Mathematical Properties of SHAP

1

Local Accuracy

$$f(x) = \varphi_0 + \Sigma\varphi_i(x)$$

Explanation matches model output exactly

2

Missingness

$$x_i = 0 \Rightarrow \varphi_i = 0$$

Absent features have zero attribution

3

Consistency

$$\Delta f_i \uparrow \Rightarrow \varphi_i \uparrow$$

Higher contribution = higher SHAP value

4

Efficiency

$$\Sigma\varphi_i = f(x) - E[f(x)]$$

Values sum to deviation from average

5

Symmetry

$$i \equiv j \Rightarrow \varphi_i = \varphi_j$$

Equivalent features get equal credit

6

Linearity

$$\varphi(f+g) = \varphi(f) + \varphi(g)$$

Additive for model ensembles

Visual Proof: Local Accuracy Property

Base Value $\varphi_0 = 100$ + Feature 1 $\varphi_1 = +30$ + Feature 2 $\varphi_2 = +20$ + Feature 3 $\varphi_3 = -10$ = Model Output $f(x) = 140$

$$100 + 30 + 20 - 10 = 140$$

✓ Local Accuracy Verified: $\varphi_0 + \Sigma\varphi_i = f(x)$