = -1$
    end if
  end if
end for
return $cluster$ as c_q =0

Algorithm 3

- An inherently explainable n-dimensional decision space is generated, such that the distance from the origin (0,0) to a unique feature cluster i.e. (1,0) or (1,1) is predetermined, with distances to the decision boundary increasing as the block distance from the minimum cluster increases
- Provides a ranked list of the most unstable features, with quantitative instability values derived from the distance, queryable by XAI methods

DFEST: Feature Stability Descent and Tensor Search

DFEST is a post-hoc XAI method to identify the k-most unstable feature-interaction clusters through:
(1) Uniform Distribution Search, followed by
(2) Informed Cluster Search

DFEST Evaluation With Synthetic Ground Truth Model

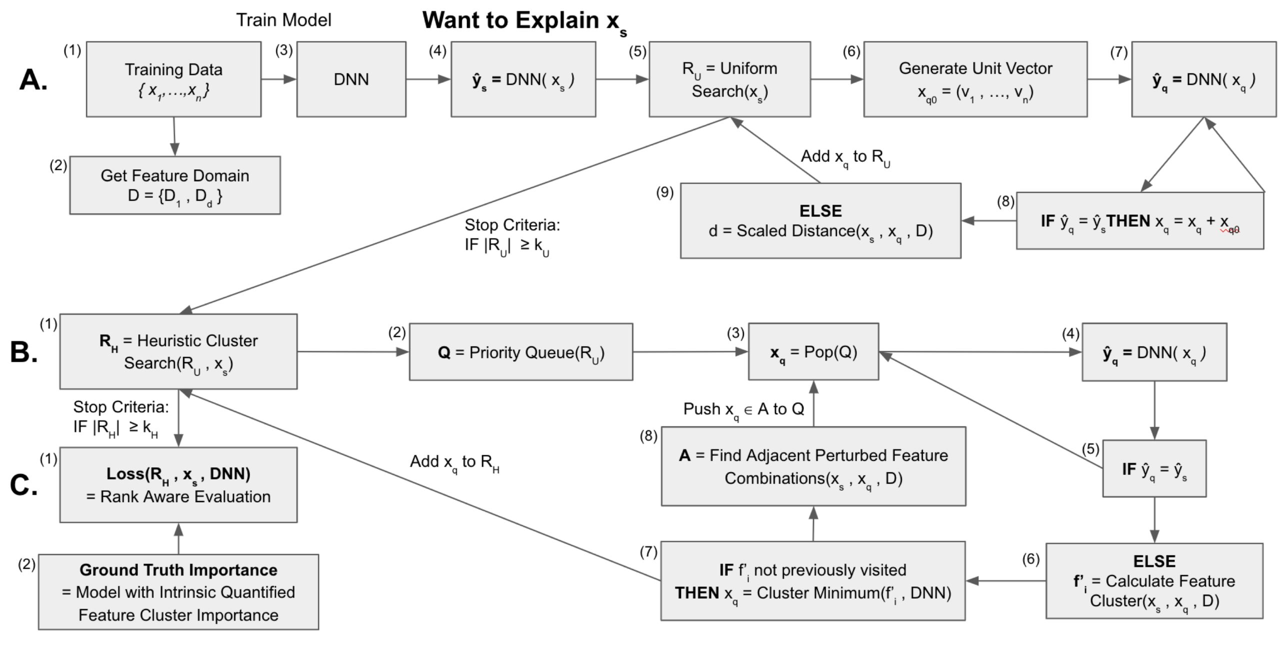


Figure A.1

Uniform Distribution Search

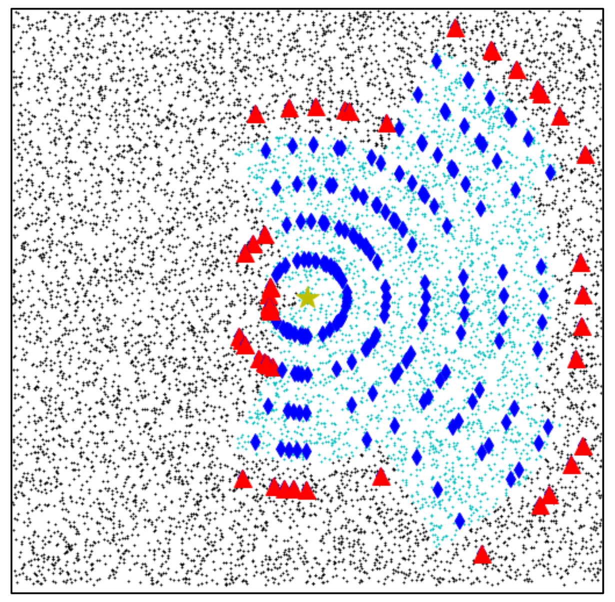


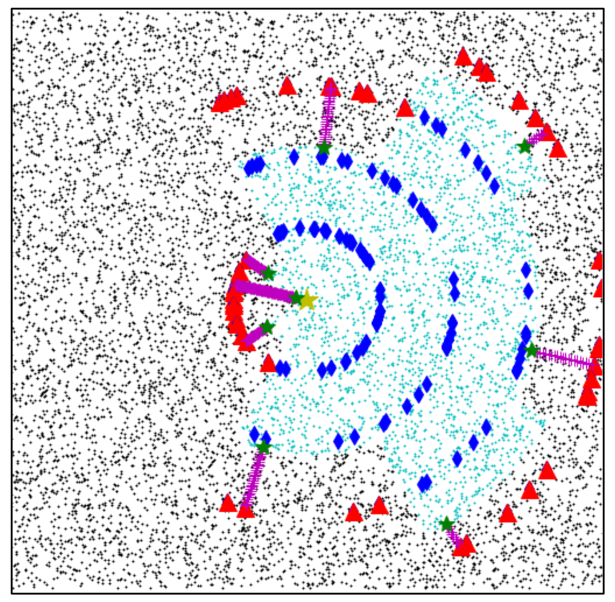
Figure 3a

D = $\sigma_3(f_i)$ - $\sigma_{-3}(f_i)$ $\forall i \in d$ and $\forall f_i \in$ training set
$\mu = \frac{d}{nSteps}$ $\forall d \in$ D
$F = \frac{1}{d}$ $\forall d \in$ D
featSteps = $(\mu_{f_0}, \ldots, \mu_{f_d})$
$\hat{y}$ = BBF(x_s)
for $0 \rightarrow k$ **do**
  x_u = Normalize(Rand(0,1) $\forall i \in n$)
  for $step \in nSteps$ **do**
    $x_q = (x_u \times featSteps[step]) + x_s$
    $\hat{y'}$ = BBF(x_q)
    if $\hat{y} \neq \hat{y'}$ **then**
      x_q.distance = $\triangle'(x_s, x_q)$
      R_U.insert(x_q)
    end if
  end for
end for
return R_U =0

Algorithm 1

- Search over the decision space has a time complexity of O(n^k)
- To gain heuristics for an informed search method, sparse solutions of clusters giving an opposite model output are identified surrounding the model output to be explained in decision space
- Heuristics are identified via generation of an even distribution of points on the surface of an n-sphere

Informed Cluster Search



- Informed cluster search is implemented as A* search over the priority queue of solutions discovered above
- Uniformly distributed heuristics enable random restarts, as adjacent cluster feature stability (gradient) descent toward the inner decision boundary is performed

