原文链接： <https://note.youdao.com/s/bIhItehb>

**机器学习的可解释性**

英文：<https://christophm.github.io/interpretable-ml-book/>

github：<https://github.com/christophM/interpretable-ml-book>

中文：<https://gitee.com/laogg/InterpretableMLBook> 第一版

中文：<https://blog.csdn.net/jiazhen/article/details/126824573> 第二版

上述是对应的书籍

本笔记按照新版章节划分，译文参考中文版本

**第五章 可解释的模型**

**5.3 GLM,GAM和其他模型**

**5.3.1 非高斯结果输出GLM**

☐ 假定伯努利分布，使用Logit函数作为连接删的GLM ，怎么得到逻辑回归公式

☐ 泊松分布的参数λ是单位时间(或单位面积)内随机事件的平均发生次数。 泊松分布适合于描述单位时间内随机事件发生的次数

☐ 二项分布是指在只有两个结果的n次独立的伯努利试验中，所期望的结果出现次数的概率

☐ 连续型数据或者数据离散性小，数据基本符合正态分布特点，或者对不符合的数据进行取对数或者样本重新排序达到正态分布特点，有具体的均数（期望）和标准差

**5.3.2 交互**

**5.3.3 非线性效应GAM**

特征的简单转换（例如对数，平方根，平方函数，指数函数）

特征分类（等距间隔？分位数？多少间隔？，需要跟多的数据，容易过拟合）

广义加性模型（GAM）

采用样条函数学习非线性函数（特征工程范畴）

**5.3.4 优点**

线性模型的所有这些扩展本⾝就是⼀个很⼤的探索空间。⽆论线性模型遇到的问题是什么，你都可能会找到可解决该问题的扩展

**5.3.5 缺点**

线性模型的⼤多数修改都会使模型的解释性变差。不是恒等函数的任何连接函数(在GAM 中) 都会使解释复杂化；交互也会使解释复杂化；⾮线性特征效应要么不太直观(如对数转换)，要么不能再由单个数字(如样条函数) 来概括

☐ 在许多情况下，随机森林或梯度提升树等基于树的集成⽅法的性能⽐最复杂的线性模型要好

**5.3.7 扩展**

|  |  |  |
| --- | --- | --- |
| **问题** | **示例** | **解决方案** |
| 数据违反了独⽴同分布(IID) 的假设 | 例如，对同⼀患者重复测量 | 混合模型(Mixed Models) 或⼴义估计⽅程(Generalized Estimating Equations) |
| 模型有异⽅差性误差(Heteroscedastic Errors) | 例如，在预测房屋价格时，昂贵房屋的模型误差通常较⾼，这违背了线性模型的同⽅差性 | 稳健回归(Robust Regression)。 |
| 有异常点会强烈影响模型 |  | 稳健回归 |
| 想预测事件发⽣前的时间 | 到达事件的时间数据通常带有经过审查的测量值，这意味着在某些情况下没有⾜够的时间来观察事件。例如，⼀家公司想预测其制冰机的故障，但只有两年的数据。⼀些机器两年后仍然完好⽆损，但可能会在以后会故障 | 参数⽣存模型(Parametric Survival Models)，Cox 回归(Cox Regression)，⽣存分析(Survival Analysis)。 |
| 要预测的结果是⼀个类别 |  | 如果结果有两个类别，则使⽤逻辑回归模型(Logistic Regression Model)，该模型对类别的概率进⾏建模。如果你有更多类别，可以考虑多分类回归(Multinomial Regression)。逻辑回归模型和多分类回归都是GLM。 |
| 预测有序的类别 | 例如，学校成绩 | ⽐例优势模型(Proportional Odds Model)。 |
| 结果是⼀个计数(就像⼀个家庭中孩⼦的数量) |  | 寻找泊松回归(Poisson Regression)。泊松模型也是⼀个GLM |
| 在上⼀个问题中，0 的计数值⾮常频繁 |  | 寻找零膨胀泊松回归(Zero-inflated Poisson Regression)，栅栏模型(Hurdle  Model) |
| 不确定模型中需要包含哪些特征才能得出正确的因果结论 | 例如，想知道药物对⾎压的影响。这种药物对某些⾎液有直接影响，⽽该⾎液会影响结果。  应该将⾎液纳⼊回归模型吗 | 寻找因果推理(Causal Inference)，中介分析(Mediation Analysis)。 |
| 缺少数据 |  | 多重填补(Multiple Imputation)。 |
| 把先前的知识整合到模型中 |  | 贝叶斯推断(Bayesian Inference)。 |

