به نام خدا

یادگیری ماشین قابل‌تفسیر

منبع Interpretable Machine Learning (Christoph Molnar)

|  |  |
| --- | --- |
| اهمیت تفسیرپذیری (Importance of Interpretability)  طبقه‌بندی روش‌های تفسیرپذیری (Taxonomy of Interpretability Methods)  دامنه تفسیرپذیری (Scope of Interpretability)  ارزیابی تفسیرپذیری (Evaluation of Interpretability)  خواص توضیحات (Properties of Explanations)  توضیحات انسان پسند (Human-friendly Explanations) | مقدمه |
| رگرسیون خطی (Linear Regression)  رگرسیون لجستیک (Logistic Regression)  مدل خطی تعمیم یافته (Generalized Linear Model)  مدل تجمعی تعمیم یافته (Generalized Additive Model)  درخت تصمیم (Decision Tree)  قوانین تصمیم (Decision Rules)  RuleFit | مدل‌های قابل‌تفسیر  Interpretable Models |
| نمودار وابستگی جزئی (Partial Dependence Plot)  نمودار اثرات محلی انباشته (Accumulated Local Effects Plot)  تعامل ویژگی (Feature Interaction)  تجزیه تابعی (Functional Decomposition)  اهمیت ویژگی جایگشتی (Permutation Feature Importance)  جایگزین کلی (Global Surrogate)  نمونه‌های اولیه و انتقادات (Prototypes and Criticisms) | روش‌های آگنوستیک مدل کلی  Global Model-Agnostic Methods |
| Individual Conditional Expectation  جایگزین محلی (Local Surrogate (LIME))  توضیحات خلاف واقع (Counterfactual Explanations)  Scoped Rules (Anchors)  Shapley Values | روش‌های آگنوستیک مدل محلی  Local Model-Agnostic Methods |