Results: Top Feature Interaction Clusters

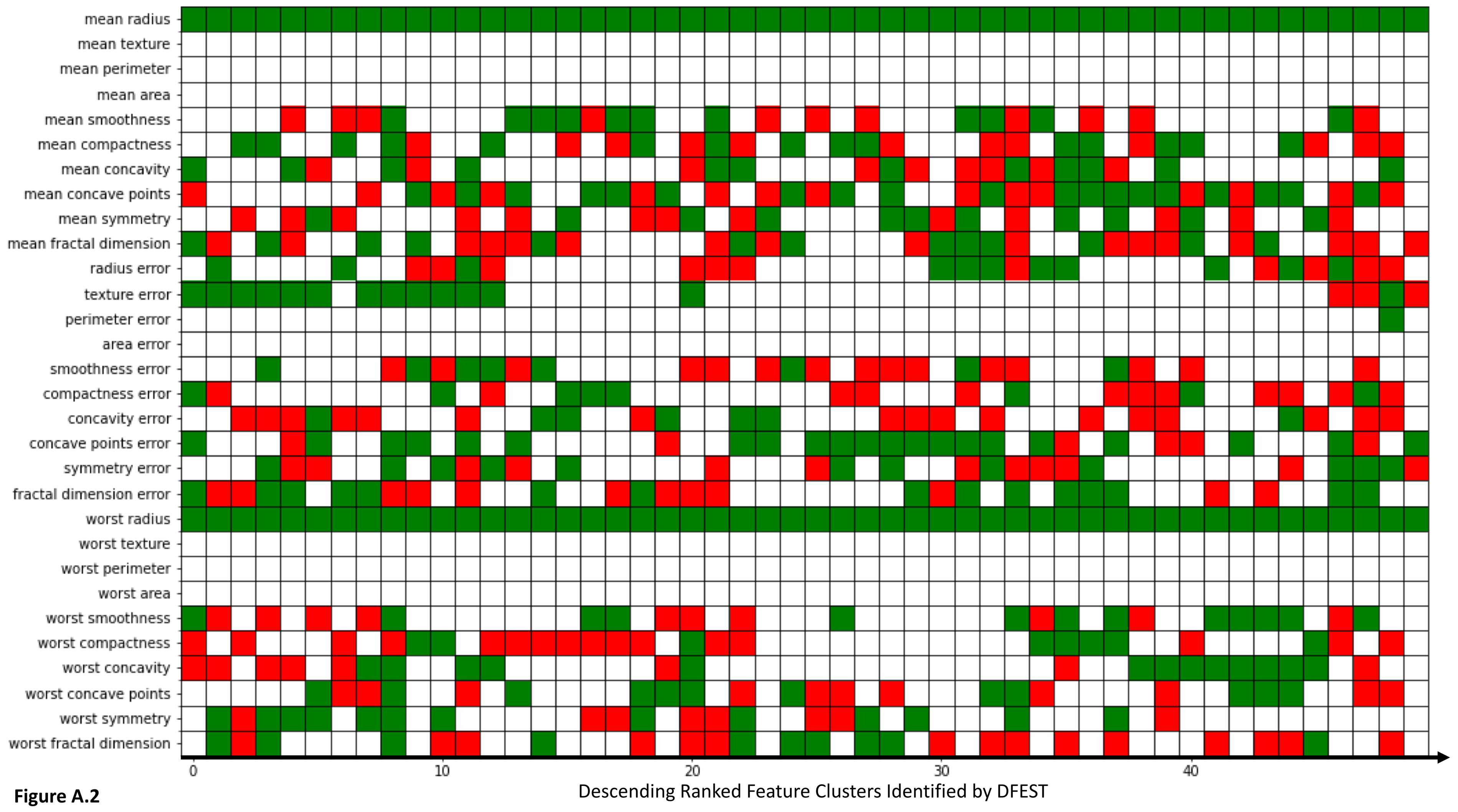


Figure A.2

The 50 most unstable feature-interaction clusters demonstrates that the clusters with the highest instability tend to have the same core unstable features, i.e. the MOST unstable features in each cluster

DFEST Feature Importance (Aggregation)

