Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

机器学习常用的可解释方法

慕星星　硕士研究生

2020年10月25日

内容提要

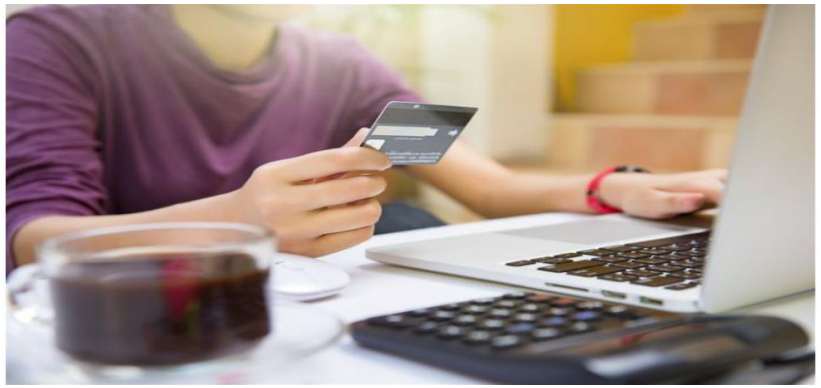
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用介绍
- 参考文献

预期收获

- 预期收获
  - 1. 了解机器学习可解释性的基本概念
  - 2. 了解机器学习可解释性方法的分类
  - 3. 了解常用的机器学习可解释性方法的基本原理和应用（规则提取、LIME和SHAP）

