阿里云天池牛年读书会

可解释机器学习

分享嘉宾：朱明超
复旦大学研究生

天池读书会

TIANCHI天池

Broadview®

www.broadview.com.cn

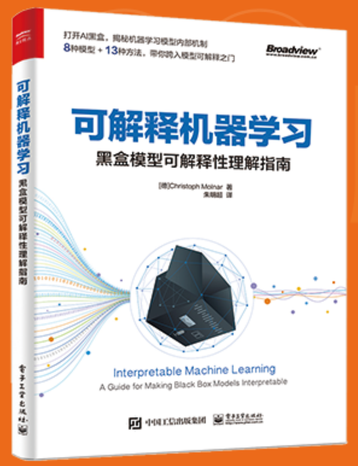
《可解释机器学习：黑盒模型可解释性理解指南》

全面介绍了可解释模型、黑盒模型的可解释性、与模型无关的方法

包含各种解释方法优缺点，以及每种方法的软件实现。

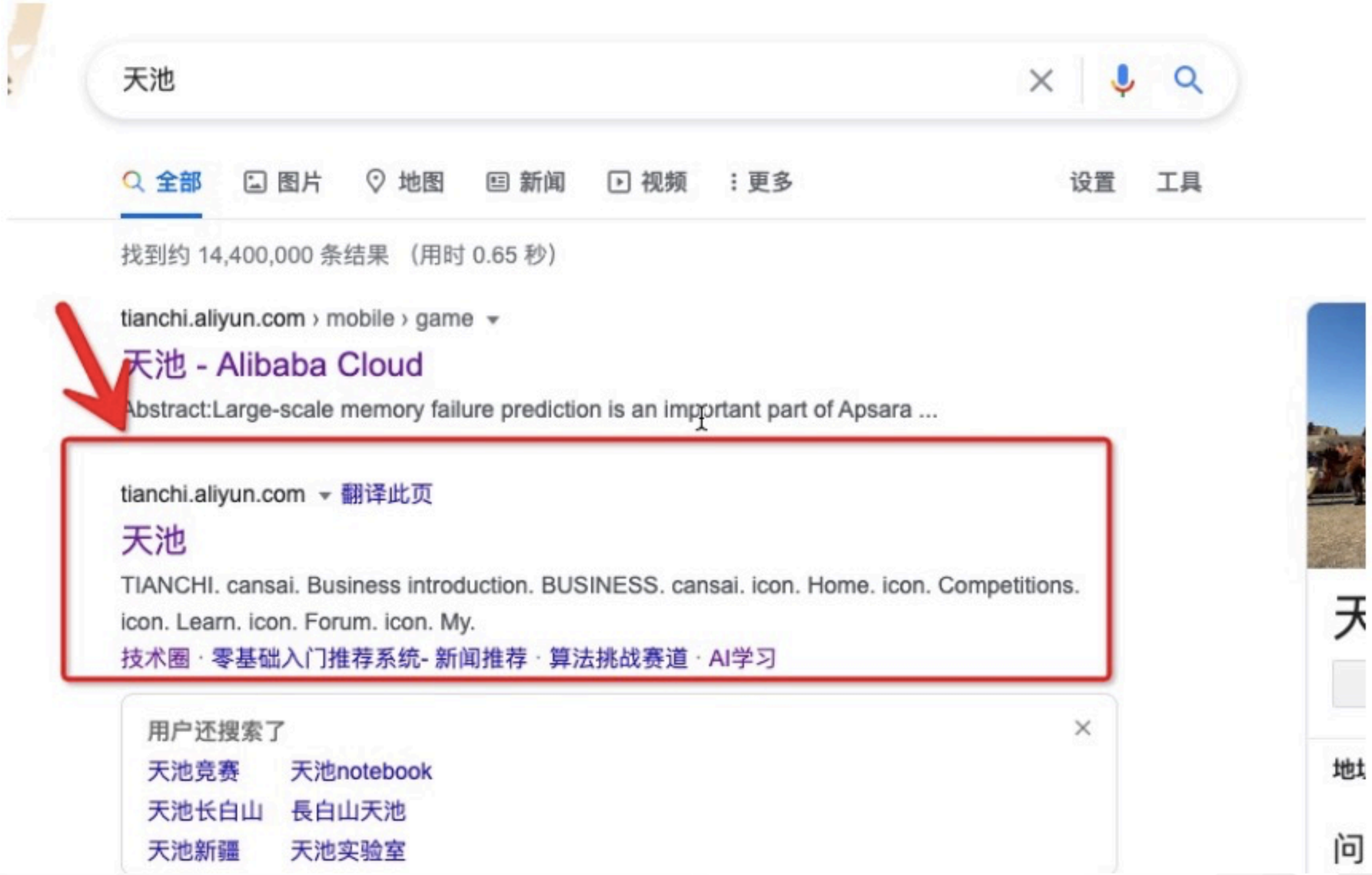
直播嘉宾：朱明超

直播时间：4月21日20:00



扫码领取读书会配套学习资源

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

朱明超 本书译者、复旦研究生

直播主题 《可解释机器学习：黑盒模型可解释性理解指南》

直播时间 2021年4月21日 20:00

学习资料 机器学习训练营

实践项目 待定

提问 | 学习训练营 | 购买地址 | PPT下载 | 实战代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

目录

TIANCHI天池

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

分享嘉宾简介

TIANCHI天池

Christoph Molnar

德国慕尼黑大学可解释机器学习博士
可解释机器学习领域著名人物之一

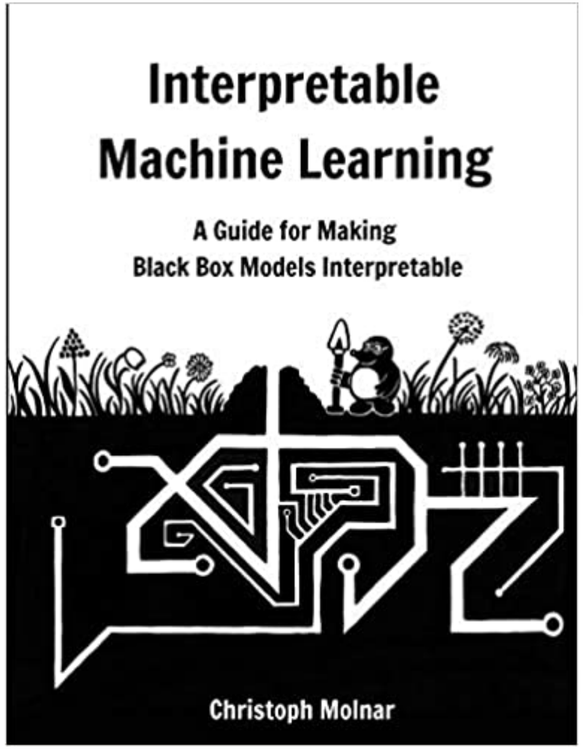
朱明超

复旦大学可解释机器学习研究生

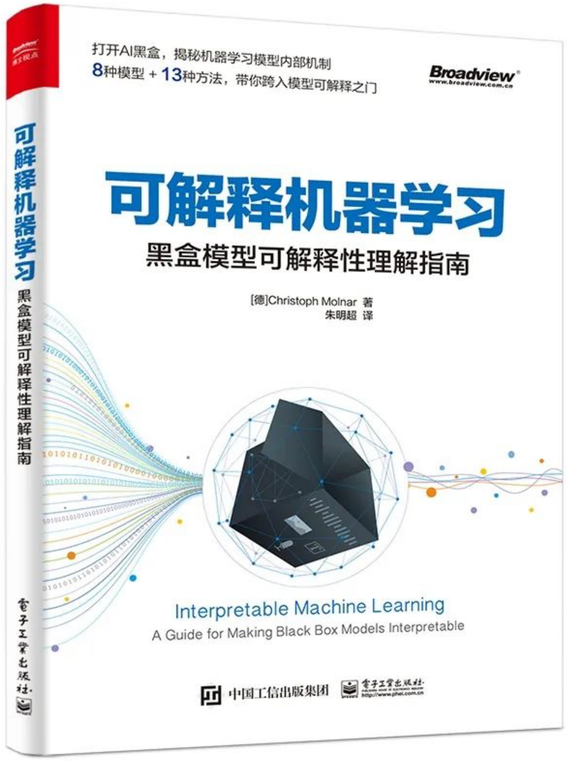
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书简介

TIANCHI天池



全球首本**可解释机器学习书籍**



国内首本**可解释机器学习书籍**

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书简介

TIANCHI天池

FloydHub评定为**2020年最佳机器学习书籍之一**

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书简介

TIANCHI天池

斯坦福大学**课程教材**

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Textbooks

- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning, 2018.
- Molnar, Christoph. Interpretable machine learning, 2019.