**5.4 决策树**

线性回归和逻辑回归在特征与结果之间的关系为⾮线性或特征交互的情况下会失败

树模型可⽤于分类和回归

☐ CART 基尼指数:基尼指数和信息熵都是用来描述系统混乱度的量 ,数学形式不一样，干的事是一样的.不纯度（impurity）--[GINI系数](https://so.csdn.net/so/search?q=GINI%E7%B3%BB%E6%95%B0&spm=1001.2101.3001.7020)：（不纯度就是混乱度）

☐ 我们希望决策树的分支结点 所包含的样本尽可能属于同一类别，即结点的“纯度 ”(purity)越来越高 ,不纯度越来越低

☐ 信息熵越小/基尼指数越小, 混乱度越小，纯度越高 <https://blog.csdn.net/qq_48864962/article/details/121003991>

**5.4.1 解释**

所有边通过“AND” 连接

特征重要性的计算方式：⼀个特征的总体重要性可以⽤以下⽅法计算：遍历使⽤该特征的所有分割，并测量它相对于⽗节点减少了多少⽅差或基尼指数

☐ 树分解方法：解释决策树的单个预测

**5.4.2 示例**

租赁回归任务

**5.4.3 优点**

树结构⾮常适合捕获数据中特征之间的交互。

**5.4.4 缺点**

树不能处理线性关系，输⼊特征和结果之间的任何线性关系都必须通过分割来近似，从⽽创建⼀个阶跃函数，这会缺乏平滑度

树也相当不稳定。训练数据集中的⼀些更改可以创建完全不同的树，这是因为每个分割都依赖于⽗分割

**5.5 决策规则**

IF-THEN 语句

决策规则的有⽤性通常概括为两个数字：⽀持度(Support)/覆盖率(Coverage) 和准确性(Accuracy)/置信度(Confidence)。规则条件适⽤的实例所占的百分⽐称为⽀持度

数据中挖掘学习规则

**5.5.1 从单个特征学习规则(OneR)**

通过选择适当的间隔来离散化连续特征。

对于每个特征：

在特征值和(分类) 结果之间创建交叉表。

对于特征的每个值，创建⼀个规则，预测具有此特定特征值(可以从交叉表中读取) 的实例的最常见类别

计算特征规则的总误差。

选择总误差最⼩的特征

OneR 倾向于具有许多可能级别的特征，因为这些特征可以更容易地覆盖⽬标

☐ 连结(Ties) 是另⼀个问题，即当两个特征导致相同的总误差时。OneR 通过使⽤具有误差最⼩的第⼀个特征或具有卡⽅检验的p 值最⼩的特征来解决连结。

**5.5.2 顺序覆盖**

Sequential covering

它重复学习单个规则以创建⼀个决策列表(或集合)，该决策列表(或集合) 按规则覆盖整个数据集

假设有一个算法可以创建一个覆盖部分数据的规则

空列表

学习一个规则

将规则添加到列表，删除刚学习的规则涵盖的所有数据点，在剩余的数据中继续学习其他规则

返回决策列表

通过束搜索beam search从决策树中学习单个规则

从根节点开始，递归的选择最纯的节点

终端节点的多数类别作为规则的预测，通过该节点的路径用作规则条件

☐ 学习单个规则是⼀个搜索问题，其中搜索空间是所有可能规则的空间。搜索的⽬标是根据某些条件找到最佳规则。有许多不同的搜索策略

☐ 爬⼭法(Hill-climbing)，束搜索，穷尽搜索(ExhaustiveSearch)，最佳优先搜索(Best-first Search)，有序搜索(Ordered Search)，随机搜索(StochasticSearch)，⾃上⽽下搜索(Top-down Search)，⾃下⽽上搜索(Bottom-up Search)

☐ RIPPER (Repeated Incremental Pruning to Produce ErrorReduction) 是顺序覆盖算法的⼀个变体；仅适用于分类