背景简介

- 机器学习的应用



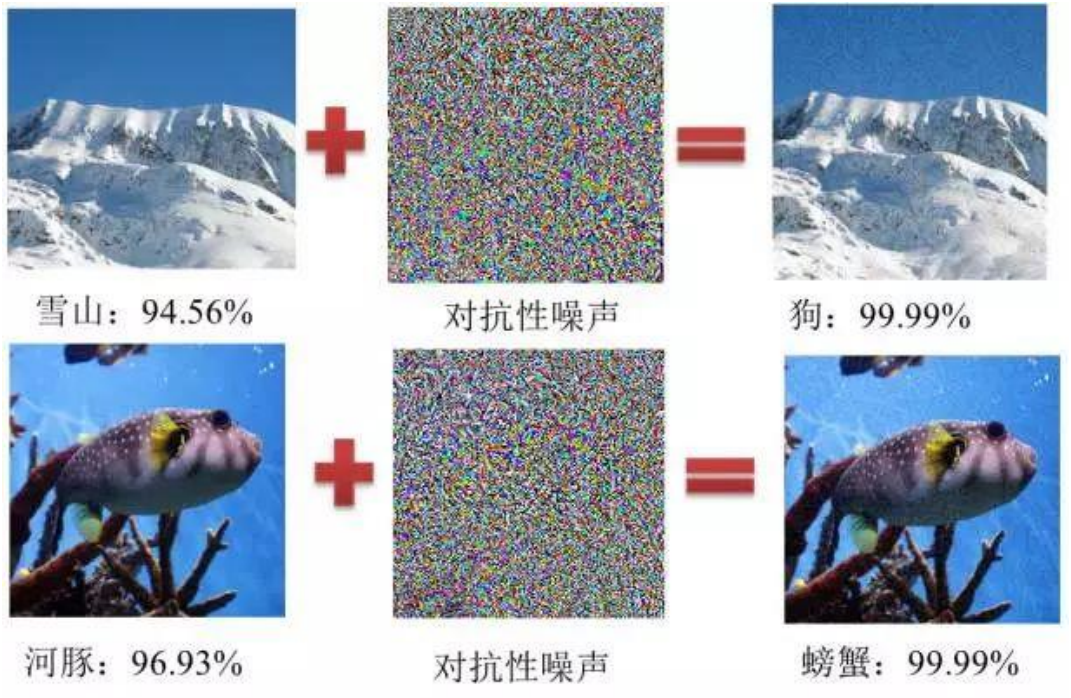
金融领域

医疗领域

法律领域

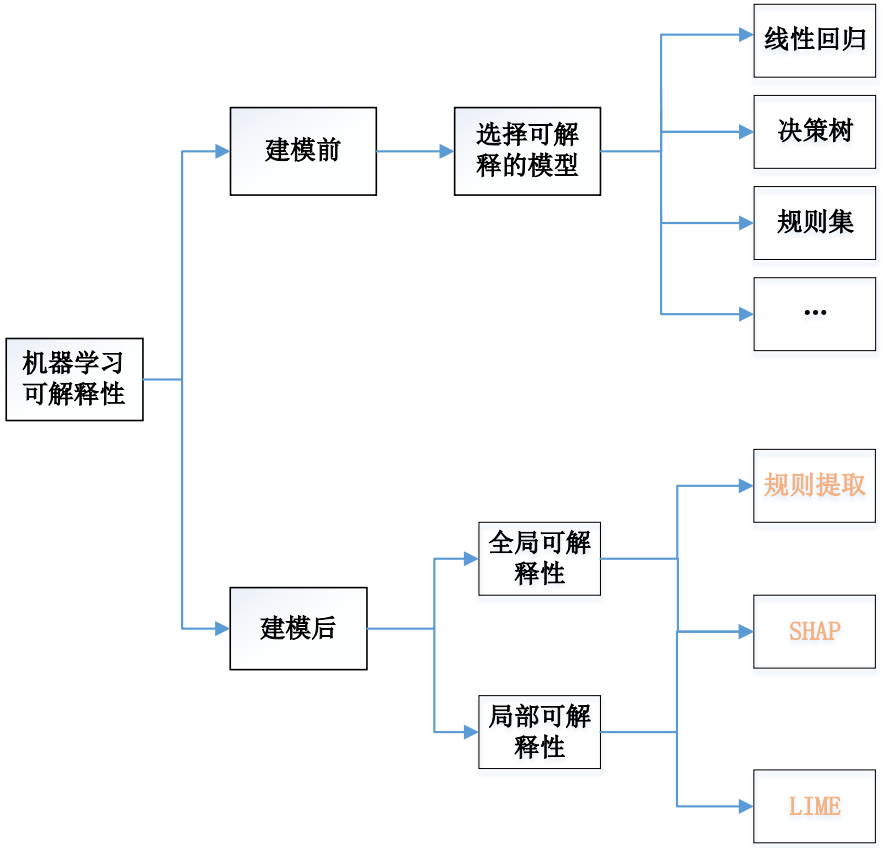
背景概念

- 黑盒问题
  - 黑盒系统自身可能存在着一些主观偏见
  - 黑盒系统抵御攻击的能力差



基本概念

- 机器学习可解释性
  - 可解释性：可解释性是人们能够理解模型决策原因的程度；另一种定义：可解释性是指人们能够一致地预测模型结果的程度。



基本概念

- 可解释模型：由于结构简单且容易理解的机器学习模型。线性回归、逻辑回归和决策树是常用的可解释模型

- 建模后（post-hoc)的可解释方法：在模型训练后运用解释方法或构建解释模型，解释学习模型的工作机制、决策行为和决策依据。
  - 全局可解释性：整体上理解模型背后的复杂逻辑以及内部的工作机制。
  - 局部可解释性：针对每一个特定输入样本的决策过程和决策依据，可以通过分析输入样本的每一维特征对模型最终决策结果的贡献来实现。

背景简介

- 可解释性的意义（重要性）
  - **建模阶段**：辅助开发人员理解模型，进行模型的对比选择，必要时优化调整模型
  - **运行阶段**：向业务方解释模型的内部机制，对模型结果进行解释。比如基金推荐模型，需要解释：为何为这个用户推荐某支基金。
  - 法律规定: **欧盟于2018年5月生效的GDPR(General Data Protection Regulation)中有条例明确规定，当机器针对某个个体作出决定时，该决定必须符合一定要求的可解释性。**

算法原理

规则提取

- 规则提取技术是早期模型可解释性研究的重点，是一种有效的开箱技术，能够提供复杂模型或黑盒模型内部工作机制的深入理解。

- 规则提取技术针对黑盒模型，利用可理解的规则集合生成可解释的符号描述，或从中提取可解释模型（如决策树、基于规则的模型等），使之具有与原模型相当的决策能力。

- 根据解释对象不同，规则提取方法可以分为针对集成树模型的规则提取和针对神经网络的规则提取

规则提取

- 规则提取方法的基本步骤（以集成树模型为例）：
  - 初始规则提取
    - 从原模型逐个中提取规则（根节点到叶子节点的每一条路径都表示一条决策规则）
    - 将提取的规则进行组合得到初始的规则集
  - 规则度量
    - 基于规则长度、规则频率、误差等指标对提取的初始规则进行排序
  - 规则剪枝
    - 基于排序结果，对规则集中的无关项和冗余项进行剪枝
  - 构建规则学习器
    - 基于剪枝后的规则构建一个可解释的规则学习器