发展历史

~1800 线性模型

~1960 基于规则的模型（AID,CHAID,ID3,CART,关联规则）

~ 2001 随机森林

~ 2012 深度学习

~ 2016 可解释机器学习，可解释人工智能

- Gartner已将可解释人工智能（XAI）技术列为数据和分析技术领域的TOP10重要趋势之一
- 2017年，美国国防部资助的美国国防高级研究计划局（DARPA）发起了XAI计划，全面开展了对可解释机器学习的研究
- 2018年，欧洲委员会向欧洲议会、欧洲理事会等欧洲官方机构发布的有关欧洲人工智能的函件中强调对可解释机器学习的需求
- 微软、谷歌、Oracle等诸多科技巨头，都在开展可解释机器学习相关技术的研发

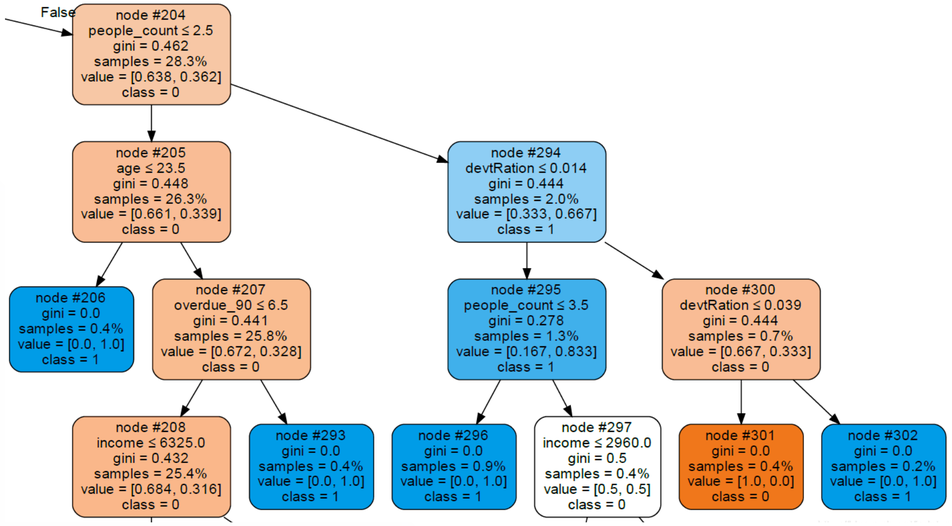
图书简介

TIANCHI天池

谷歌趋势

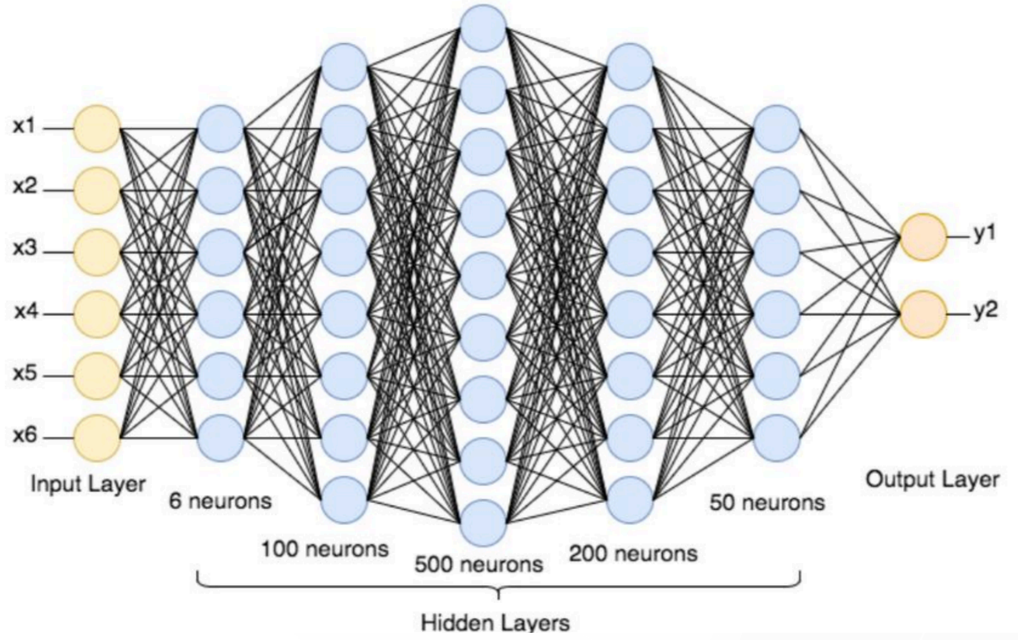
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书内容



可视化树模型

工具：Graphviz



神经网络模型

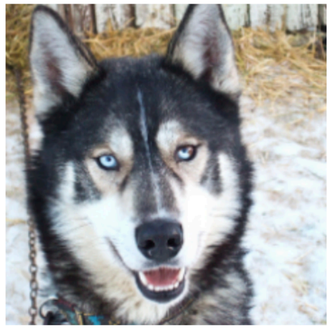
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

为什么我们需要可解释机器学习？

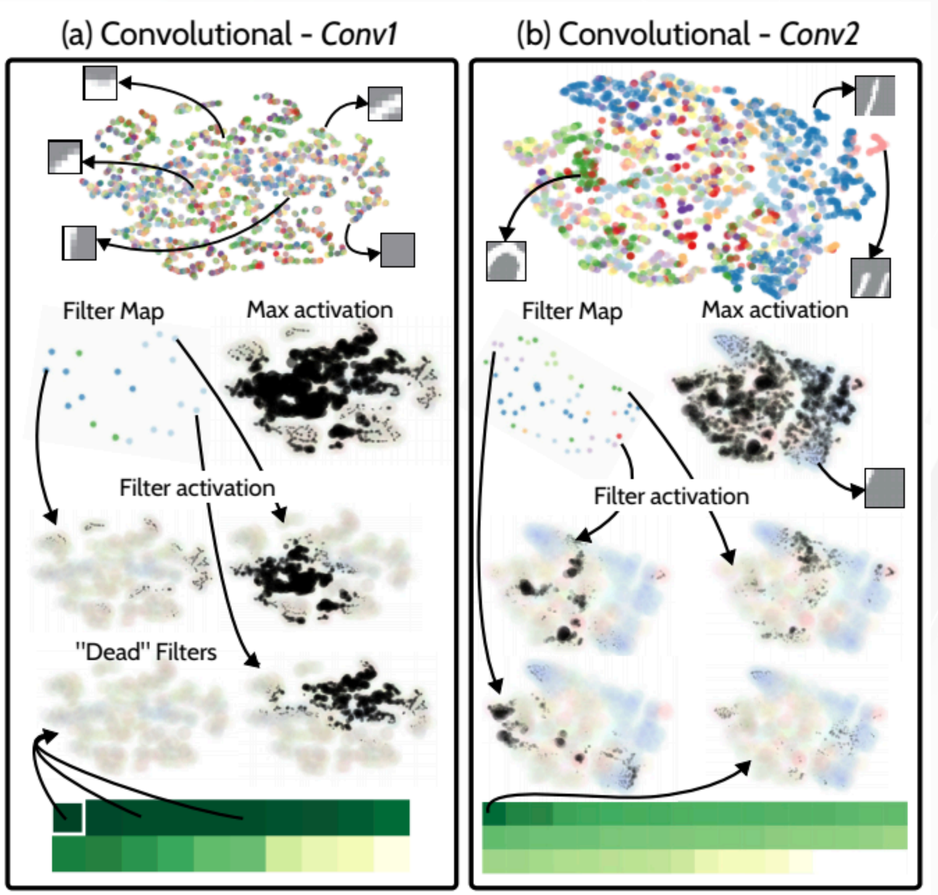
- 社会对AI的依赖性：无人驾驶、安全、金融
- 用户：增强信任、理解决策的后果，例如：隐私性、公平性
- 监管机构：遵循法规、审计和责任
- 模型设计者：诊断和调试模型
- 科学：科学知识发现

调试

- 哮喘和肺炎之间的奇怪联系
- ID是预测过程中最重要的特征
- 狼和狗的分类中，背景中的雪是最重要的特征
- 识别退化过滤器



图源于“ “Why Should I Trust You?” Explaining the Predictions of Any Classifier”



图源于“DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks”