**5.5.3 贝叶斯规则列表**

Bayesian Rule List 难懂

从数据中预先挖掘(pre-mined) 频繁使⽤的模式，这些模式可⽤作决策规则的条件。

使⽤特征(在此步骤中不需要⽬标结果) 并从中提取频繁出现的模式

从预先挖掘的规则中选择决策列表。

☐ 狄利克雷分布

**5.5.4 优点**

IF-THEN 规则很容易解释。它们可能是可解释模型中最容易解释的

**5.5.5 缺点**

关于IF-THEN 规则的研究和⽂献主要集中在分类上，⼏乎完全忽略了回归

这些特征通常也必须是分类的

**5.5.6 软件和替代方法**

《Foundations of Rule Learning》(2012)[ürnkranz12]

Weka 规则学习器 <http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/package-summary.html>，该学习器实现了RIPPER，M5Rules，OneR，PART 等。IF-THEN 规则可⽤于线性模型中，

**5.6 RuleFit**

**5.6.1 解释与示例**

RuleFit 最终会估计⼀个线性模型，因此其解释与“常规” 线性模型的解释相同。

**5.6.2 理论**

规则生成;从决策树创建“规则”

稀疏线性模型:以原始特征和新规则作为输⼊⽤线性模型拟合

特征重要性：计算原始特征和新规则

☐ 不仅提升树，任何树集成算法都可以⽤作RuleFit ⽣成树

**5.6.3 优点**

RuleFit ⾃动将特征交互添加到线性模型

RuleFit 可以处理分类和回归任务（这意味着树集成中树的最⼤深度为3）

☐ RuleFit 提出了许多有⽤的诊断⼯具。这些⼯具与模型⽆关，因此你可以在本书的模型⽆关部分中找到它们：特征重要性，部分依赖图和特征交互。

**5.6.4 缺点**

有时RuleFit 创建许多规则，这些规则在Lasso 模型中的权重⾮零。可解释性随着模型中特征数量的增加⽽降低

**5.6.5 软件和替代方法**

<https://github.com/christophM/rulefit>

skope-rules <https://github.com/scikit-learn-contrib/skope-rules>

**5.7 其他可解释模型**

**5.7.1 朴素贝叶斯分类器**

**5.7.2 k-最近邻**

k-最近邻⽅法可⽤于回归和分类，并将数据点的最近邻⽤于预测

对于分类，分配实例到最近邻的类。对于回归，采⽤邻居结果的平均值

**第六章 模型无关的方法**

将解释与机器学习模型分离，这样可以使用各种机器学习模型，而用相同的解释方法

解释模型要灵活，可以与原模型一起使用；

解释方法的形式要灵活，可以是含数值公式形式，可以是含数值的图形形式；

解释的特征表示要灵活，不一定是具体的数值，也可以是单词

**局部解释和全局解释**

**第七章 基于样本的解释**

基于样本的解释⼤多与模型⽆关

基于样本的⽅法通过选择数据集的实例⽽不是通过创建特征概要(例如特征重要性或部分依赖性) 来解释模型

基于样本的解释的蓝图是：事物B 与事物A 类似，事物A 导致Y，因此我预测事物B 也将引起Y

**基于样本的解释可以用来：解释模型好坏，以及分析脏数据**

反事实解释告诉我们实例必须如何改变才能显着改变其预测。通过创建反事实实例，我们了解模型如何做出预测并可以解释各个预测。

对抗样本是⽤来欺骗机器学习模型的反事实。重点是翻转预测⽽不是解释它。

原型是从数据中选择具有代表性的实例，⽽批评是那些原型⽆法很好地表⽰的实例。

有影响力的实例是对预测模型的参数或预测本⾝影响最⼤的训练数据点。识别和分析有影响⼒的实例有助于发现数据问题、调试模型并更好地了解模型的⾏为。

k-最近邻模型是基于样本的(可解释的) 机器学习模型。

**第八章 全局模型无关法**

全局解释方法描述了平均表现，因此当建模人员想要理解数据中的一般机制或调试模型时，它们特别有用。

全局方法通常表示为基于数据分布的预期值

部分依赖图是一种特征效应方法。

累积局部效应图是另一种在特征依赖时起作用的特征效应方法。

