Duke UNIVERSITY

Yale SCHOOL OF MANAGEMENT

NEURAL INFORMATION PROCESSING SYSTEMS

Improving Decision Sparsity

Yiyang Sun, Tong Wang, and Cynthia Rudin

Sparsity is important for interpretability

WIKIPEDIA
The Free Encyclopedia

Create account Log in

The Magical Number Seven, Plus or Minus Two

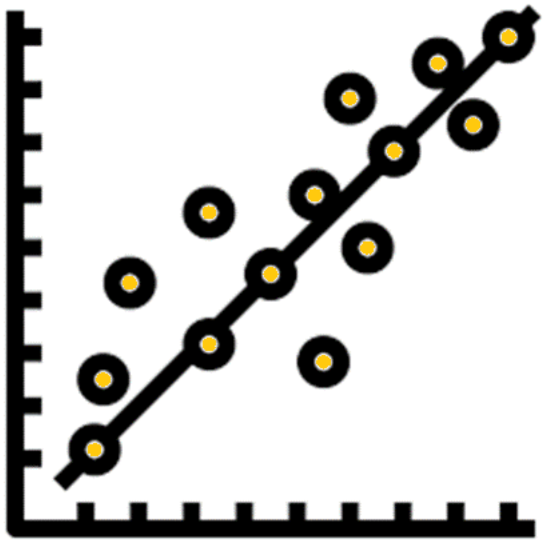
18 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

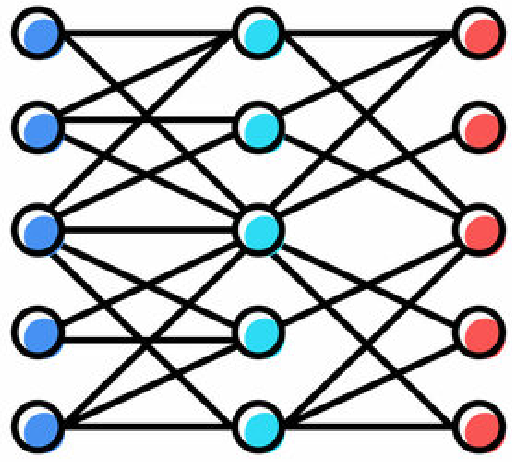
"**The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information**"[1] is one of the most highly cited papers in psychology.[2][3][4] It was written by the cognitive psychologist George A. Miller of Harvard University's Department of Psychology and published in 1956 in *Psychological Review*. It is often interpreted to argue that the number of objects an average human can hold in short-term memory is 7 ± 2. This has occasionally been referred to as *Miller's law*.[5][6][7]