图书内容

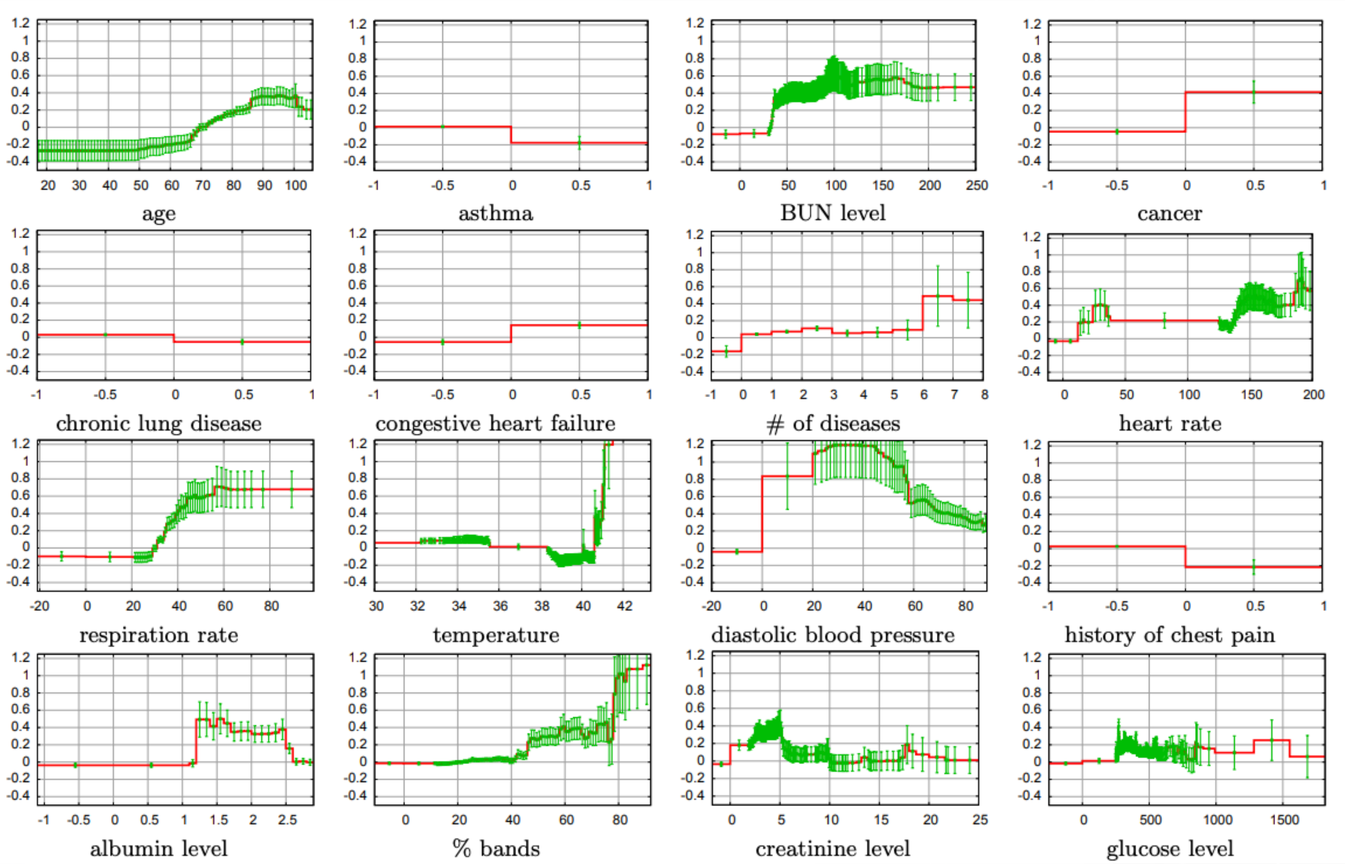
TIANCHI天池

预测肺炎患者的死亡概率

- 高概率->医院/ICU
- 低概率->门诊治疗

模型出错

- 数据集中存在哮喘的偏差
- 哮喘是一种严重的情况
  必须住院甚至重症监护室
- 90年代中期，哮喘的神经网络错误
  阻止了临床试验

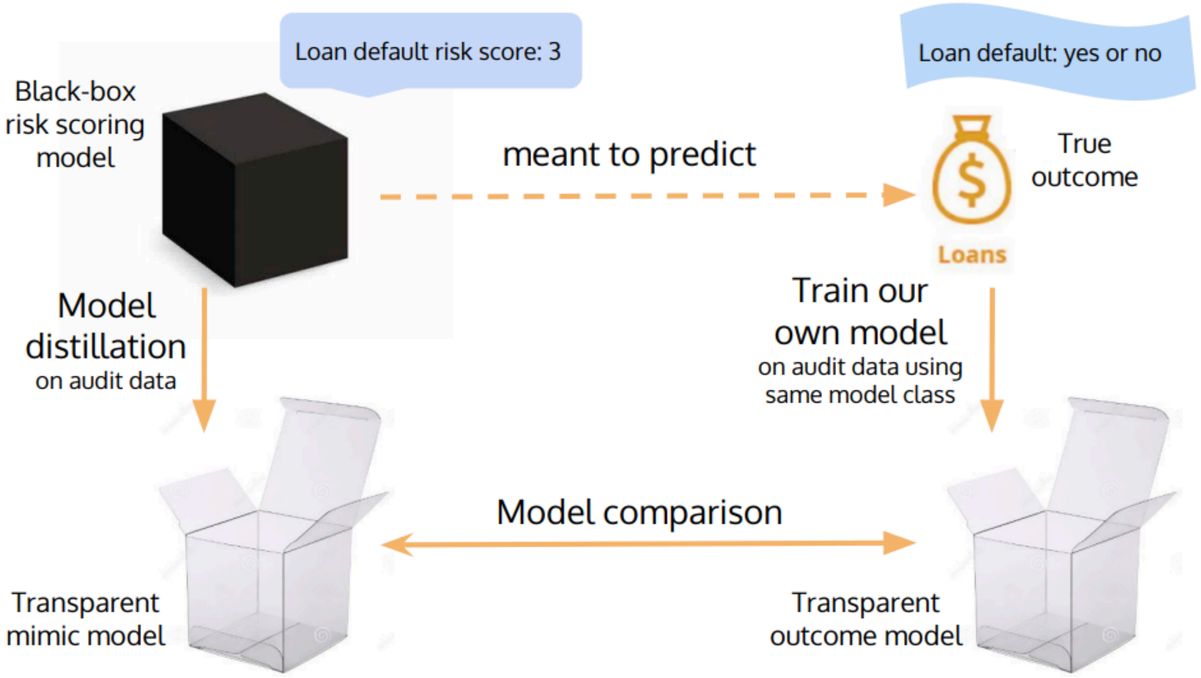


图源于“Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission”

图书内容

TIANCHI天池

法律软件 COMPAS



图源于“Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation”

一起案件对 COMPAS 提出挑战

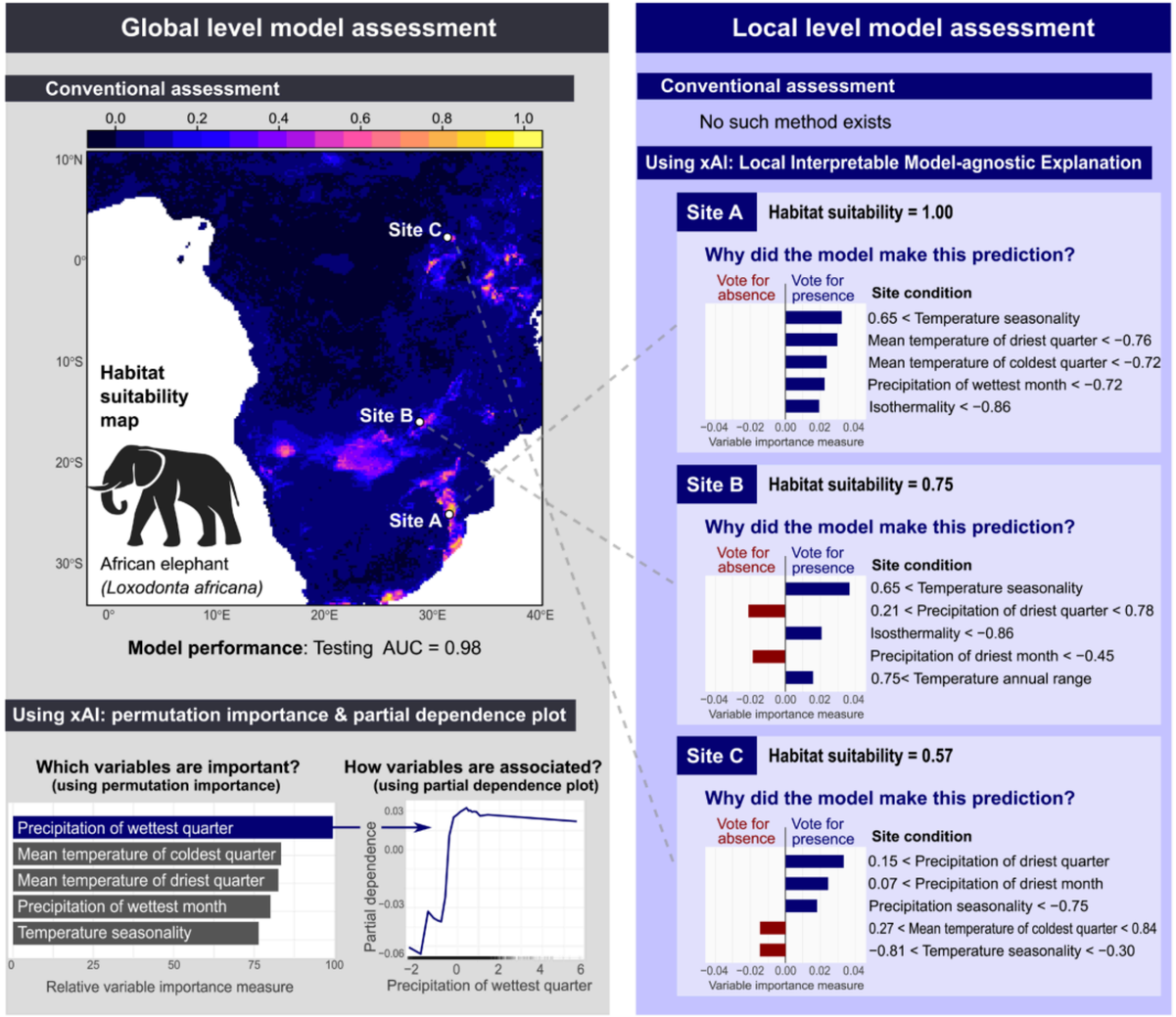
- 在决策过程中使用性别和种族特征
- 算法细节属于商业机密，不透明

案件内容“Damned Lies & Criminal Sentencing Using Evidence-Based Tools”