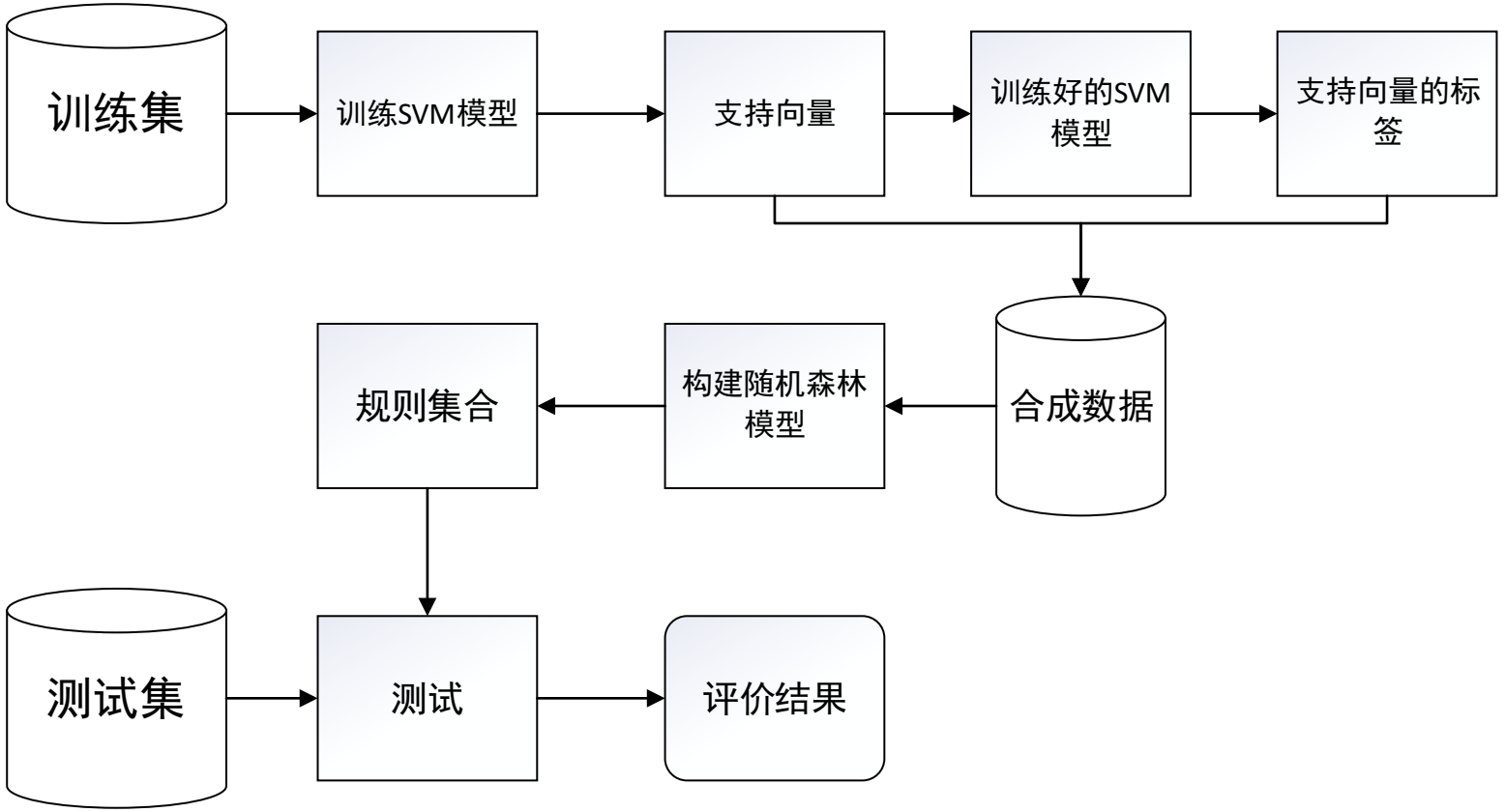
算法原理

T	增强糖尿病风险评估模型的可解释性
I	训练数据和SVM模型
P	1.通过训练集训练SVM，获得支持向量 2.用SVM模型预测支持向量的标签，产生人工合成数据 3.利用人工合成数据训练一个RF模型 4.提取决策树比重最大的前n维作为初始规则集合 5.通过规则度量、冗余规则剪枝，获得最终的规则集合
O	可理解的规则集合