Common metrics for evaluating sparsity of the model



# of terms
in linear regressions

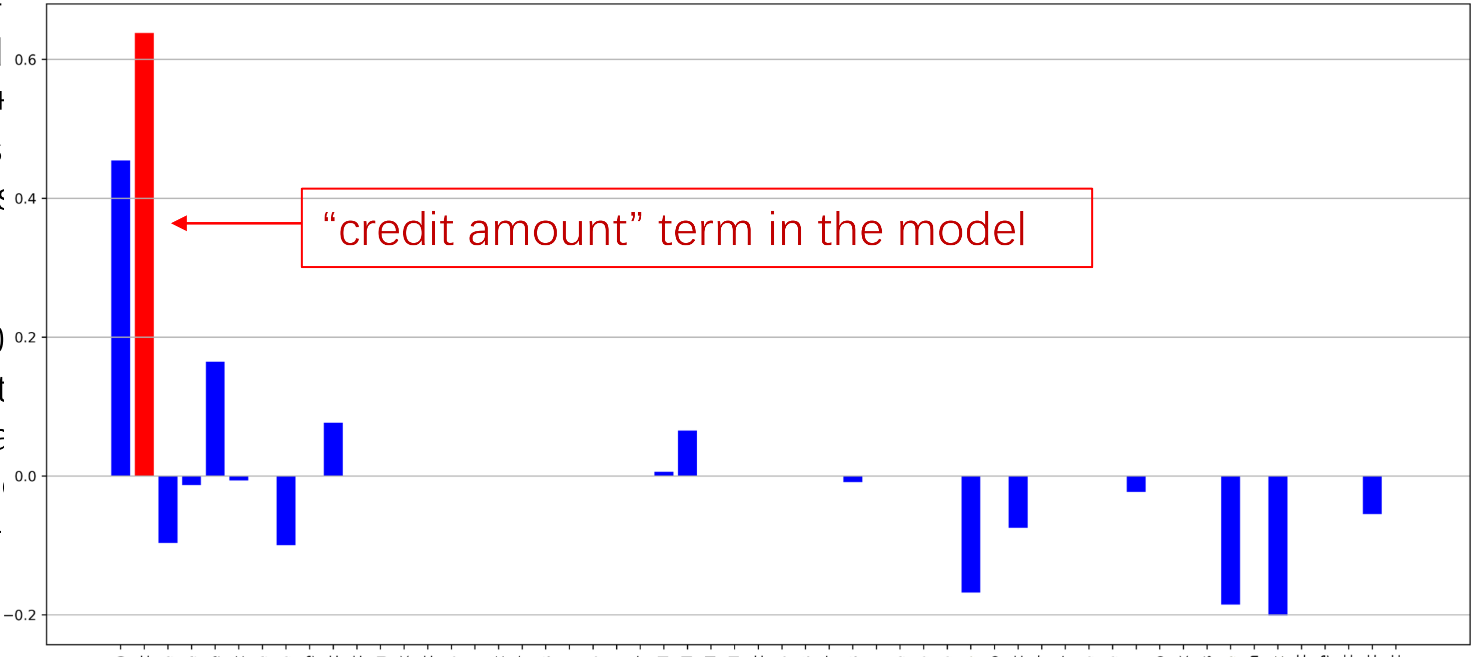
# of leaf nodes
in decision trees



# of parameters
in neural networks

This is not a sparse model.

(duration(month))*0.21+(credit_amount)*0.33+(installment_rate)*0.13+(present_residence_since)*-0.01+(age)*-0.11 +(number_of_credit)*0.03 +(number_of_people_being_liable)*-0.01 +(provide_maintenance)*-0.09+(telephone)*-0.07 +(check_acocunt_existence=0-200DM)*0.06 +(checking account = none)*0.35 +(credit_history=all_credit)*0.08 +(credit_history=critical_account)*-0.19
+(credit_history=delay_in_paying)*0.01 +
+(purpose=business)*0.02+(purpose=ed ... ɔose=others)*-
0.01 +(purpose=radio/television)*-0.12 +
+(saving_account=100-500DM)*0.02 +(s
+(saving_account=>=1000DM)*-0.04 +(s
+(present_employment=4-7years)*-0.11
+(present_employment=unemployed)*0. ... ed_male)*0.03
+(personal_status_sex=married_male)*-0 ... ɔtors=co-
applicant)*0.02 +(other_debtors=guarant ... te)*-0.13
+(property=unknown)*0.04 +(other_insta ... stores)*0.03
+(housing=free)*0.0 +(housing=doesn't ...).06
+(job=unemployed_non-resident)*-0.0+



Why was Arun's loan application denied?

