Supplementary material
BOSSE: BOosting SphereS Explanations

Anonymous

Anonymous Institution

1 Appendix

1.1 Algorithms

$$N_r = 100nr\pi^{n/4}, \tag{1}$$

where n is the dimensionality and r the radius.

The **sphereSearch** function is a method based on the growing spheres algorithm by Laugel et al. [2] and is illustrated in Alg. 1. This function increases r_{small} and decreases r_{big} recursively. The process is repeated until both radii converge and the absolute difference between them is lower than or equal to ϵ. Lines 1 and 2 randomly sample N (following Eq. 1) examples on the surfaces of the spheres of radius r_{small} and r_{big}. Line 3 verifies whether $sphere_{r_{small}}$ contains a CF using a function called **containsPlausibleCF**. This function determines whether the sphere surface points generated are close to the plausible feature values. For example, in a 3-dimensional feature space, where the first two features are binary and the third is continuous, a possible generated point could be [0.8, 0.2, 0.4], while another could be [0.96, 0.04, 0.4]. In this case, the former point is further from being plausible, because 0.8 is further from 1 then 0.96, and 0.2 is further from 0 than 0.04. We use 0.05 as the threshold for binary and ordinal features to indicate whether a given surface point is plausible or not. If the small sphere contains a plausible CF, then the closest CF must be there (because r_{small} started as close as possible to the IOI p), and line 4 returns the small sphere surface examples. Line 5 constructs a new sphere with radius $(r_{small} + r_{big})/2$. If the mean-radius sphere contains a plausible CF, then $r_{big} = (r_{small} + r_{big})/2$ (line 7). If the mean-radius sphere does not contain a plausible CF, then $r_{small} = r_{mean}$ (line 9). Line 10 makes a recursive call of the **sphereSearch** algorithm, until $|r_{small} - r_{big}| < \epsilon$ or the sphere of r_{small} contains a plausible CF.

1.2 Datasets

1. **German**: UCI ML Repository dataset available at the website[1] for credit risk prediction. The dataset initially contains 20 features, which are preprocessed to obtain a final set of 4 features, 1 binary, namely Sex ∈ {Male, Female} and 3 continuous, namely Age, Credit and LoanDuration.

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Algorithm 1: sphereSearch Pseudoalgorithm

input : p: IOI, r_{big}: big sphere radius, $r_{pre-big}$: previous big sphere radius, r_{small}: small sphere radius, $f(p)$, ϵ
output: $sphere_{instances}$: instances on the smallest sphere containing a CF

$sphere_{r_{small}} \leftarrow$ `sphereConstruction`$(p, r_{small}, f(p))$
$sphere_{r_{big}} \leftarrow$ `sphereConstruction`$(p, r_{big}, f(p))$
if `containsPlausibleCF`$(sphere_{r_{small}})$ *or* $|r_{big} - r_{small}| < \epsilon$ **then**
  return $sphere_{r_{small}}$
$sphere_{r_{mean}} \leftarrow$ `sphereConstruction`$(p, (r_{big} + r_{small})/2, f(p))$
if `containsPlausibleCF`$(sphere_{r_{mean}})$ **then**
  $r_{big} \leftarrow (r_{big} + r_{small})/2$
else
  $r_{small} \leftarrow (r_{big} + r_{small})/2$
return `sphereSearch`$(p, r_{big}, r_{pre-big}, r_{small}, f(p))$

2. **Compass**: Propublica dataset for recidivism prediction, available at the Propublica website[2]. The dataset used is the compass-scores-two-years.csv. The dataset is processed to contain only 5 features, 3 binary, namely Sex $\in$ {Male, Female}, Race $\in$ {African-American, Caucasian}, ChargeDegree $\in$ {Misdemeanor, Felony}, and 2 continuous, namely PriorsCount and Age. The target variable is a new criminal sentence in the next two years.
3. **Ionosphere**: UCI ML Repository dataset available at the website[3] for the prediction of ionospheric condition prediction. A RF model is implemented to obtain 5 features out of the 34 available continuous features according to Mean Decrease in Impurity (MDI) measure. Fig. 1 shows the MDI for all 34 features. The features corresponding to the highest MDI are the most important. In this case, features 2, 4, 5, 6 and 26 are selected.

The preprocessing of the **Compass** and **German** datasets is carried out according to the pipeline presented by Karimi et al. [1].

[2] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[3] https://archive.ics.uci.edu/ml/datasets/ionosphere

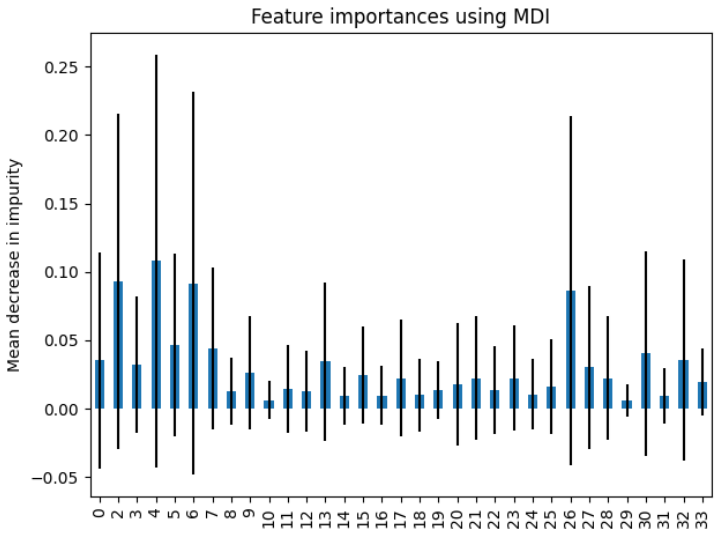

Fig. 1: MDI for all Ionosphere dataset features.

References

1. Karimi, A.H., Barthe, G., Balle, B., Valera, I.: Model-Agnostic Counterfactual Explanations for Consequential Decisions
2. Laugel, T., Renard, X., Lesot, M.J., Marsala, C., Detyniecki, M.: Defining Locality for Surrogates in Post-hoc Interpretablity. arXiv:1806.07498 [cs, stat] (Jun 2018), http://arxiv.org/abs/1806.07498, arXiv: 1806.07498