Motivation and summary

- A counterfactual explanation seeks the minimal change to a given feature vector that will change a classifier’s decision in a prescribed way.

- Counterfactual explanation is important to interpret a black-box decision for a given instance.

- Mathematically, it has the same formulation as classifier inversion and adversarial examples: given a source instance \( \mathbf{x} \), target class \( y \) and a classifier \( F \), find the closest instance \( \mathbf{x} \) to \( \mathbf{\bar{x}} \) such that \( F(\mathbf{x}) = y \).

- Given an input instance \( \mathbf{\bar{x}} \in \mathbb{R}^D \), predictive function \( F \), and set of target predictions \( S \), the problem can be formulated as:

\[
\min_{\mathbf{x} \in \mathbb{R}^D} d(\mathbf{x}, \mathbf{\bar{x}}) \quad \text{s.t.} \quad F(\mathbf{x}) \in S.
\]

where \( d(\mathbf{x}, \mathbf{\bar{x}}) \) is the cost of changing features of \( \mathbf{\bar{x}} \).

- Here, we focus on decision forests (axis-aligned and oblique).

- With decision forests \( F \) is not differentiable, this makes problem nondifferentiable and non-convex, and gradient-based methods are not applicable.

- For decision trees, whether axis-aligned or oblique, the problem can be solved exactly and efficiently by finding an optimal CE within each leaf’s region and picking the closest one.

Counterfactual explanations in decision forest

- Decision forests define a piecewise constant function with an exponential number of regions in feature space, so searching for a counterfactual explanation exhaustively is impractical.

- LIRE (LIve REgion search) restricts the search to only those regions containing at least an actual data point, producing a very good approximate solution with a runtime suitable even for interactive use.

- LIRE also generates realistic counterfactuals because restricted regions act as a nonparametric density estimate.

Realistic solution

- Decision forests define a piecewise constant function with an exponential number of regions in feature space, so searching for a counterfactual explanation exhaustively is impractical.

- LIRE (LIve REgion search) restricts the search to only those regions containing at least an actual data point, producing a very good approximate solution with a runtime suitable even for interactive use.

- LIRE also generates realistic counterfactuals because restricted regions act as a nonparametric density estimate.