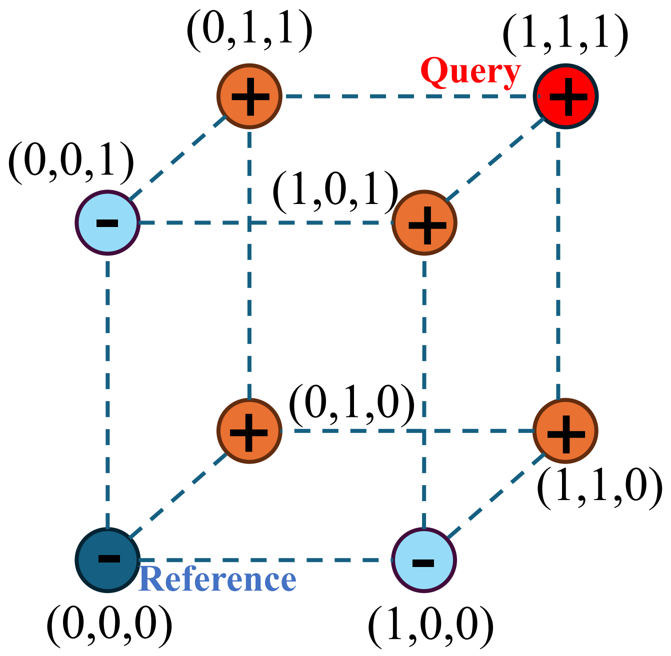
The prediction is determined by *one feature*!

He asked for over $10K!

Recap of Decision Sparsity: SEV$^-$

- Step 1: Define reference (usually either **0** or average of negative class)

- Step 2: Define Boolean hypercube for query ***x***:
  - coordinate$_j$ is 1 if feature j is x_j
  - coordinate$_j$ is 0 if feature j is at reference value r_j