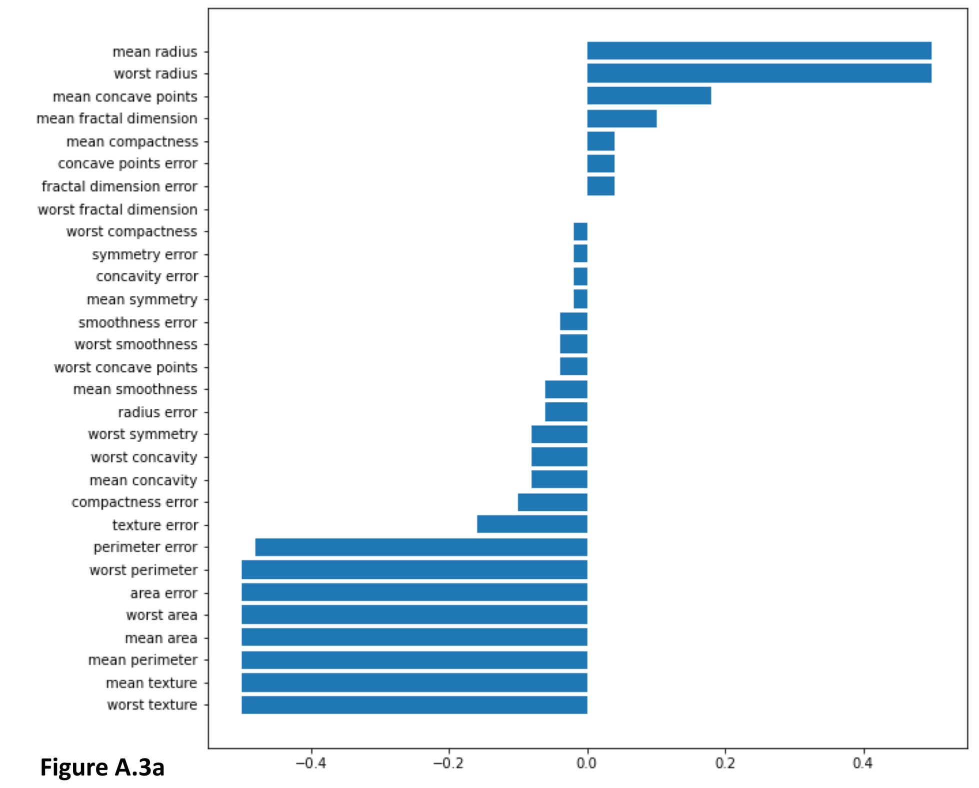


Figure A.3a

LIME Feature Importance

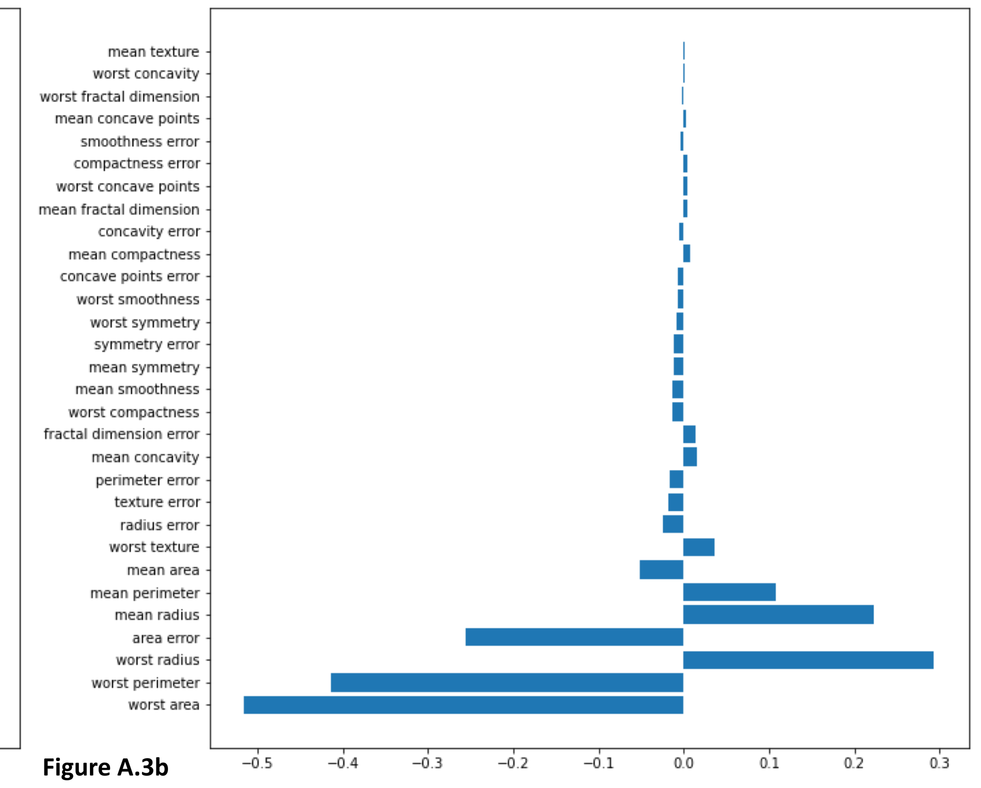


Figure A.3b

Comparison of DFEST & LIME Loss

	d Dimensions	Ranking Loss	k Precursor Solutions	Time Precursor Solutions (s)	k A* Solutions	Time A* Solutions (s)
DFEST		0.167	1,000	0.14	1,000	0.06
LIME	2	3.3125	5,000	1.08		
Random		0.9375				
DFEST		0.0625	1,000	0.31	100	0.15
LIME	4	1.594	5,000	1.7		
Random		0.75				
DFEST		0.141	10,000	1.35	1,000	3.94
LIME	8	0.718	10,000	3.59		
Random		0.875				
DFEST		0.0234	100	0.054	1,000	4.0449
LIME	16	0.7031	10,000	7.825		
Random		0.90625				
DFEST		0.4065	100,000	48.68	1,000	11.3
LIME	32	0.6718	10,000	14.35		
Random		0.855				
DFEST		0.5351	100,000	33.221	5,000	156.87
LIME	64	0.648	1,000,000	28.23		
Random		0.867				
DFEST		0.634	300,000	412.34	5,000	1138.85
LIME	128	0.675	1,000,000	55.21		
Random		0.875				

Figure A.2

$$loss = \sum_{i=1}^{k} \left[min \sum_{j=1}^{k} \left[(|c_{i_{index}} - g_{j_{index}}|+1) * (\sum_{x}^{g_i} |c_i - g_j|+1) \right] - 1 \right] * \frac{1}{k}$$

Equation 3

Conclusion

- The loss function accounts for both **cluster variability** and **feature instability magnitude,** given continuous input features
- The synthetic ground truth model with deterministic feature instability offers a trustworthy benchmark for the development and evaluation of future XAI methods
- DFEST demonstrates a method to quantify the impact of multi-way feature interactions on a model's output, which is inherently out of scope for current feature attribution methods which perform local linear function approximation