解释 COMPAS

- 模型蒸馏的思想，学习模型行为
- 可解释模型用于解释黑盒模型的行为

图书内容



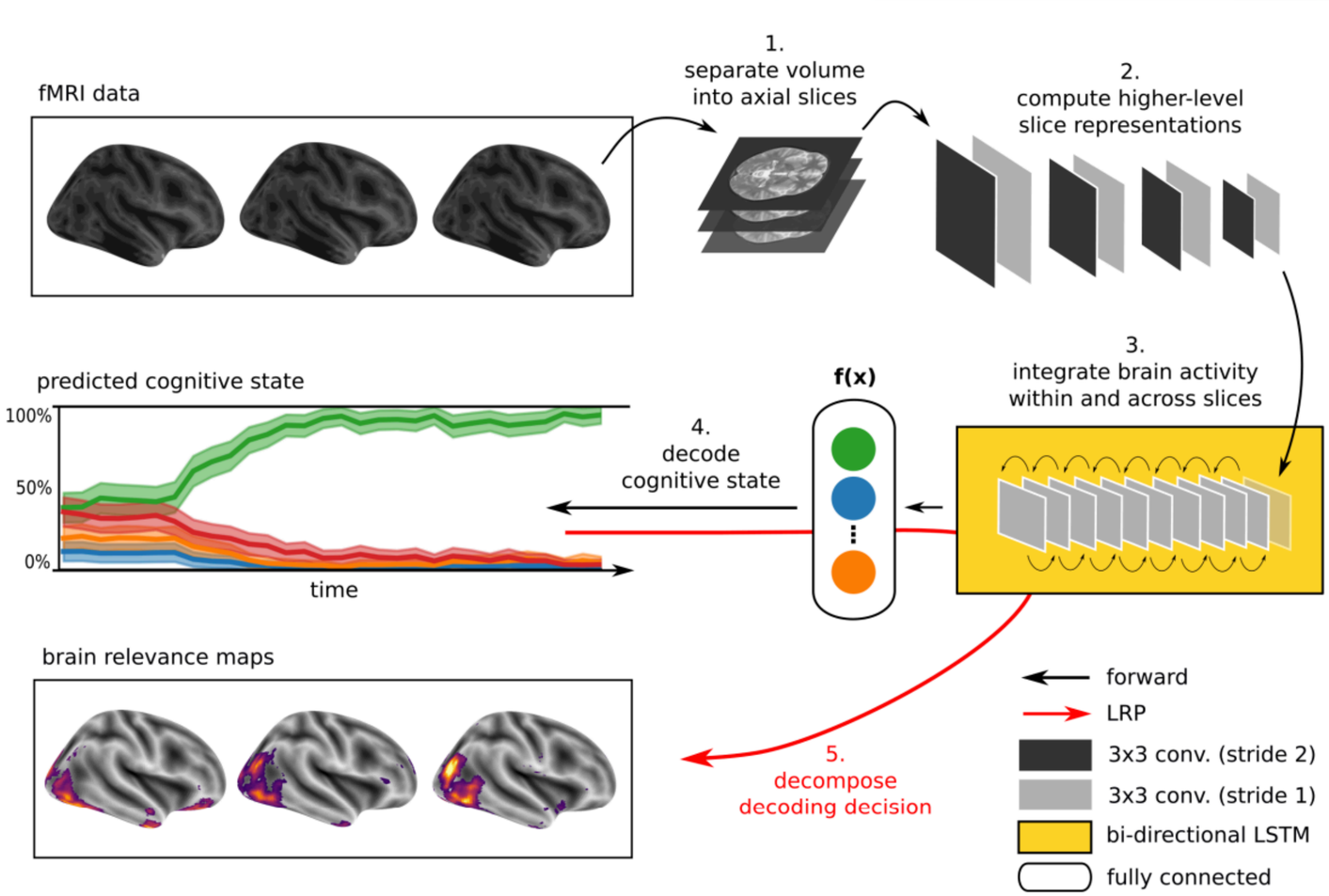
（科学）见解

使用XAI工具解释非洲象🐘

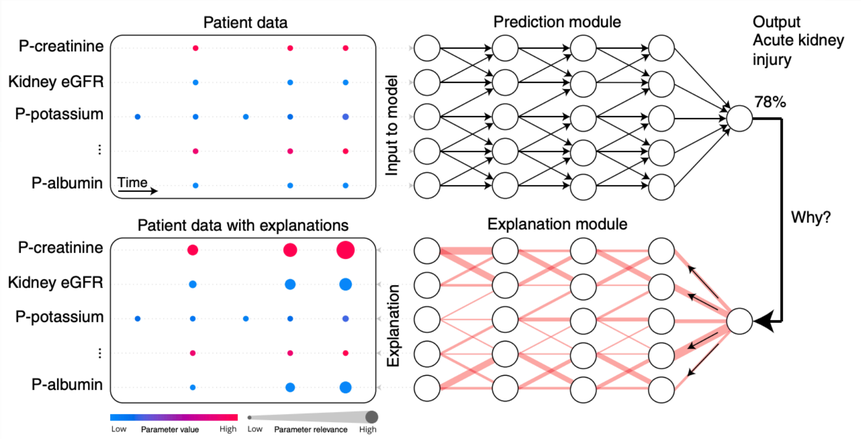
- 全局性的置换重要性、部分依赖图
- 局部性的LIME

图源于“Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models”

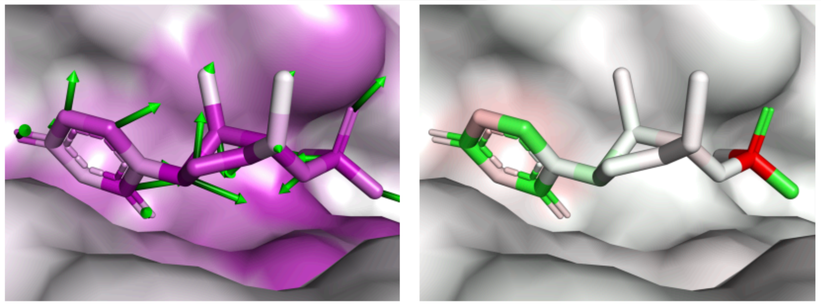
图书内容



图源于“Analyzing Neuroimaging Data Through Recurrent Deep Learning Models”



图源于“Explainable artiÞcial intelligence model to predict acute critical illness from electronic health records”



图源于“Visualizing Convolutional Neural Network Protein-Ligand Scoring”

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

可解释机器学习分类

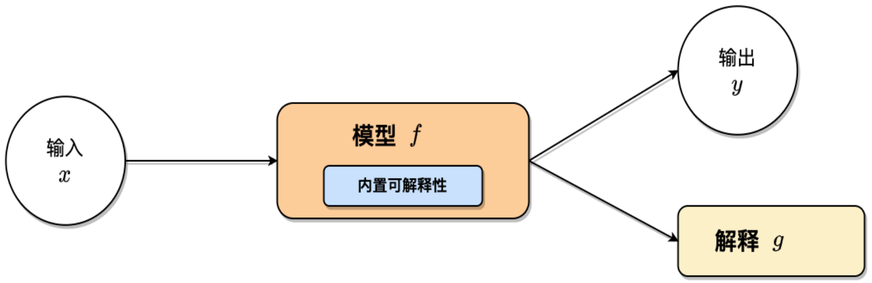
1. 内在/内置可解释性和事后可解释性

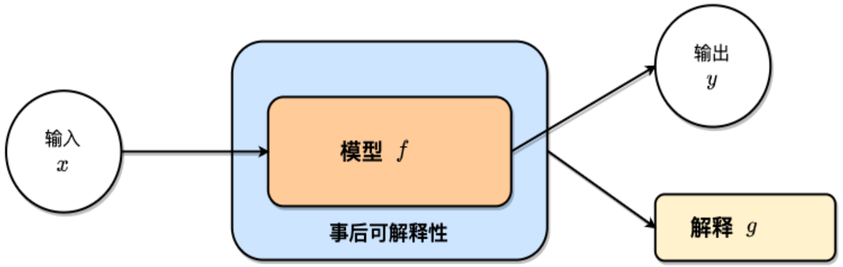
- 内在/内置可解释性

  -> 可解释性通过模型设计实现

  -> 模型本身是可以解释的

  -> 可解释性通常是模型训练的副产品





- 事后可解释性

  -> 模型是在模型训练后得到的

  -> 模型通过外部方法实现的

可解释机器学习分类

2. 特定于模型的解释和与模型无关的解释

- 特定于模型的解释
  - -> 用于某个特定架构的技术
  - -> 可能会损害模型的性能
  - -> 需要使用数据集训练模型
  - -> 根据定义，内置可解释性方法是特定于模型的

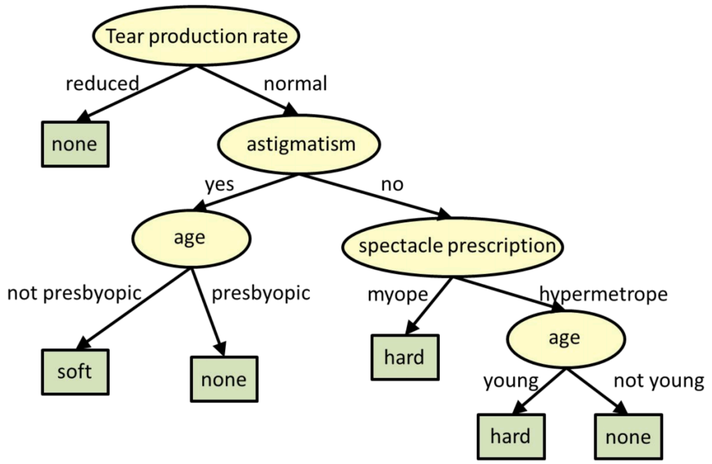
- 与模型无关的解释
  - -> 可以跨许多黑盒模型使用的技术
  - -> 与模型无关的解释方法不会影响模型的性能
  - -> 不需要训练模型
  - -> 事后解释方法通常是与模型无关的

可解释机器学习分类

3. 全局解释和与局部解释

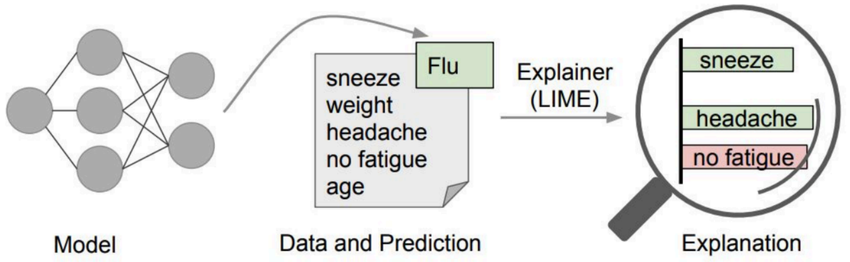
- 全局解释

  -> 解释整个模型行为



- 局部解释

  -> 解释单个实例的模型预测（单个实例或者一组实例）



内置可解释性的模型

线性回归模型

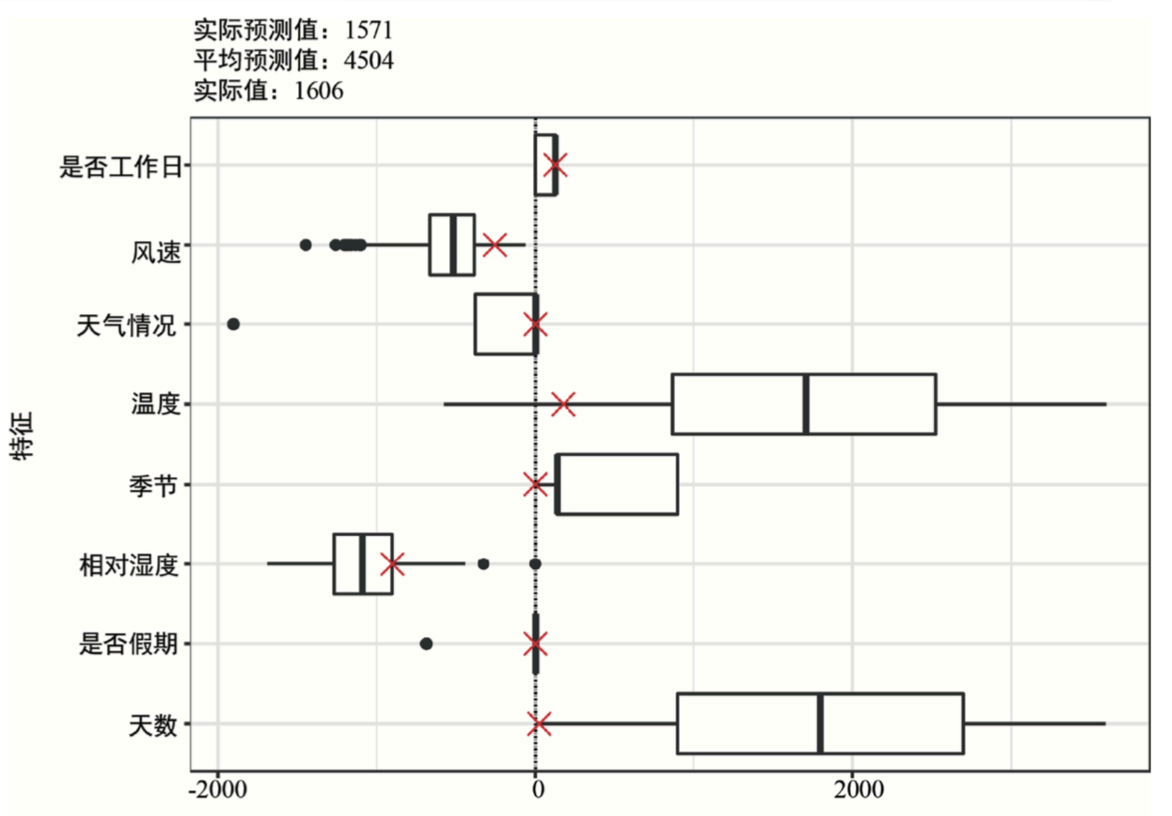
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