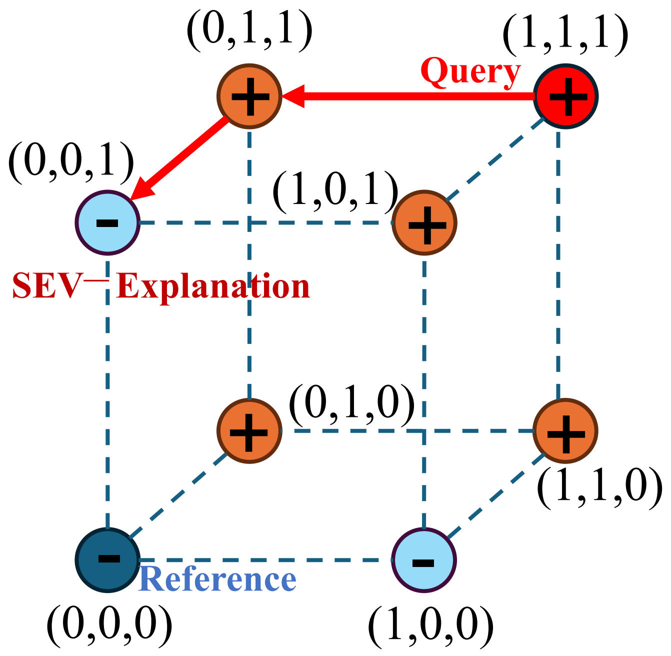


Recap of Explanation Sparsity: SEV^{-}

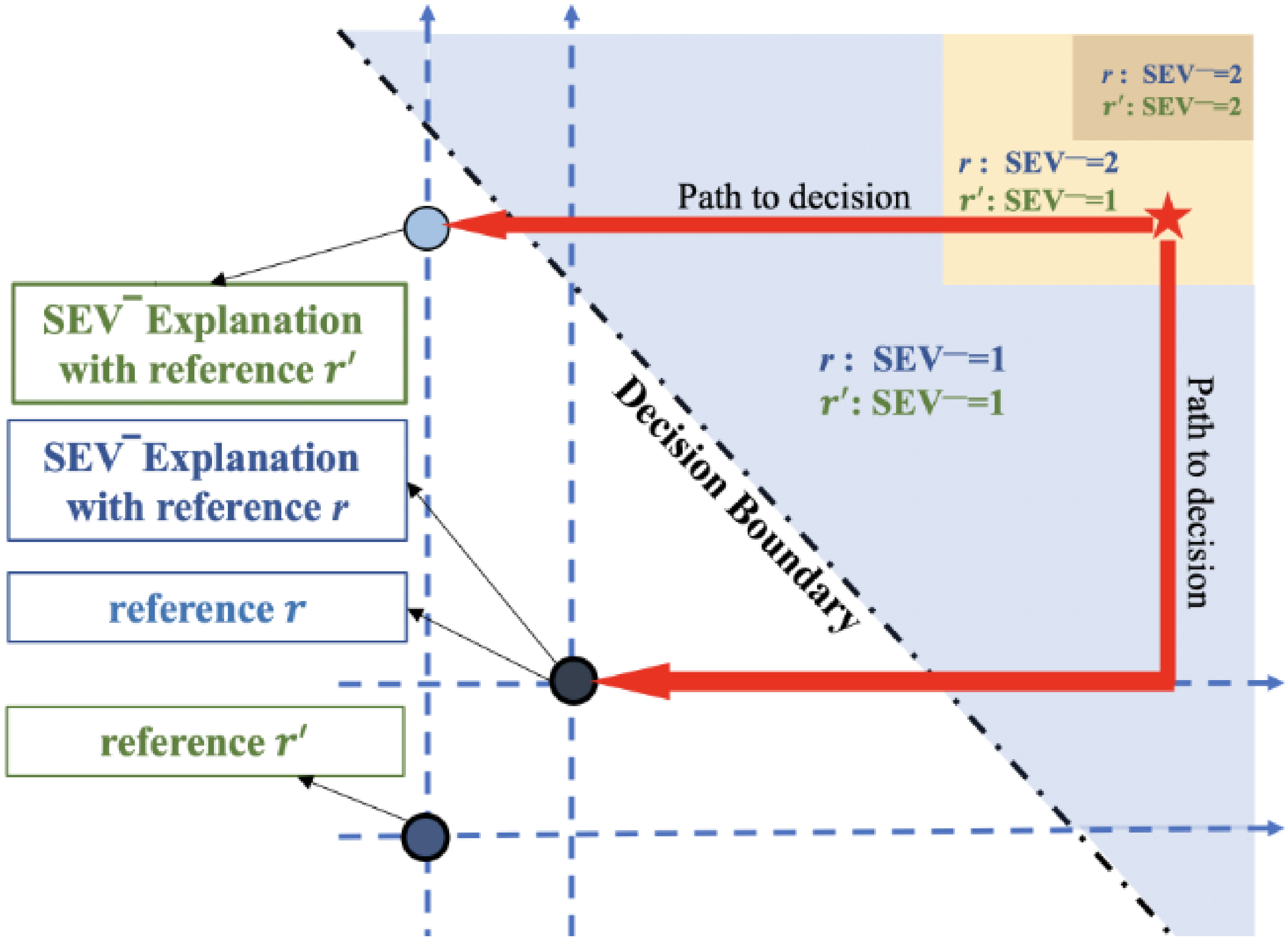
- Step 1: Define reference (usually either **0** or average of negative class)

- Step 2: Define Boolean hypercube for query $\boldsymbol{x}$:
  - coordinate$_j$ is 1 if feature j is x_j
  - coordinate$_j$ is 0 if feature j is at reference value r_j

- Step 3: Define SEV^{-} for $\boldsymbol{x}$. Moving from $\boldsymbol{x}$ towards reference, SEV^{-} is the minimum l_0 distance to a negative label.