特征交互（H-统计量）量化了预测的变化 在多大程度上取决于特征的交互作用。

函数分解是可解释性的核心思想，也是一种将复杂的预测函数分解为更小的部分的技术。

置换特征重要性衡量特征的重要性，作为置换特征时损失的增加（方差变化）。

全局代理模型用更简单的模型代替原始模型进行解释。

原型和批评是分布的代表性数据点，可用于增强可解释性

**8.1 部分依赖图**

Partial Dependence Plot，简称PDP 或PD

部分依赖图观察某个特征如何影响预测结果，特征重要性展示的是哪个变量对预测影响最大

其他特征不变，改变其中一个或两个特征，根据原模型预测**平均**结果得到依赖图

<https://zhuanlan.zhihu.com/p/100454991>

<https://www.cnblogs.com/zhouyc/p/15685920.html>

具体操作就是：先固定住其他特征，把你感兴趣的这个特征全部批量的改成某个值，不断从某个低值改到某个高值，看所有样本预测结果会做什么变化。

**8.1.2 优点**

部分依赖图的计算很直观，容易实现

当查看的特征与其他特征不相关时，PDP可以完美的表示该特征如何平均影响预测，解释很清楚

部分依赖图的计算具有因果关系

**8.1.3 缺点**

特性数量最大我2

独立性假设是个问题，会出现不正常的组合。当特征相关时的解决办法：适⽤于条件分布⽽⾮边际分布的累积局部效应图或简称ALE 图

异质效应可能被隐藏

**8.1.4 软件**

PDPBox

PDP 替代⽅法是ALE 图和ICE 曲线

**8.2 累积局部效应图**

Accumulated Local Effects Plot， ALE

累计局部效应和PDP一样，同样描述了特征如何平均影响机器学习模型的预测

<https://blog.csdn.net/shuaibuzhi1mian/article/details/119971907>

**8.3.1 动机和直觉**

计算有强相关的特征部分依赖图，会生成实际中不可能出现的人工数据，对这样的数据进行平均预测会影响估计的特征效应

ALE通过基于特征的条件分布计算预测差异，而不是平均值；差异的使用会阻止其他特征的影响

PDP:基于特征的边际分布计算平均值，每个数据实例都含特征的某个具体值

M：基于特征的条件分布计算平均值，在特征的某个具体值附近的数据实例 指定该具体值，计算小范围内的平均值

ALE：基于特征的条件分布，计算围绕特征具体值的小范围内的预测变化

**8.3.2 理论**

各公式对比

**8.3.3 估计**

☐ 估计非中心化效应

☐ 效应中心化

间隔定义：使用特征分布的分位数，确保每个间隔内有相同数量的实例，但可能间隔长度不同

☐ 用于两个特征交互作用的ALE图

**8.3.5 优点**

ALE 图是无偏的，这意味着在特征相关时它们仍然有效

ALE 绘图的计算速度比PDP 更快

ALE 图的解释很清楚

☐ ALE 图以0 为中心

ALE 绘图仅显示交互作用：如果两个特征不交互，则图不显⽰任何内容

在大多数情况下，我宁愿使用ALE 图而不是PDP 图，因为特征通常在某种程度上相关

**8.3.6 缺点**

ALE 图可能会变得有些不稳定，减少间隔数可以估计的更稳定，但不好确定间隔数量

☐ ALE图不附带ICE曲线

二阶ALE 估计在整个特征空间中具有不同的稳定性，这是不以任何方式可视化

二阶效应图解释起来有点烦人

即使ALE 图在相关特征的情况下没有偏差，但当特征强相关时，解释仍然困难

**8.3 特征交互**

Feature Interaction

**8.3.1 特征交互**

估计交互强度的一种方法时衡量 预测的变化 在多大程度上取决于特征的交互作用，这种衡量方法称为H统计量

**8.3.2 理论：弗里德曼的H统计量**

我们将处理两种情况：⾸先，采⽤双向交互度量，它告诉我们模型中的两个特征是否交互以及在何种程度上交互；其次，是⼀个总体交互度量，它告诉我们某个特征在模型中是否与所有其他特征发⽣交互以及在何种程度上的交互。从理论上讲，可以测量任意数量的特征之间的任意交互，但这两个是最感兴趣的情况。