可解释组件：β_j

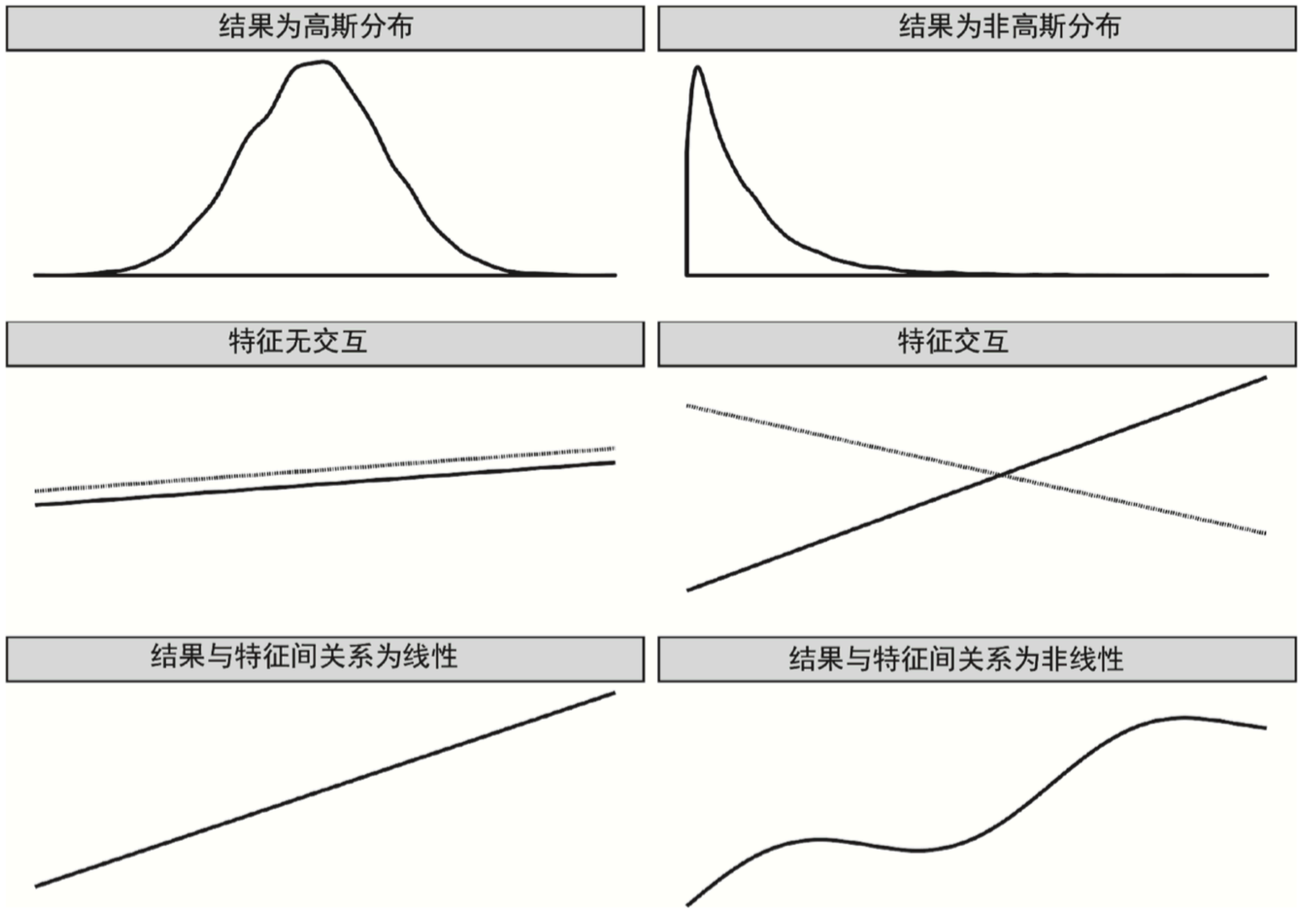
逻辑回归模型

$$P(y^{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}))}$$

可解释组件：几率



图书内容



->广义线性模型

->特征交互

->广义加性模型

图书内容

TIANCHI天池

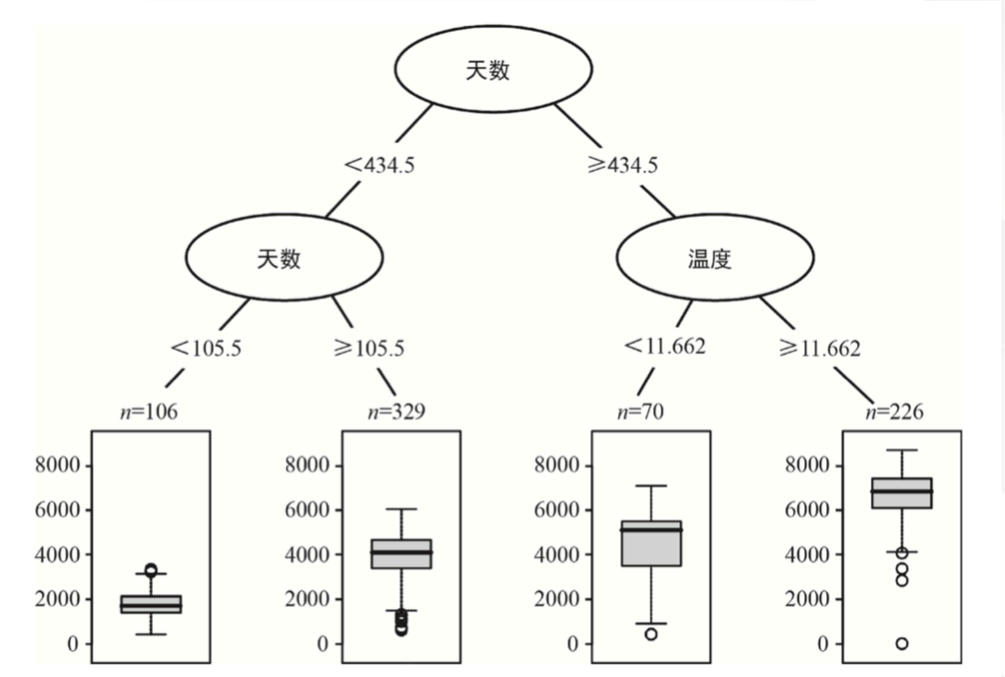
决策规则

即 IF-THEN 规则，学习 IF-THEN 规则的三种代表性方法：

- OneR： 基于单个特征学习规则
- 顺序覆盖：迭代地学习规则并删除新规则覆盖的数据点
- 贝叶斯规则列表：使用贝叶斯统计将预挖的频繁模式组合到决策列表中

RuleFit模型、贝叶斯模型、k近邻模型

决策树模型

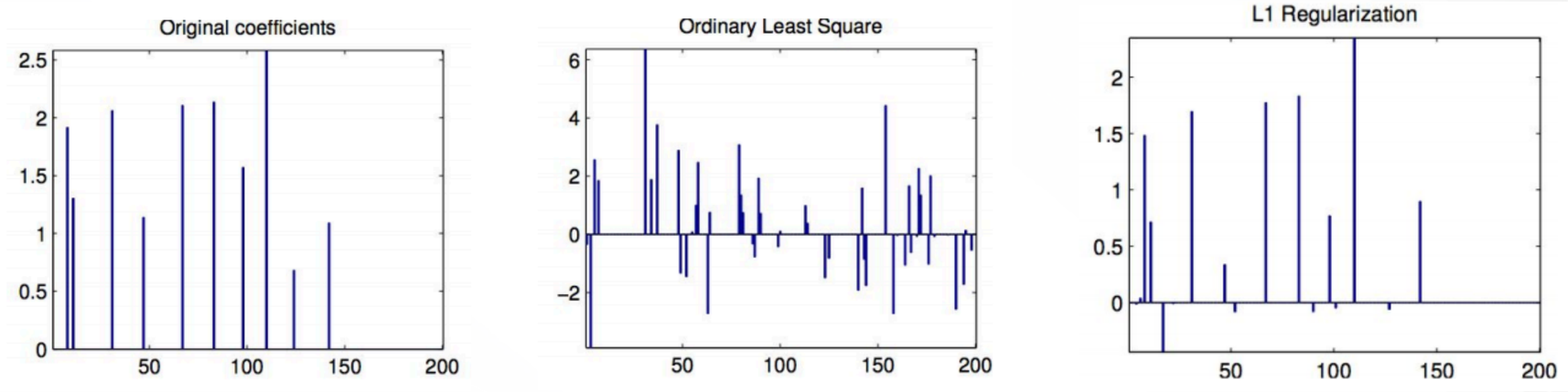


稀疏性

可以控制学习模型时的稀疏性，考虑线性模型的情况下：

- Lasso

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{x}^{(i)\top} \boldsymbol{\beta})^2 \right) + \lambda \|\boldsymbol{\beta}\|_1$$