SEV⁻ is sensitive to the reference selection



Selection criteria for references

Reference and Query are under the same subpopulations

Minimum number of features used for explaining decisions

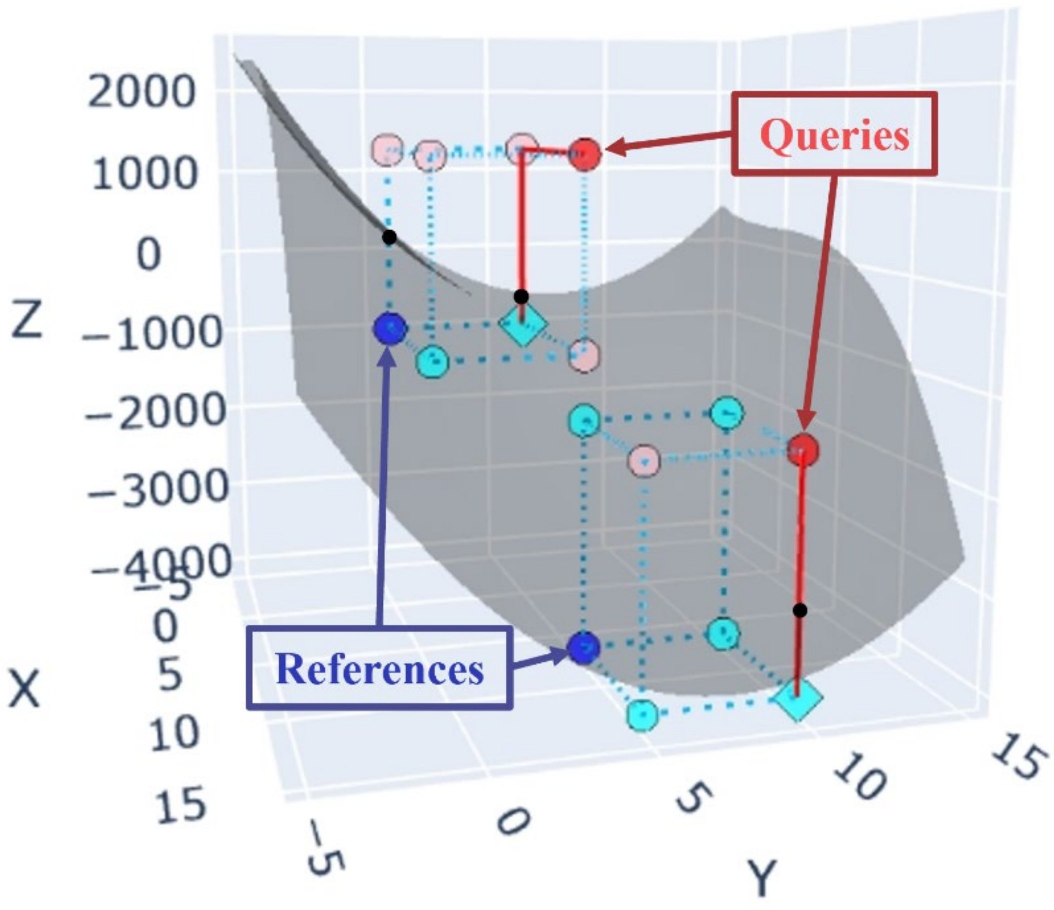
$$\|\boldsymbol{x}_i - \boldsymbol{r}_i\|, \boldsymbol{r}_i \in \mathcal{R} \quad \text{(Closeness)}$$

$$\text{SEV}^-(f, \boldsymbol{x}_i, \boldsymbol{r}_i), \boldsymbol{r}_i \in \mathcal{R} \quad \text{(Sparsity)}$$

$$-P(\boldsymbol{x}_i^{\text{expl},\boldsymbol{r}_i} | X^-) \quad \text{(Negated Credibility)}$$