P	如何增强模型的可解释性
C	基于非线性的SVM模型
D	规则提取方法提取的规则不够精确，只能提供近似解释
L	SCI 2区 2015

算法原理框图

- 首先根据训练数据获得SVM模型，然后根据支持向量生成合成数据，并通过决策树的方法进行规则提取。



算法原理

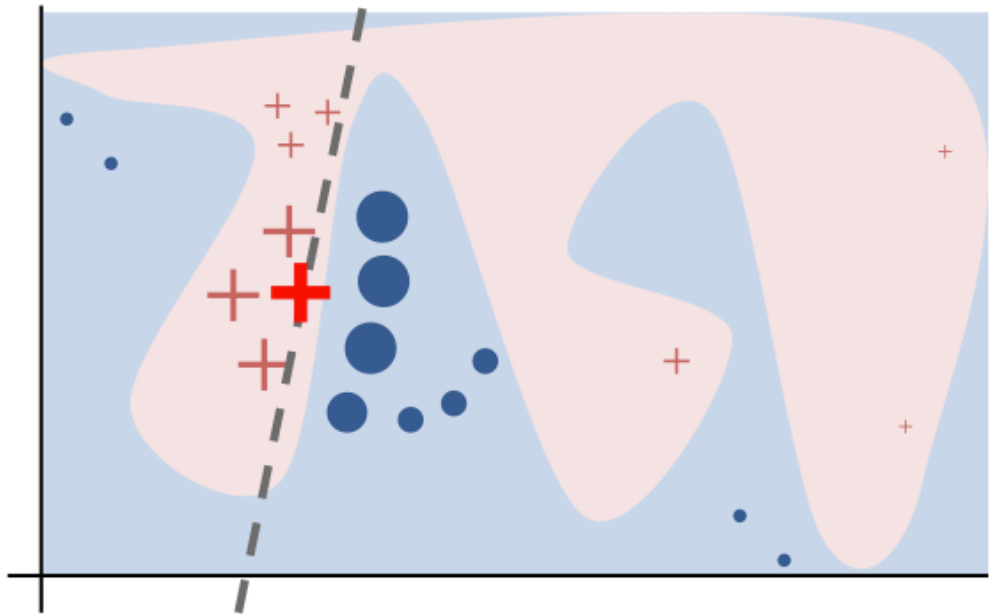
LIME

T	解释黑盒模型对单个样本的预测结果
I	一个待解释的预测样本；训练好的黑盒模型
P	1.重复输入样本同时进行微小扰动，形成样本集 2.用黑盒模型在样本集上进行预测，样本集和预测结果形成新的数据集 3.用简单模型（一般是线性模型）拟合新数据集
O	输出特征重要性（简单模型的权重）

P	解释黑盒模型对单个样本的预测结果
C	需要用户定义样本的邻域范围大小以及简单模型的复杂度
D	解释不稳定，邻域范围不同，得到的局部可解释性模型可能会有很大的差别
L	KDD 2016

LIME

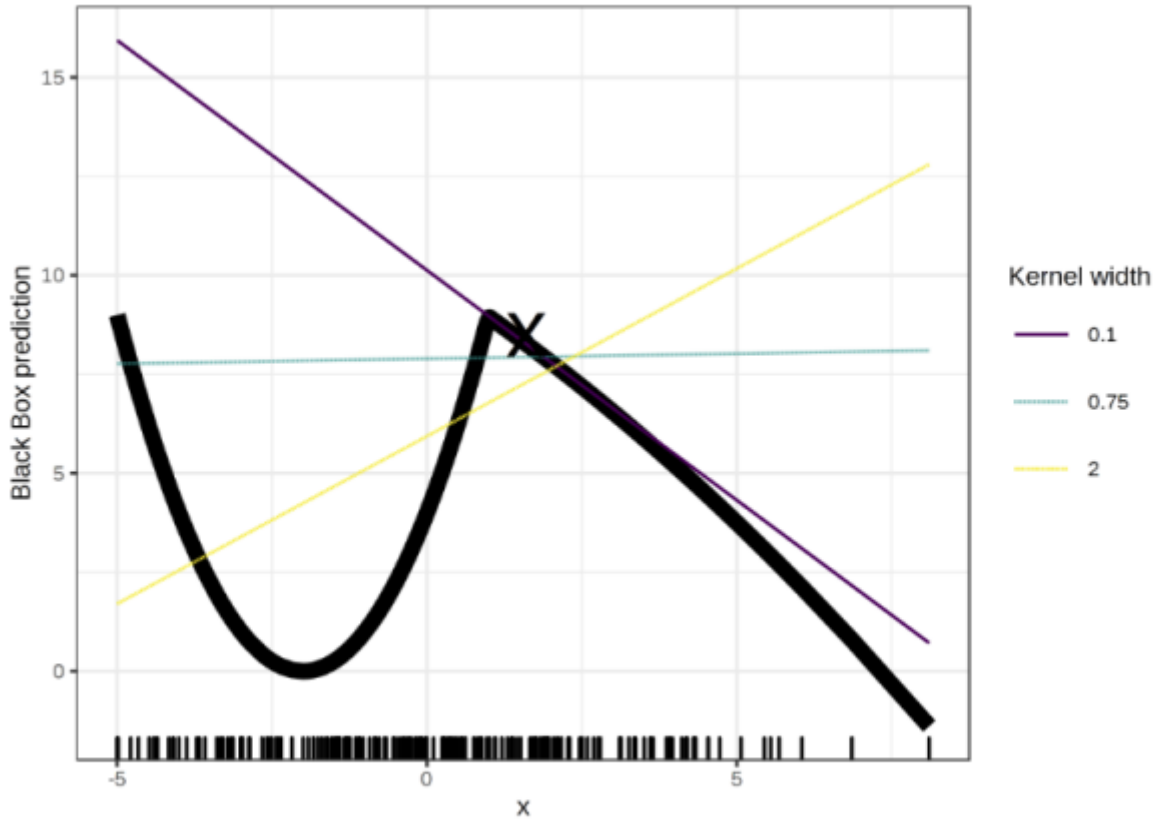
- LIME（Local interpretable model-agnostic explanations)的主要思想是利用可解释性模型（如线性模型，决策树）局部近似黑盒模型的预测，通过对输入进行轻微的扰动，观察黑盒模型的输出变化，根据这种输入到输出的变化，在输入点训练一个可解释性模型。