H 统计量的评估成本很⾼，因为它会在所有数据点上进⾏迭代

**8.3.4 优点**

交互作⽤H 统计量通过部分依赖分解具有理论基础

H 统计量具有有意义的解释：交互作⽤定义为由交互作⽤解释的⽅差份额。

由于统计信息是无量纲的，并且总是在0 和1 之间，因此它在各个特征甚⾄模型之间都具有可⽐性。

统计信息会检测各种类型的交互，⽆论它们的特殊形式如何。

使⽤H 统计量，还可以分析任意更高阶的交互作用，例如3 个或更多特征之间的交互作⽤强度

**8.3.5 缺点**

计算量大

涉及估计边际分布，如果不用所有点可能估计值有差异，则进行点采样时估计结果有偏差，结果可能不稳定；建议重复几次H统计看是否稳定

H统计量只能说明交互的强度，但如何交互不清楚

**8.3.7 替代方法**

Greenwell 等⼈(2018) 基于部分依赖性的特征交互测量了两个特征之间的交互

**8.4 函数分解**

Functional Decomposition

**8.5 置换特征重要性**

Permutation Feature Importance

<https://cloud.tencent.com/developer/article/1680674>

**8.5.1 理论**

特征交互，或者特征下的数值打乱

**8.5.2 应该计算训练集还是测试集的重要性**

模型训练过拟合，在训练集上效果好，特征重要，但在测试集上效果不好，特征就不重要，通过误差来判断特征重不重要 不好确定（因为本身预测就不准确）

**8.5.3 示例和解释**

**8.5.4 优点**

很好的解释：特征重要性是当特征信息被破坏时模型误差的增加。

置换特征重要性同时考虑了主要特征效应和交互效应

置换特征重要性不需要重新训练模型

**8.5.5 缺点**

训练数据和测试数据计算特征重要性结果可能不同

置换特征的重要性与模型的误差有关，有时候不需要考虑特征多重要，而是考虑哪个特征对输出的影响

无监督数据无法计算置换特征重要性

置换特征，如果特征相关，会出现不切实际的数据实例

**8.6 全局代理模型**

Global Surrogate

全局代理模型是⼀种可解释的模型，经过训练可近似于黑盒模型的预测。

**8.7 原型与批评-基于样本的解释**

Prototypes and Criticisms

**第九章 局部模型无关法**

局部解释方法解释个别预测。在本章中，你将了解以下局部解释方法：

单个条件期望曲线是部分依赖图的构建块，描述了改变特征如何改变预测。

局部代理模型（LIME）通过将复杂模型替换为局部可解释的代理模型来解释预测。

范围规则（锚点）是描述哪些特征值锚定预测的规则，从某种意义上说，它们将预测锁定到位。

反事实解释通过检查需要更改哪些特征以实现所需的预测来解释预测。

Shapley 值是一种归因方法，可以将预测公平地分配给各个特征。

SHAP是 Shapley 值的另一种计算方法，但也提出了基于数据中 Shapley 值组合的全局解释方法。

LIME 和 Shapley 值是归因方法，因此单个实​​例的预测被描述为特征效应的总和。其他方法，例如反事实解释，是基于示例的。

**9.1 个体条件期望**

Individual Conditional Expectation，简称ICE

个体条件期望 图为每个实例显⽰⼀条线，该线显⽰了特征更改时实例的预测如何改变。

单个数据实例的PDP就是ICE，PDP是ICE图的线的平均值

**9.1.1 示例**

中心化个体条件期望图（Centered ICE Plot / c-ICE),判断个体间的不同，选择一个基准

导数个体条件期望图（Derivative ICE Plot / d-ICE） 发现异质区域

**9.1.2 优点**

和PDP相比更直观

可以发现异质区域

**9.1.3 缺点**

只能有意义的显示一个特征

如果特征相关，和PDP同样的问题

全部分析单个实例，图太拥挤

**9.2 局部代理 LIME**

参考[机器学习&深度学习可解释性报告](note://WEB5190ef8cee6e6def416bfdc630f7c0bc)