Sparse explanations are under the original data distribution

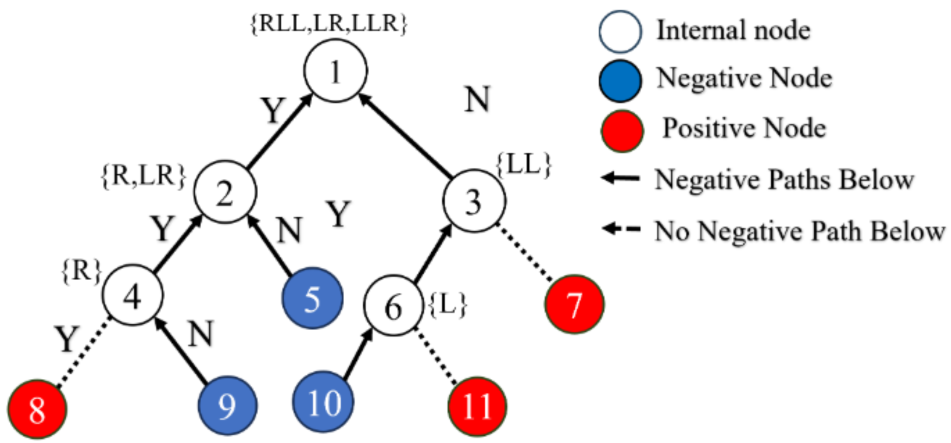
Improving Closeness: Cluster-based SEV (SEV-C)



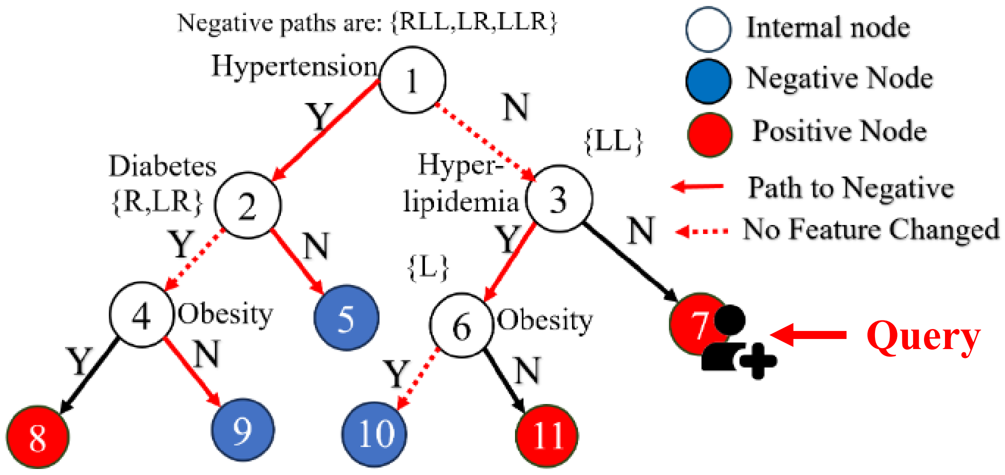
The process for calculating the SEV-C:

- **Step 1:** Selects the references by clustering the negative populations, and regarding the cluster centroid points as the references.
- **Step 2:** Assign each query their closest cluster centroid as their reference points.
- **Step 3:** Go over the original SEV Calculation

Special case for SEV-C: Tree-based SEV



Step 1: SEV-T Preprocessing: Collect negative leaf node information for each internal nodes



Step 2: Efficient SEV-T Calculation: Go over internal node of decision path, and check all negative paths

It have many useful properties and computational benefits!

Optimizing SEV Variants for Models

Gradient-based Optimization (AllOpt)

- Maximize the fraction of points with $SEV^{-}=1$

Search-based Optimization (TOpt)

- Find a model with the lowest SEV with in a set of classification models with the best performance

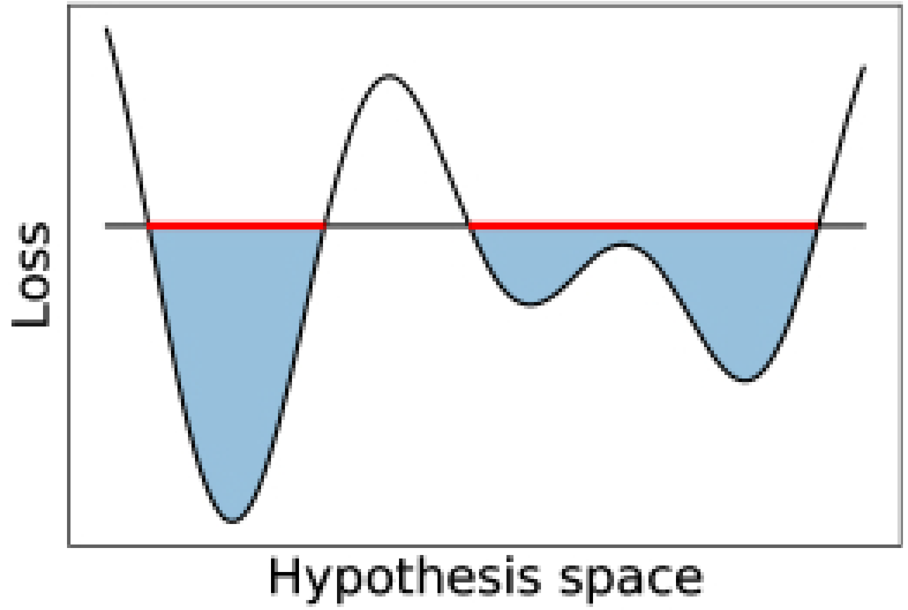
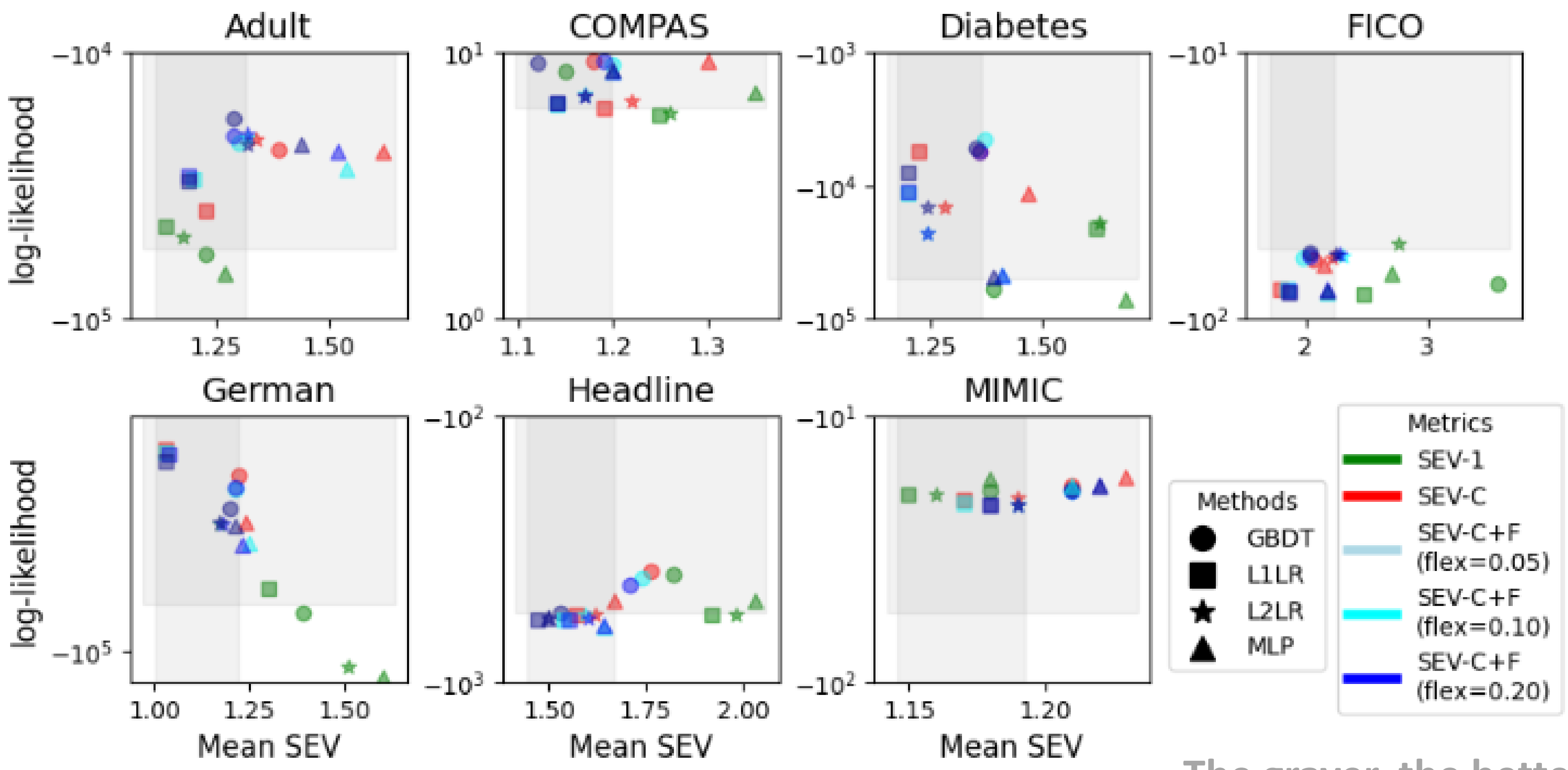


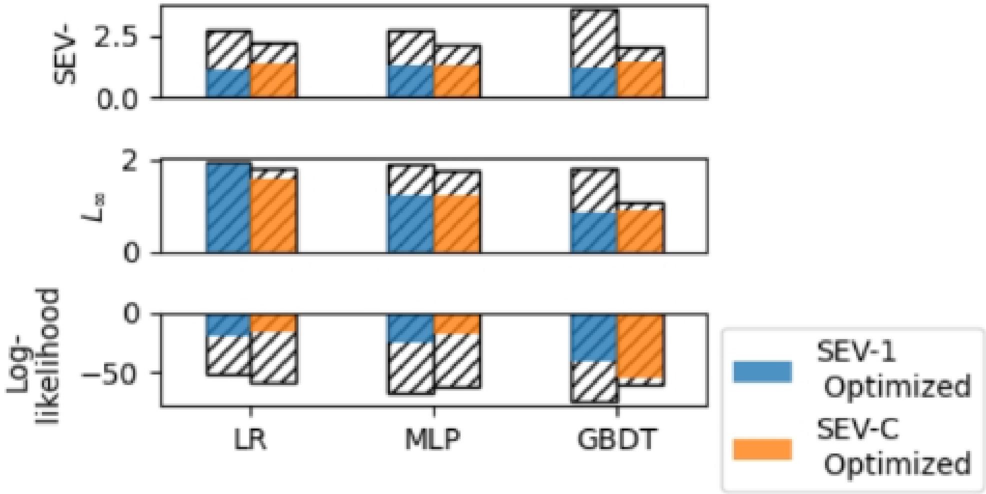
Figure 1: Rashomon Sets

SEV-C improves in credibility and closeness



The grayer, the better.

Optimization improves explanations while preserving model performance



(a) All-Opt^{-} Performance

	Train Acc	Test Acc	Mean SEV^{T}
CART	0.71 ± 0.01	0.71 ± 0.01	1.10 ± 0.03
C4.5	0.71 ± 0.01	0.71 ± 0.01	1.13 ± 0.05
GOSDT	0.70 ± 0.01	0.70 ± 0.01	1.08 ± 0.02
TOpt	0.70 ± 0.01	0.70 ± 0.01	$\boxed{1.00 \pm 0.00}$

(b) SEV^{T} performance on different tree-based models

The model performance are not changed before and after optimization

More in the paper

- SEV Variants for further improving the sparsity and credibility of the explanations.
- Sparsity and Credibility comparison with counterfactual explanation methods.
- Score-based soft K-Means for avoiding positive predicted references
- Detailed Algorithms for tree-based SEV
- Timing experiments
- …

SEV Paper NeurIPS Paper