- LIME算法的具体流程：
  - 选择需要解释的样本实例x
  - 对x进行扰动得到新数据，计算原始模型对新的数据的预测值
  - 求出新数据的权重即数据点与要解释的数据之间的距离
  - 根据新的数据集，预测值和权重训练出简单模型g
  - 利用简单模型g的特征权重对原模型f在x点进行局部解释，公式如下：

$$\text{eplanation}(\text{x}) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

- 需要确定邻域范围：邻域范围不同，得到的局部可解释性模型可能会有很大的差别。



LIME

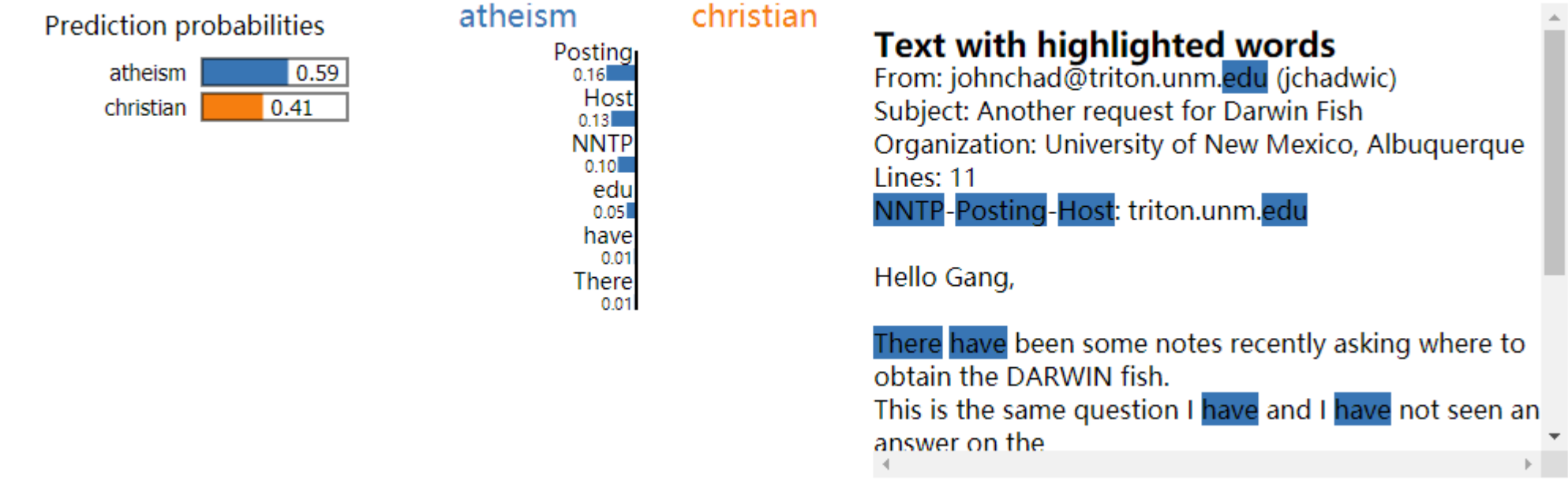
- 优点
  - 原理简单，可解释任何黑盒模型
  - Python中LIME库非常易于使用，适用于表格数据，文本数据和图像数据

- 缺点
  - 需要正确定义邻域和解释模型的复杂度
  - 解释不稳定，在模拟环境中两个非常接近的点的解释差异可能会很大

LIME用于文本分类

- 数据集：The 20 newsgroups text dataset
- 选用的分类：基督教（christian)和无神论（atheism)

```
exp.show_in_notebook(text=True)
```



算法原理

SHAP

T	对模型预测的结果进行解释
I	训练好的模型和训练数据
P	计算每个样本中的每个特征变量的Shapley值
O	Shapley值，能反映出样本中的每个特征的影响力，而且还表现出影响的正负性

P	如何解释模型的输出
C	基于博弈论的最佳Shapley值
D	计算效率低
L	NIPS 2017

SHAP

- SHAP是由Shapley值启发的可加性解释模型。

- 基本原理：根据联盟博弈理论计算每个特征的Shapley值，然后通过每个特征对预测结果的贡献来解释实例的预测。

- Shapley值是联盟博弈论的一种方法，把数据集的每一个特征变量当成一个玩家。用数据集的特征去训练模型得到的预测结果，可以看成众多玩家合作完成一个项目的收益。Shapley值通过考虑各个玩家做出的贡献，来公平地分配合作的收益。

Shapley值

- 特征的Shapley 值是对所有可能的联盟中特征的平均边际贡献。 以求“禁止猫进入”的Shapley值为例：

用于计算“禁止猫进入”的Shapley值的联盟

SHAP

- 优点
  - 在博弈论中具有扎实的理论基础，预测结果在特征值中公平分配
  - 不仅考虑单个变量的影响，而且考虑变量组的影响，变量之间可能存在协同效应

- 缺点
  - 计算效率低，需要大量时间（Shapley值的精确计算成本很高）
  - 不能用于对输入变化的预测做出变化的陈述

利用SHAP解释Xgboost模型

- 数据集：2018年足球球员身价数据

In [4]:
```
data.head()
```

Out[4]:

	id	club	league	birth_date	height_cm	weight_kg	nationality	potential	pac	sho	...	st	lw	cf	cam	cm	cdm	cb	lb	gk	y
0	0	293	25	10/4/96	177	72	78	73	65	60	...	63.0	64.0	64.0	64.0	63.0	57.0	53.0	56.0	NaN	70.0
1	1	258	24	9/21/84	178	70	51	62	56	39	...	52.0	60.0	57.0	59.0	61.0	64.0	61.0	64.0	NaN	24.0
2	2	112	3	6/8/99	177	69	52	68	68	57	...	56.0	54.0	55.0	53.0	45.0	34.0	31.0	36.0	NaN	17.0
3	3	604	9	7/25/88	181	81	54	81	76	74	...	77.0	76.0	77.0	77.0	79.0	78.0	77.0	78.0	NaN	1750.0
4	4	80	37	8/4/80	179	75	96	72	40	62	...	62.0	66.0	65.0	68.0	71.0	70.0	66.0	64.0	NaN	97.5

5 rows × 65 columns

In [10]:
```
# 选择特征，这里只是举例，未必是最佳组合
# 特征依次为身高（厘米）、潜力、速度、射门、传球、带球、防守、体格、国际知名度、年龄
cols = ['height_cm', 'potential', 'pac', 'sho', 'pas', 'dri', 'def', 'phy', 'international_reputation', 'age']
```

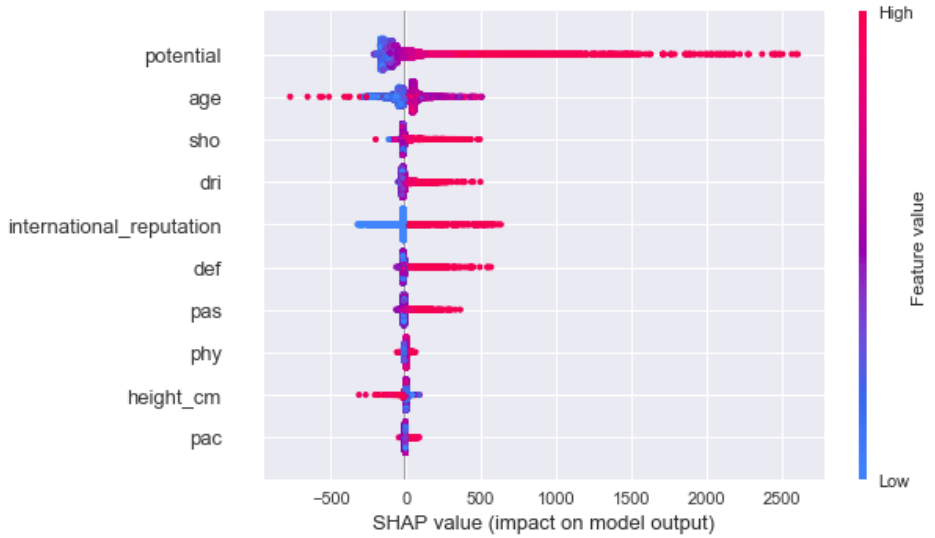
In [12]:
```
#训练xgboost回归模型
model=xgb.XGBRegressor(max_depth=4,learning_rate=0.05,n_estimators=150)
model.fit(data[cols],data['y'].values)
```