与模型无关的解释方法

- 部分依赖图
- 个体条件期望图
- 累积局部效应图
- 特征交互
- 置换特征重要性
- 全局代理模型
- 局部代理-LIME
- Anchors (*)
- Shapley 值
- SHAP

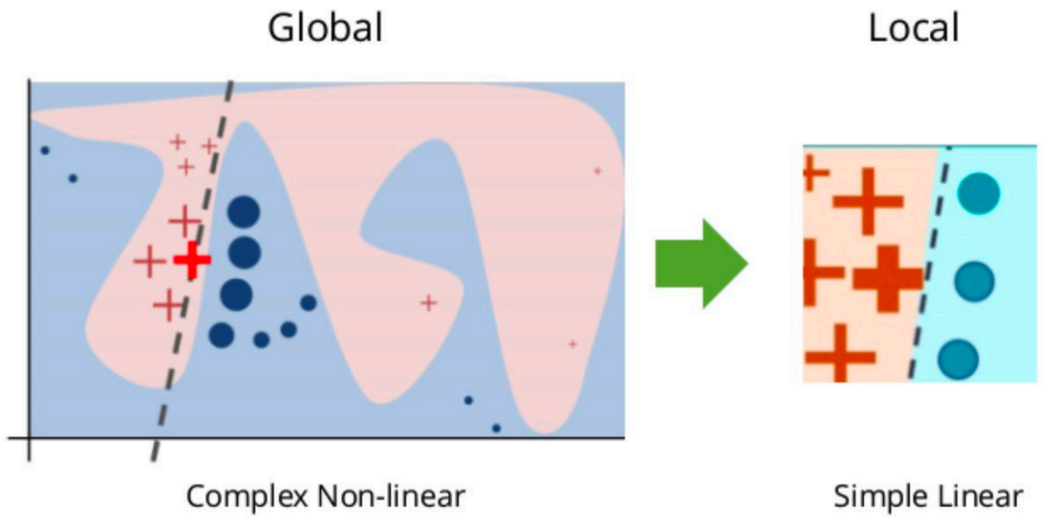
图书内容

TIANCHI天池

LIME

局部代理模型

- 局部代理方法的目的是根据模型保真度寻找解释 g 来近似目标函数 f 在实例 x 附近的行为
- 使用简单的可解释模型（例如线性模型）来寻求复杂模型的局部可解释性



图源于“Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission”

图书内容

TIANCHI天池

LIME

线性可解释的模型

- 我们用线性模型解释，$g(z') = w_g \cdot z'$ ，
- z′ 是特征掩码，指示扰动样本中是否包含该特征
- 基于扰动掩码得到扰动样本 z

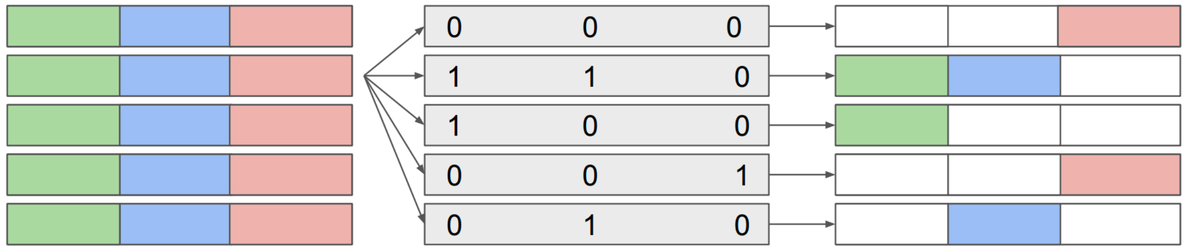
黑盒模型　　解释

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

局部代理损失　　接近度　　复杂度惩罚项

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

采样过程

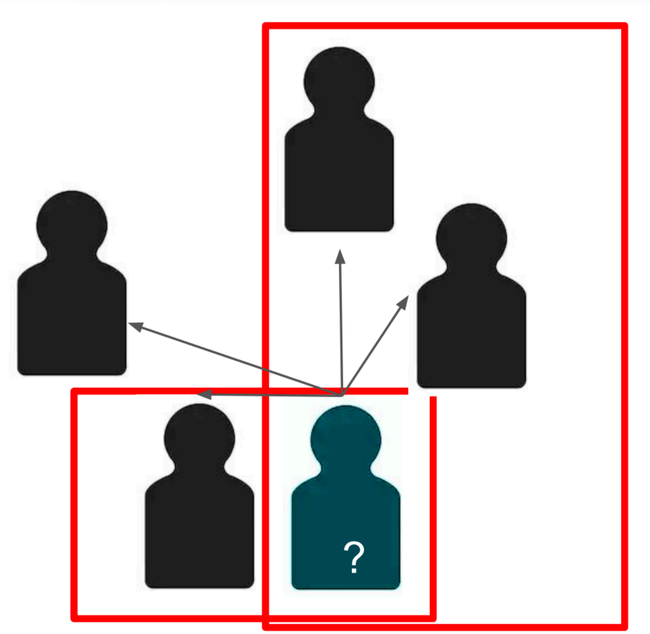


被解释的实例 x　　扰动掩码 z′　　扰动样本 z

Shapley 值

我们如何公平地给玩家分配重要性分数?

- 考虑与所有其他玩家的交互



图书内容

TIANCHI天池

要求

- 效益性。特征贡献必须加起来等于实例预测和平均预测的差。
- 对称性。如果两个特征值 x_j 和 x_k 对所有可能的联盟均贡献相同，则它们的 Shapley 值应相同。
- 虚拟性。如果将特征 j 的特征值 x_j 无论添加到哪个联盟中都不会改变预测值， 则它的 Shapley 值应为 0。
- 可加性或线性。

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{S!(M-S-1)!}{M!} \left(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})\right)$$

图书内容

Kernel SHAP

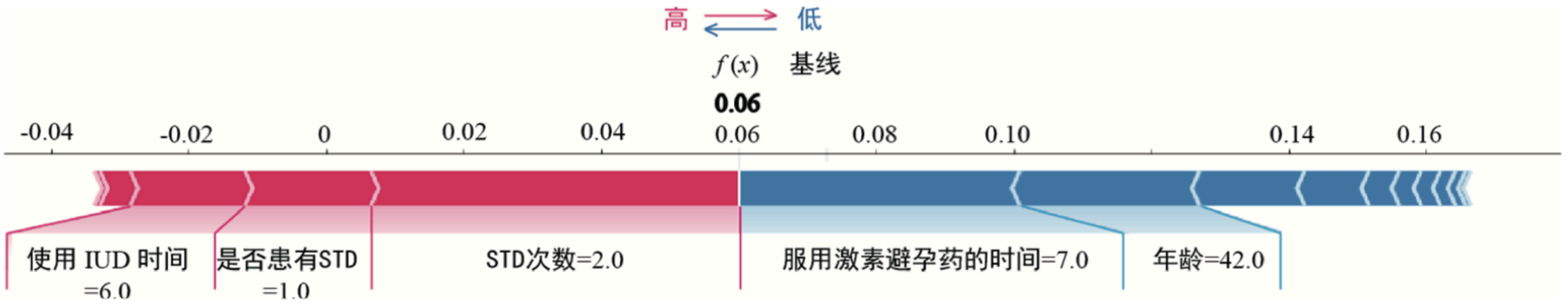
- 把每个特征 i 当作一个玩家来对待
- 通过 Shapley 值估计特征 i 的值

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x(z')) - g(z') \right]^2 \pi_{x'}(z')$$

$$\Omega(g) = 0$$

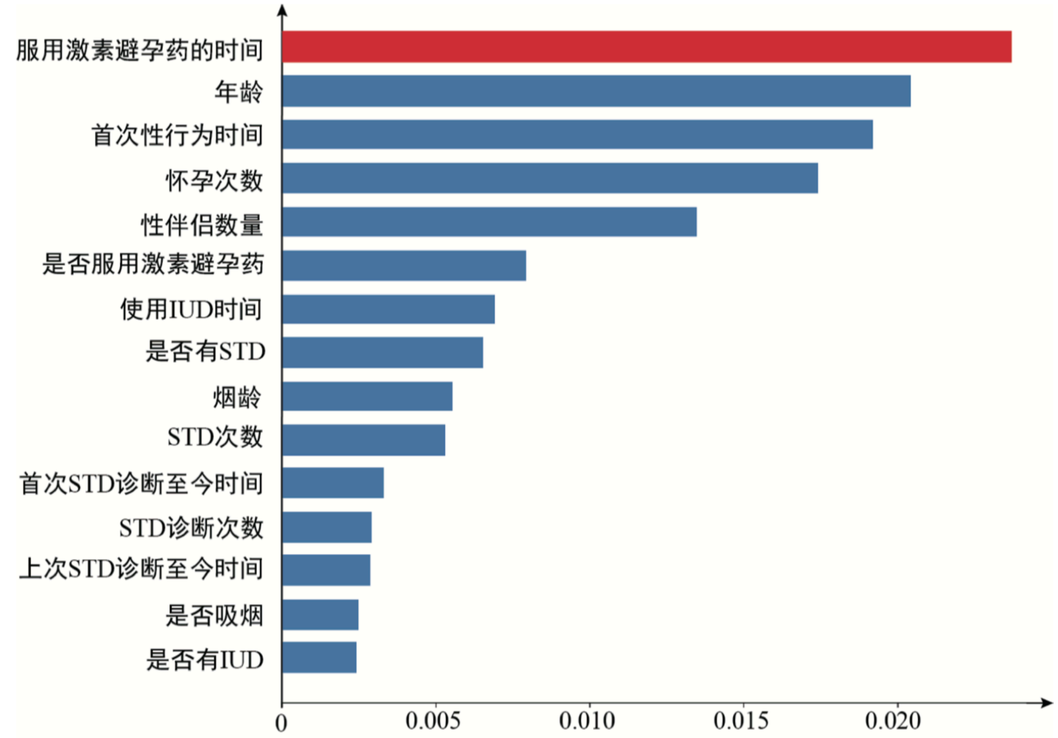
$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} |z'| (M - |z'|)}$$