LIME 是少数适用于表格数据，文本和图像的⽅法之⼀。

定义邻域是个问题，必须尝试不同的核设置，并亲自查看解释的合理性

改善采样方法

**9.3 反事实解释-基于样本的解释**

Counterfactual Explanations

反事实解释描述了一种因果关系：如果没有发生A ，就不会发生B

不考虑输入导致的预测结果是对是错，只关心输入是预测的原因

预测的反事实解释：我们想要得到什么样的预测结果时，特征值的最小变化

定义反事实解释实例

定义的实例必须能输出想要的预测结果，

该实例的特征变化尽量小，可以和原实例用距离度量

定义的实例要合理，不要出现不正常的特征值，可以根据数据的联合分布产生反事实

**9.3.1 生成反事实解释**

一种方法:随机改变感兴趣的实例的特征值，直到输出预期的结果停止

损失函数：反事实的预测结果和期望结果的距离，反事实和要解释的实例间的距离（预测距离和特征值距离）

绝对中位差实际求法是用原数据减去中位数后得到的新数据的绝对值的中位数

**9.3.2 示例**

想得到什么结果，至少需要改哪项特征的值

**9.3.3 优点**

解释清楚

模型无关，只需要模型的预测结果，不需要关心那种类型的模型

反事实解释本质上是一种损失函数

**9.3.4 缺点**

罗生门效应：每个实例有多个反事实解释

最小化预测的公差，可能找不到反事实实例 （可能是数据问题）

分类特征多，每个分类特征级别又多，组合方式多（Alibi包解决了）

**9.3.5 软件和替代方法**

Martens提出了文档分类的解释方法

通过搜索而不是通过优化损失函数的方法找反事实实例 Growing Spheres

**9.4 范围规则**

Scoped Rules (Anchors)

**9.5 Shapley 值**

博弈论方法

机器学习中的 Shapley 值怎么理解？<https://zhuanlan.zhihu.com/p/91834300> 理解更细

<https://www.zhihu.com/question/23180647/answer/49724759> 这个更好理解

shapley值用于计算每个单独的特征对模型输出的贡献

**9.5.3 详细的Shapley值**

线性模型中每个特征对预测的贡献：特征效应减去平均效应

shapley值是满足多条性质的归因方法

☐ 估计shapley值的方法：蒙特卡罗抽样的近似值

**9.5.4 优点**

理论扎实，可以对比解释

**9.5.5 缺点**

计算时间长，特征越多计算时间越久

使⽤Shapley 值⽅法创建的解释始终使用所有特征，Shapley 值为每个特征返回⼀个简单值

Shapley 值⽅法在特征相关时会遇到不现实的数据实例

**9.6 SHAP**

SHapley Additive ExPlanations

SHAP 基于博弈论上的最佳Shapley 值

kernalshap:基于核的代理方法，根据局部代理模型对Shapley值进行估计

treeshap:基于树的模型的有效估计方法

基于shapley值聚合的全局解释方法

先参考这篇：<https://zhuanlan.zhihu.com/p/85791430> SHAP知识点全汇总

模型可解释性(2)-SHAP计算过程 <https://zhuanlan.zhihu.com/p/186204351>

**9.6.1 SHAP定义**

待定

**第十章 神经网络可解释性**

**10.1 Learned Features**

**10.2 Pixel Attribution (Saliency Maps)**

**10.3 Detecting Concepts**

**10.4 对抗样本-基于样本的解释**

Adversarial Examples

对抗样本是反事实实例，目的是欺骗模型而不是解释模型

**10.4.1 方法与示例**

Intriguing Properties of Neural Networks 基于梯度的优化⽅法来寻找深度神经⽹络的对抗样本

损失函数和反事实解释一样，对抗样本与期望类别距离+对抗样本与原始图像距离

快速梯度符号方法(Fast Gradient Sign Method)

通过向每个像素添加或减去⼀个⼩的扰动ϵ 来操纵原始图像x。加或减ϵ 取决于像素的梯度符号是正还是负

单像素攻击(1-pixel attacks)

对抗补丁(Adversarial Patch)

消除了对抗图像必须⾮常接近原始图像的限制

：鲁棒的对抗样本(RobustAdversarial Examples)

黑盒攻击(Black Box Attack)

**10.4.2 网络安全视角**

**10.5 有影响力的实例-基于样本的解释**

**心得**

基于NLP的深度学习模型，如何解释（对于非结构化数据怎么处理）

对结果评估分两步: 先对数据进行解释，再对解释进行合理性评估

深度学习如何利用结构化数据进行模型学习-《深度学习处理结构化数据实战》

了解目前项目中各环节使用到的算法，规则；了解各环节输入输出；了解各环节识别结果的指标

了解现有库说明，文本类的局部模型无法解释说明

读一些论文

**阅读次数**

☐ 第一次：2023/01/30 10:34 - 2023/02/02 08:46

引用：

[explaining-decisions-made-with-ai](note://WEB8b99abd5719311260a88cc686c677901)

[可解释性库Captum](note://WEB38c13ede062b35d10fb8b60f4dc02f57)

[可信TrustAI](note://WEB7880ef444d0d8cae48dec97a534f31ba)

[机器学习&深度学习可解释性报告](note://WEB5190ef8cee6e6def416bfdc630f7c0bc)