

Integrated Gradients



Axiomatic Attribution for Deep Networks
(Jun 2017) by Sundararajan et al

Axioms

Sensitivity

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

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

Default Prediction Baseline

baseline vector
 x'

$$\begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix}$$

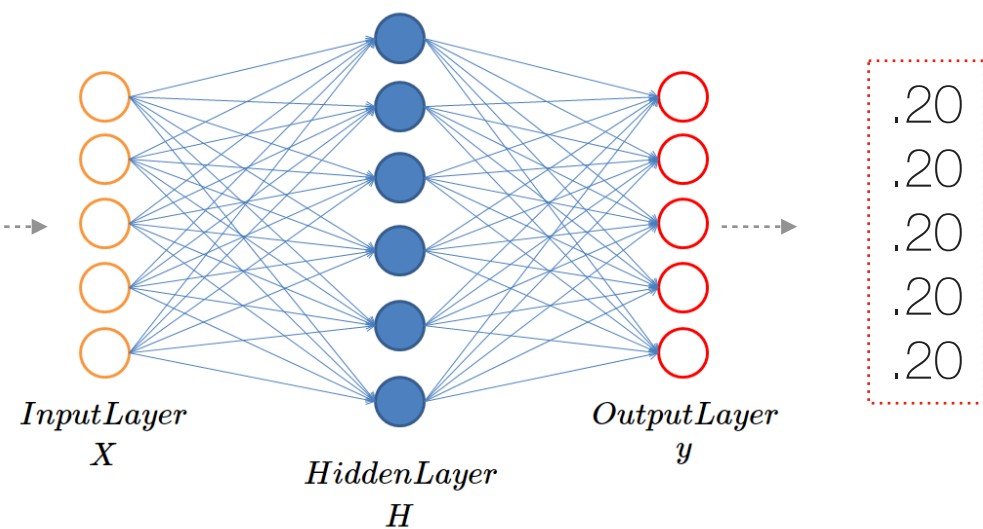


Image Classification Baseline