图书内容

TIANCHI天池

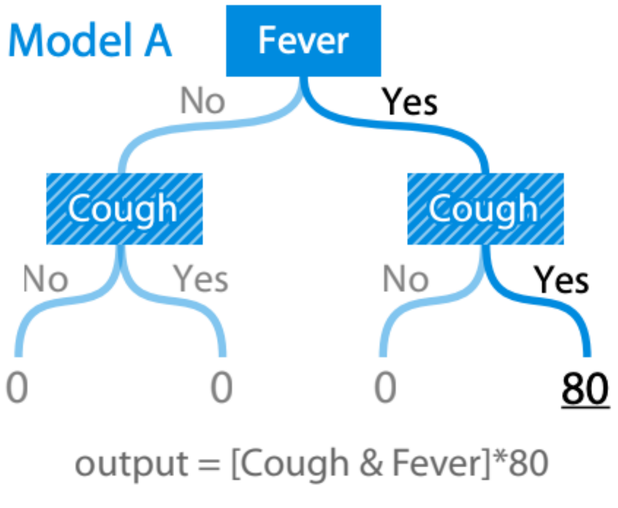
SHAP 特征重要性



Tree SHAP

将Shapley值合并到基于树的算法

- 在XGBoost和LightGBM中实现



图源于“Consistent Individualized Feature Attribution for Tree Ensembles”

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

基于样本的解释方法

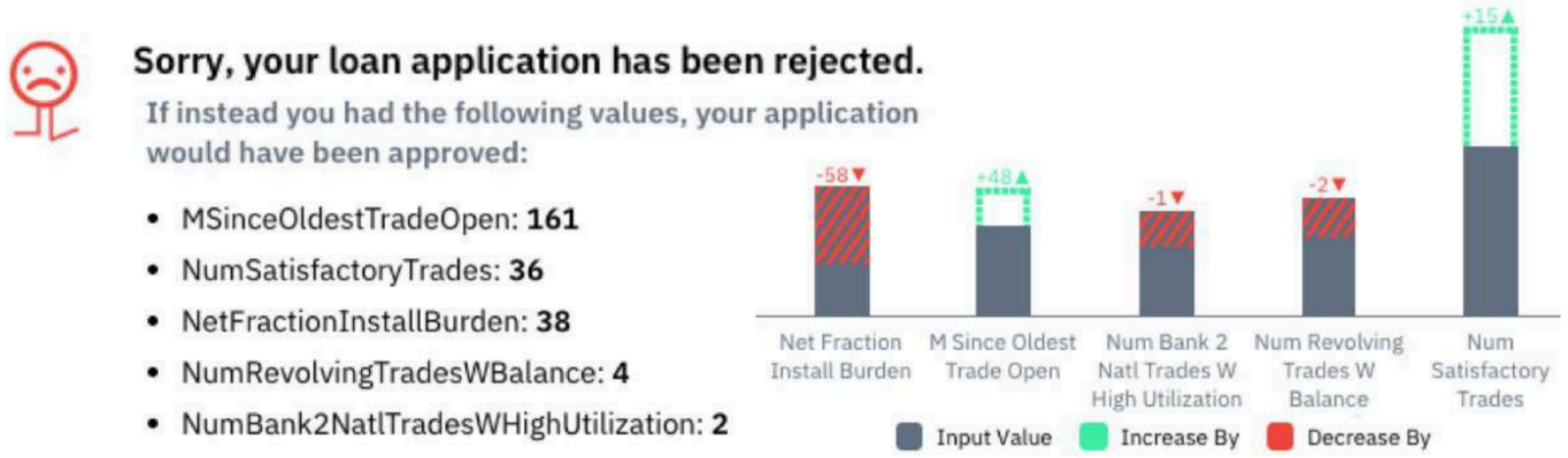
- 反事实解释
- 对抗样本
- 原型和批评
- 有影响力的实例

图书内容

反事实解释

反事实的示例

- 让模型回答“假如”的问题
- 生成反映目标结果的模型决策的样本



图源于“Interpretable Credit Application Predictions With Counterfactual Explanations”

