Integrated Gradients



Sensitivity

is the attribution sensitive to **relevant features**: inputs differ by a single feature

Axioms

Implementation Invariance

the attributions are always identical of **functionally equivalent** networks

outputs are identical for all inputs

Integrated Gradients



Sensitivity

is the attribution sensitive to **relevant features**: inputs differ by a single feature

Axioms

The attribution for LIME is zero for features with zero gradient at the input despite a non-zero gradient at the baseline.