Out[12]:
```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.05, max_delta_step=0, max_depth=4,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=150, n_jobs=0, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```

利用SHAP解释Xgboost模型

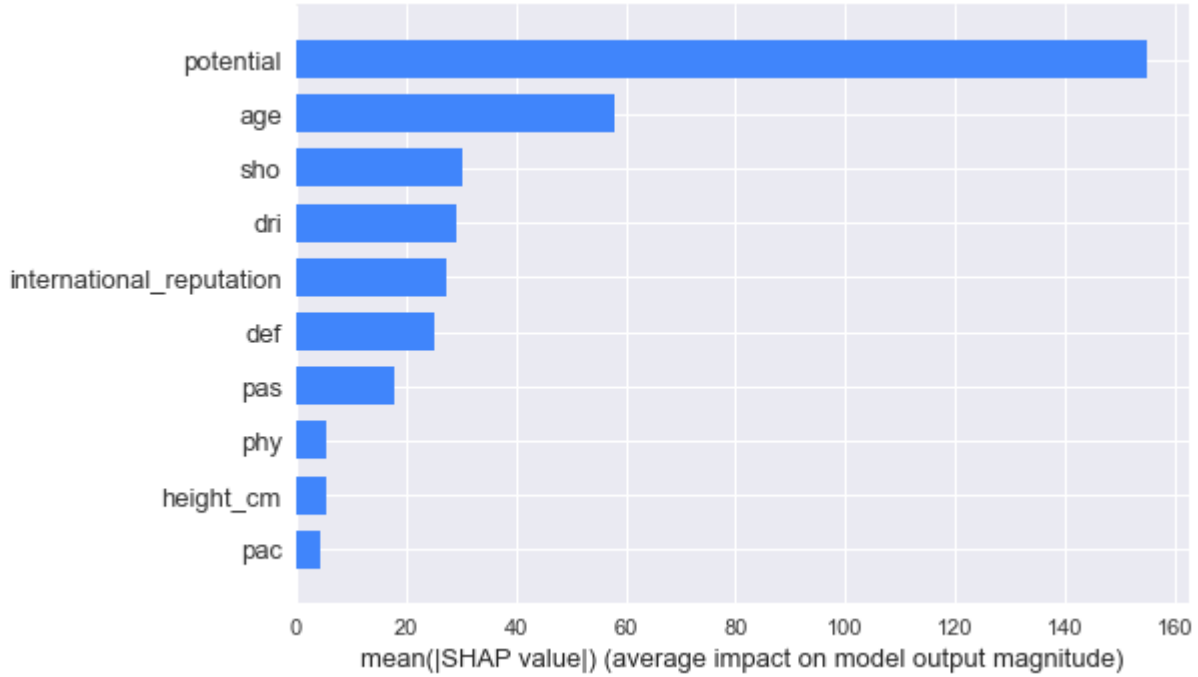
- SHAP具有强大的数据可视化功能
  - 单个样本的解释（对一个球员预测结果的展示）

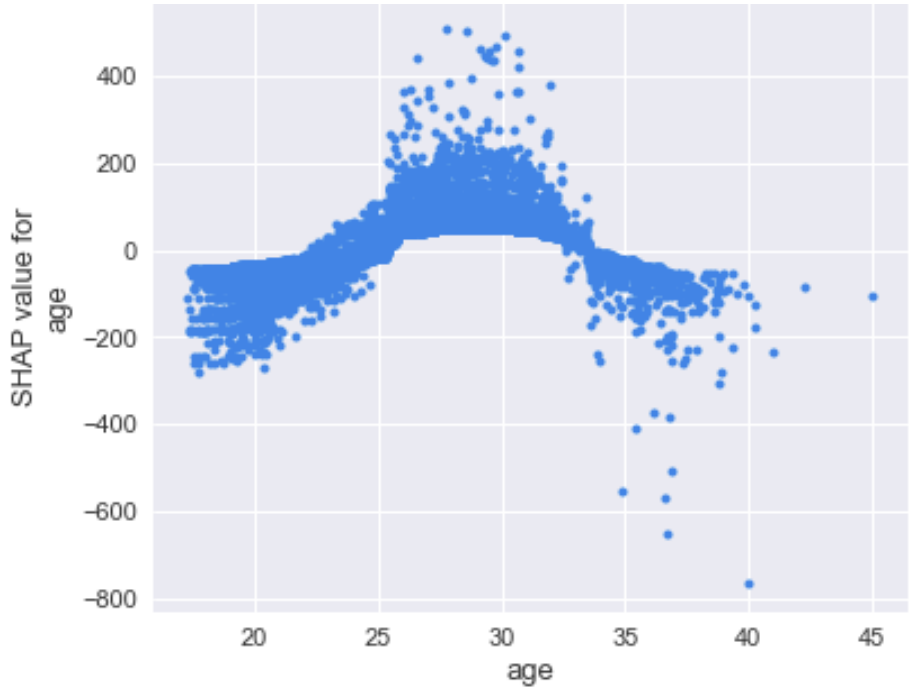
  - 对特征的总体分析



利用SHAP解释Xgboost模型

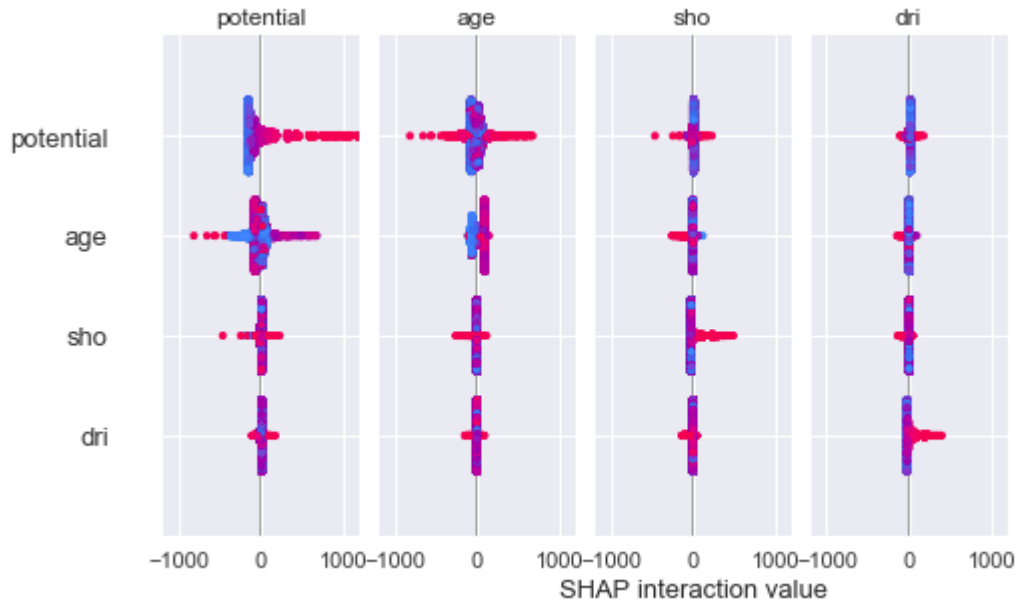
- SHAP特征重要性：具有较大的Shapley绝对值的特征很重要
- SHAP依赖图：显示一个特征的每个数据实例对应的Shapley值

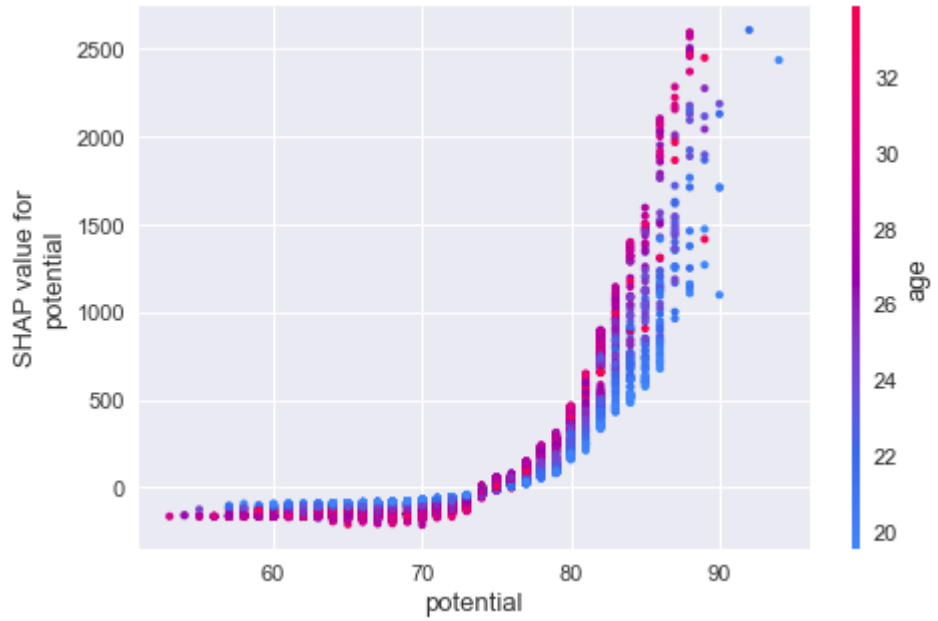




利用SHAP解释Xgboost模型

- 多变量交互作用分析





总结

- 规则提取
  - 模型可解释性早期研究的主要方法，是一种有效的开箱技术
  - 提取的规则往往不够精确，只能提供近似解释
  - 提供的可解释性的质量受规则本身复杂度的制约
- LIME
  - 模型无关的局部可解释性方法
  - LIME作为具体实现的局部代理模型非常有应用前景，但是需要解决正确定义邻域和解释不稳定等问题，才能安全应用
- SHAP
  - 具有扎实的理论基础，预测结果在特征值中公平分配
  - 计算精确Shapley值的效率低

参考文献

- [1]Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning,volume 1. Springer series in statistics New York, 2001.
- [2] Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?" : Explaining the Predictions of Any Classifier[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:1135-1144.
- [3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in neural information processing systems, pages 4765 - 4774, 2017.
- [4] Han, Longfei , et al. "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes." Biomedical and Health Informatics, IEEE Journal of 19.2(2015):728-734.

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！