反事实解释

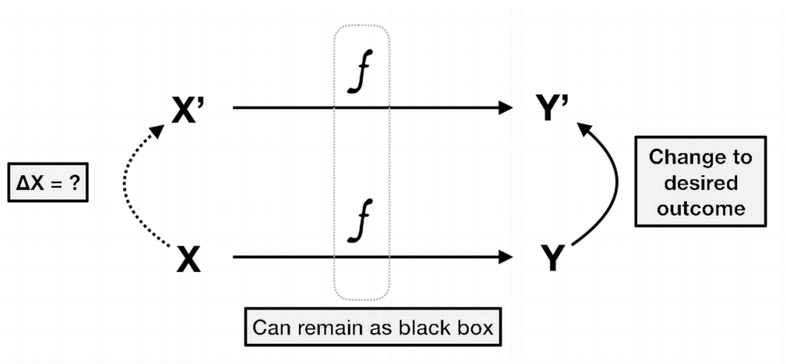
损失函数：

$$\mathcal{L}(x, x', y', \lambda) = \lambda(\hat{f}(x') - y')^2 + d(x, x')$$

反事实解释：

$$x' = arg\min_{x'} \lambda(\hat{f}(x') - y')^2 + d(x, x')$$

反事实样本　　期望输出　　距离函数



图书内容

TIANCHI天池

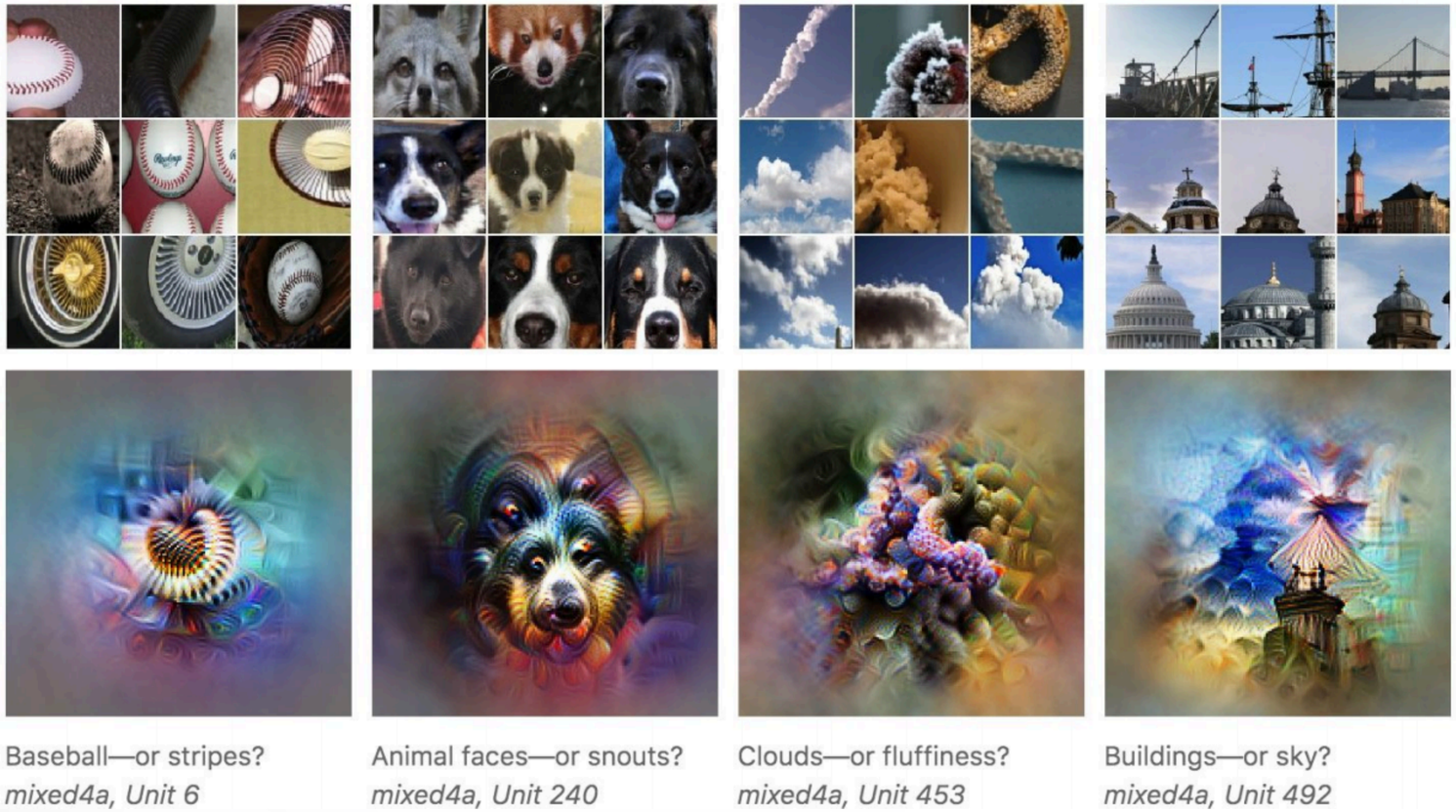
神经网络解释方法（*）

- 特征可视化
- 概念
- 特征归因
- 模型蒸馏

图书内容

TIANCHI天池

特征可视化



Baseball—or stripes?
mixed4a, Unit 6

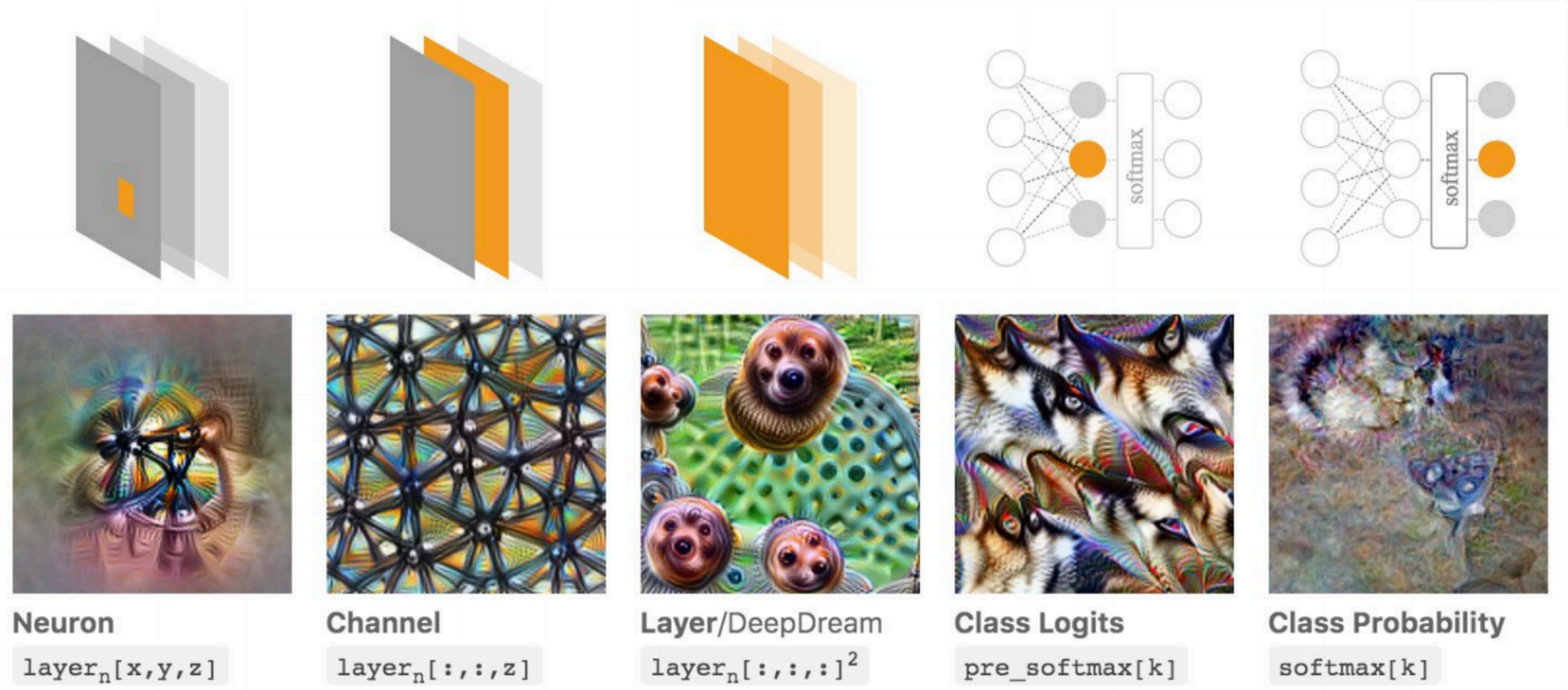
Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

Buildings—or sky?
mixed4a, Unit 492

图书内容

特征可视化

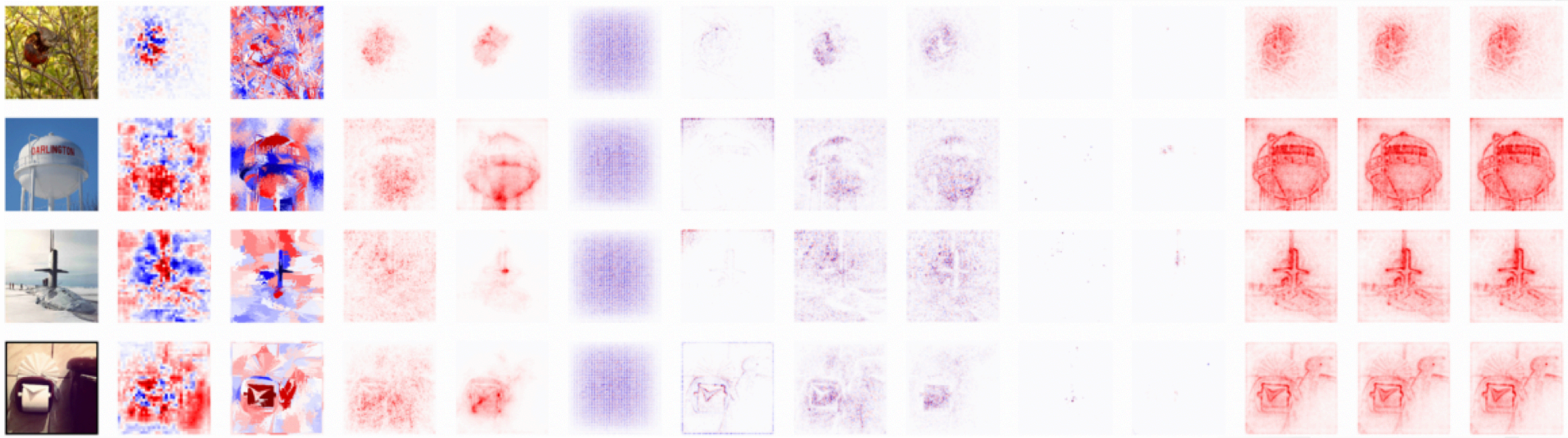


图源于“https://distill.pub/2017/feature-visualization/#enemy-of-feature-vis”

图书内容

TIANCHI天池

特征归因



图书特色

TIANCHI天池

可解释性相关知识

数据集介绍

- 自行车租赁（回归）
- YouTube 垃圾评论（文本分类）
- 宫颈癌的危险因素（分类）

方法介绍

全方面多角度，包括：

- 对方法直观的理解
- 对方法数学层面的理解
- 在数据集上的测试和解释过程
- 方法的优点和缺点分析
- 方法的实现工具和软件

R语言实现：https://github.com/christophM/iml

实战演示、互动交流

大家可以使用**电脑**访问下方地址进入天池读书会页面，点击今天读书会中的 **动手实践** 和我一起进行项目实践学习，天池为大家准备好了代码和运行环境，非常方便。

https://tianchi.aliyun.com/specials/promotion/activity/bookclub

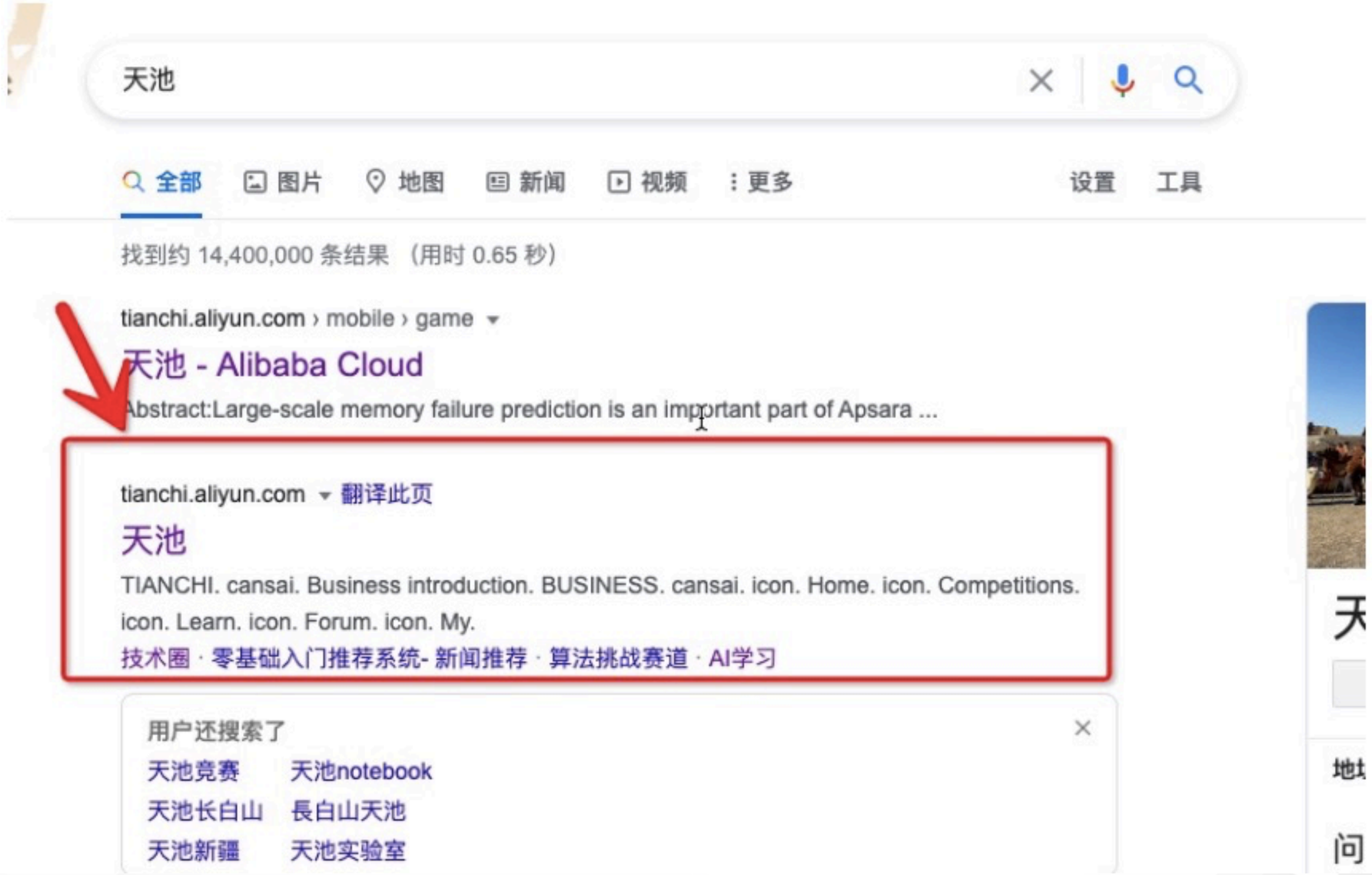
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI天池

Q&A

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

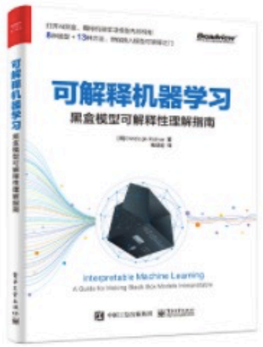
朱明超 本书译者、复旦研究生

直播主题 《可解释机器学习：黑盒模型可解释性理解指南》

直播时间 2021年4月21日 20:00

学习资料 机器学习训练营

实践项目 待定



提问 | 学习训练营 | 购买地址 | PPT下载 | 实战代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

谢谢观看

TIANCHI天池

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub