在机器学习与人工智能领域，关于模型可解释性的研究逐渐成为该领域的关注重点。因为模型效果已无法很好地满足实际应用的需要，模型产生效果的原因和模型的可解释性也是非常重要的。对于机器学习的用户而言，模型的可解释性是一种较为主观的性质，无法通过严谨的数学表达方法形式化定义可解释性。通常认为，机器学习的可解释性刻画了人类对模型决策或预测结果的理解程度[1]，用户可以更容易地理解解释性较高的模型做出的决策和预测。对于机器学习任务而言，最令人感兴趣的两类问题是“为什么会得到该结果”和“为什么结果应该是这样”。在理想状态下，如果能够通过溯因推理的方式恢复模型计算出输出结果的过程，就可以实现较强的模型解释性。对模型可解释性的探索有助于模型和特征的优化，帮助人们更好地理解模型本身，提升模型的服务质量[2]。

虽然在许多任务上，机器学习算法都已经超过了人类，但缺乏可解释性仍然是很多机器学习模型无法被广泛使用的一个重要原因。在进行模型选择时，通常会面临准确率和模型复杂度之间的权衡。一个简单的线性回归(Linear Regression，LR)非常好解释，因为它只考虑了自变量与因变量之间的线性相关关系，但正因为如此，它无法处理更复杂的关系，因此模型在测试集上的预测精度有可能比较低。而一个模型越复杂就越难以解释。比如，深度神经网络则处于另一个极端，因为它们能够在多个层次进行抽象推断，因而能够处理因变量与自变量之间非常复杂的关系，并且达到非常高的精度。但是这种复杂性也使模型成为黑箱，我们无法获知所有产生模型预测结果的特征之间的关系，所以只能用准确率、错误率这样的评价指标来评估模型的可信性。另外一个例子是随机森林：一个随机森林模型由数百个决策树组成，并通过“投票”的方式得到最后的预测结果。为了理解随机森林模型是如何做出决策的，就需要使用者查看数百棵树中每棵树的树结构和预测结果，而这基本是上很难做到的[3]。

在工业界中，机器学习的主要焦点是更偏“应用”地解决复杂的现实世界中至关重要的问题，因此可解释性就显得尤为重要。在一些领域，特别是在医疗、法律、金融等涉及高风险决策的领域，由于复杂模型通常难以解释，数据科学家通常不得不使用更传统的机器学习模型（线性或基于树的），因为需要从更为详细和具体的角度理解模型得出结论的原因。由此可见，模型可解释性对于所采取的每个决策是非常重要的，也在很大程度上决定了模型能否被广泛地应用在实际

场景中。此外，对可解释性的需求来自问题形式化的不完整性[4]，即对于某些问题或任务，仅仅获得预测结果是不够的，机器学习流程中还应该包括对模型的解释。为模型赋予可解释性也有利于确保其公平性、鲁棒性、隐私保护性能，提升用户对模型的信任程度。可解释性需求的来源可以总结为以下三方面：（1）促进模型的完善：可解释性提供对模型输入、特征、预测的解释，对我们理解为什么一个机器学习模型会做出这样的决定、什么特征在决定中起最重要作用，有助于判断模型是否符合人类的认知，进而对模型进行诊断和完善。（2）提升模型可信性与透明度：理解机器学习模型在提高模型可信度和提供预测结果透明度上是非常必要的。在医学领域，机器学习模型的预测结果可能会直接决定病人的生死。例如，现有模型在区分恶性肿瘤和不同类型的良性肿瘤方面是非常准确的，但是我们依然需要专家对诊断结果进行解释。如果一个模型能回答“为什么它将某个患者的肿瘤归类为良性或恶性”这样的问题，将会大大提升医生对其的信任程度，并节省大量时间。如果一个模型做出了错误的判断，那么在应用这个模型之前，就能通过对解释结果及时发现并阻止不良影响的发生。（3）提升模型公平性：由于机器学习高度依赖于训练数据，而训练数据往往并不是无偏的，会产生对于人种、性别、职业等因素的偏见。为了保证模型的公平，用户会要求模型具有检测偏见的功能，能够通过对自身决策的解释说明其公平性。因此，具有强可解释性的模型会具有更高的社会认可度，更容易被公众所接纳。对数据科学家和决策制定者来说，理解模型是如何做出决策的，并事先预防偏差的出现，也是一项应尽的义务。

通常，可以根据不同标准对机器学习可解释性的方法进行分类[5]：

（1）本质的(Intrinsic)和事后的(Post-hoc)。该标准通过限制机器学习模型的复杂性(本质的，亦可称内在的)或在训练后分析模型的方法(事后的)来区分是否实现了可解释性。本质的可解释性是指由于结构简单而被认为是可解释的机器学习模型，如较浅的决策树或稀疏线性模型；事后解释性是指模型训练后运用解释方法，例如，置换特征重要性（Permutation Feature Importance）[6, 7]就是一种事后解释方法。事后也可以应用于本质上可解释的模型，例如，计算决策树的置换特征重要性。

（2）特定于模型的(Model-specific)和模型无关的(Model-agnostic)。特定

于模型的解释方法仅限于特定的模型类，例如，对线性模型中回归权重的解释就是特定于模型的解释。此外，仅应用于解释神经网络的方法也是特定于模型的。相对应的，模型无关的方法可以用于任何机器学习模型，并且在模型训练完成后应用（事后的）。这类方法一般无法访问模型的内部信息，如权重或结构信息，通常通过对模型的输入和输出进行分析来提供可解释性。

（3）局部的(Local)和全局的(Global)。这种分类方式的标准是可解释方法解释单个预测还是整个模型行为，或者介于两者之间。

（4）根据可解释方法的输出来大致区分：(a)特征概要统计量(Feature Summary Statistic)。该类解释方法为每个特征提供概要统计量。比如，为每个特征返回一个值，表示特征重要性，或者更复杂的输出，例如成对特征交互强度。（b）特征概要可视化(Feature Summary Visualization)。由于有些特征概要只有在可视化的情况下才有意义，这类方法输出特征概要的可视化结果。例如，部分依赖图（Partial Dependence Plot，PDP）是一种显示特征和平均预测结果之间关系的曲线[8]。（c）模型内部(Model Internals)。对于本质上可解释模型的解释属于该类方法，如线性模型中的权重或决策树的树结构。对于卷积神经网络，输出模型内部结构的一种方法是对学习到的特征检测器进行可视化。因此，输出为模型内部的可解释方法是特定于模型的。（d）数据点(Data Point)。这类方法的输出是数据集中的数据点（实例）或者新生成的数据点。一种典型的方法是反事实解释(Counterfactual Explanations)，该方法通过改变实例的特征值以改变预测结果（例如，预测类别的改变），然后分析预测结果是如何变化的，从而找到该实例的反事实解释。另一种方法是识别数据中的原型（Prototypes）和批评（Criticisms）。原型和批评可以用于描述数据，创建可解释的模型或使黑盒模型可解释。（e）本质上可解释模型。这类方法通常采用可解释模型全局地或局部地近似黑盒模型。由于可解释模型本身可以通过查看模型内部参数或特征概要统计量来解释，因此可以通过近似的可解释模型对黑盒模型进行解释。

机器学习模型的解释方法的分类标准众多，以下从可解释模型、模型无关和基于样本的解释三个角度，介绍一些典型的模型解释方法。

采用可解释模型是实现可解释性的最简单的方式。一些传统的机器学习方法，如线性回归、逻辑回归和决策树都是可解释模型。在线性回归模型中，目标预测

等于输入特征的线性加权和。对于第i个实例，可以写成：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

其中，参数β_j是特征权重或者系数，第一项β_0称为截距，ϵ表示误差。通常使用最小二乘法计算真实结果和预测结果之间的差距并优化权重：

$$\hat{\beta} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} (y^{(i)} - (\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)}))^2$$

特征的线性组合使得线性回归模型是一种可解释模型，而特征的解释取决于特征的类型。对于数值特征，当所有其他特征保持不变时，特征x_k增加一个单位，预测结果y增加β_k；对于二分类特征，当所有其他特征保存不变时，将特征x_k从参照类别改为其他类别时，预测结果y增加β_k。特征重要性可以用其t-统计量（t-statistic）的绝对值来衡量。t-统计量是以标准差为尺度的估计：

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

可以看出，特征的重要性与权重大小成正比，与估计权重的方差成反比。对于逻辑回归模型，其回归目标是二分类中正负例的对数几率：

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

对于数值特征，当所有其他特征保持不变时，特征x_k增加一个单位，估计的几率将乘以$\exp(\beta_x)$；对于二分类特征，当所有其他特征保存不变时，将特征x_k从参照类别改为其他类别，估计的几率将乘以$\exp(\beta_x)$。此外，其他可解释模型包括朴素贝叶斯、决策树、K 近邻、广义线性模型、广义加性模型等。

对于可解释模型，可解释性是其固有的属性。然而，我们希望能够找到一些方法，对任何的黑盒子机器学习模型提供解释，即模型无关的方法。模型无关的可解释方法有以下优点[9]：（1）模型的灵活性：可以用于任何机器学习模型；（2）解释的灵活性：不限于特定形式的解释；（3）表示方式的灵活性：解释方法能够使用与所解释模型不同的特征表示方式。一种典型的模型无关解释方法是部分依赖图，它显示一个或两个特征对机器学习模型的预测结果的边际效应[8]。例如，回归任务中的部分依赖函数定义为：

$$\hat{f}_{x_S}(x_S) = E_{x_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C)\, d\mathbb{P}(x_C)$$

其中，$\hat{f}$表示机器学习模型，x_S是部分依赖函数应该被绘制的特征集合，是我们想要了解其对预测的影响的特征，x_C是其他特征集合。通过在集合x_C中的特征分布上边缘化机器学习模型的输出，展示集合x_S中的特征与预测结果之间的关系。采用蒙特卡洛估计法来计算$\hat{f}_{x_S}$：

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right)$$

其中，n是数据集中的实例数。图 1 展示了一个随机森林模型在自行车数量预测任务上，天气特征对预测结果的影响[5]。可以看到，PDP 较好地展示了不同天气特征（温度、湿度和风速）对自行车数量预测结果的影响。类似的方法还有个体条件期望(Individual Conditional Expectation，ICE)图[10]、累积局部效应(Accumulated Local Effects，ALE)图[11]等。

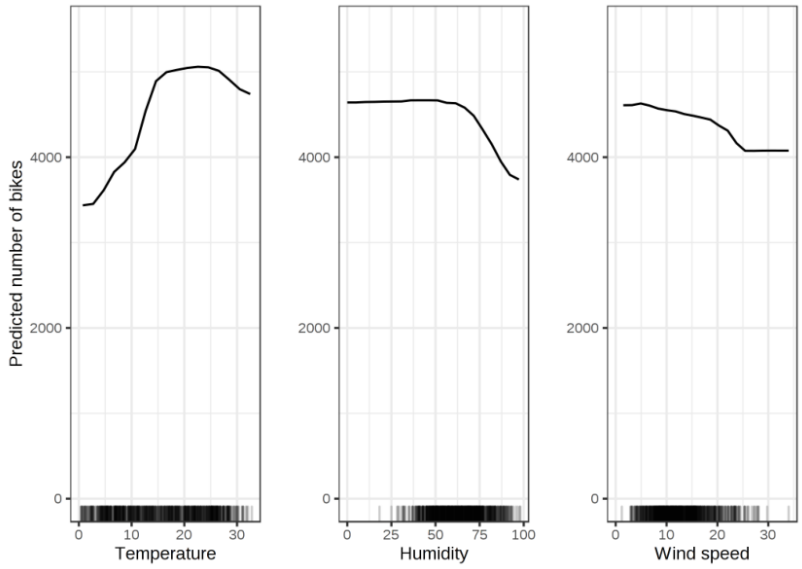

图 1 自行车数量预测与温度、湿度和风速的 PDP。

代理模型（Surrogate Models）是另一类模型无关的解释方法。代理模型通常是简单模型，用于解释复杂的、不可解释的模型。常用的代理模型有线性模型和决策树模型，主要是由于这些模型易于解释。可解释的代理模型的目的是，尽可能准确地近似不可解释模型的预测，同时进行解释。以下步骤说明了如何为复杂的黑盒模型构建代理模型：（1）选择一个数据集X（可以是和训练黑盒模型相同的数据集，也可以是来自同一分布的新数据集）；（2）获取黑盒模型在数据集X上的预测结果Y；（3）选择一种可解释模型（线性模型、决策树等）；（4）在数

据集 X 及其预测 Y 上对可解释模型进行训练，得到代理模型；（5）衡量代理模型对黑盒模型的近似效果；（6）通过代理模型对黑盒模型进行解释。（5）中，衡量代理模型对黑盒模型的近似效果的一种方法是计算R-平方度量：

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

其中$\hat{y}_*^{(i)}$和$\hat{y}^{(i)}$分别是代理模型和黑盒模型的第i个实例的预测，$\bar{\hat{y}}$是黑盒模型预测的平均值。如果R^2接近 1，说明可解释模型能很好地近似盒模型的预测。例如，用 CART 决策树作为 SVM 的代理模型，计算得到$R^2 = 0.77$，这表明 CART 决策树能很好地近似 SVM 模型。如果足够接近，则可用代理模型替换黑盒模型[5]。代理模型可分为全局代理模型(Gloal Surrogate Models)和局部代理模型(Local Surrogate Models)。全局代理模型在整个数据集上构建代理模型，而局部代理模型只对部分区域的预测构建代理解释。局部可解释的模型无关解释(Local interpretable model-agnostic explanations，LIME)[12]使用局部代理模型来对单个样本进行解释。LIME 对于需要解释的黑盒模型，取关注的实例样本，在其附近进行扰动生成新的样本点，并得到黑盒模型的预测值，使用新的数据集训练可解释的模型（如线性回归、决策树），得到对黑盒模型良好的局部近似。LIME 中，具有可解释性约束的局部代理模型表示为：

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

其中，x代表实例，g是代理模型（如线性回归模型），G是简单代理模型的集合（如所有可能的线性模型），f是原始黑盒模型，损失L（如均方误差）衡量两个模型预测的接近程度，接近度π_x定义了考虑解释时实例x附件邻域的大小，$\Omega(g)$表示模型的复杂度。可以看到，LIME 希望达到可解释性和局部可信度之间的平衡。LIME 训练局部代理模型的方法是：（1）采样局部感兴趣区域的样本点，这些样本点可以从数据集中直接检索，也可以人工生成；（2）通过邻近的感兴趣区域对新样本进行加权，在新数据集上训练加权的、可解释的代理模型；（3）通过解释局部模型来解释黑盒模型的预测。

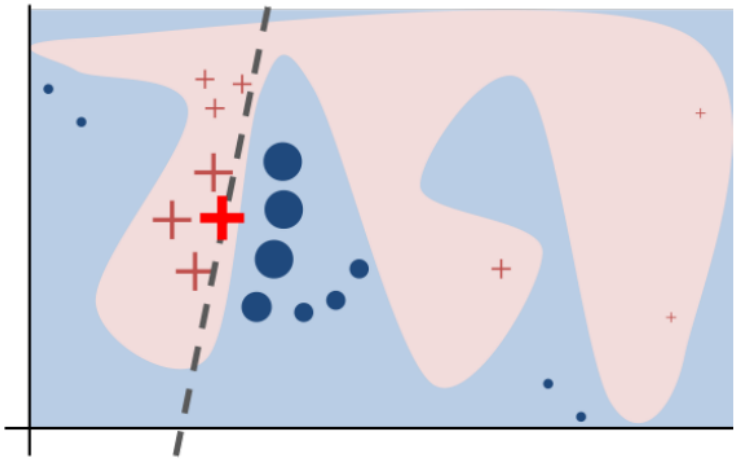

图 2 LIME 示例。蓝/粉背景表示黑盒模型f的复杂决策面（无法用线性模型很好地逼近），粗体红色十字表示需要被解释的实例，十字/圆表示不同类别实例，大小代表权重，虚线是学习到的局部可解释模型的决策面。

另外一种模型无关的可解释方案来自博弈论：Shapley 值[13]。假定数据的每一个特征是游戏中的一个玩家，每个玩家对于预测的结果都有一定的贡献。Shapley 值是一种针对任何机器学习模型的单个预测计算特征贡献的解决方案。对于每一个实例的预测结果，Shapley 值给出每一个特征对于这个预测结果的贡献度。每个特征值的 Shapley 值是其对预测的贡献在所有可能的特征值组合上的加权和：

$$\phi_j(val) = \sum_{S\subseteq\{x_1,\dots,x_p\}\setminus\{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(val\left(S\cup\{x_j\}\right)-val(S)\right)$$

其中S是模型中使用的特征的子集，x是要解释的实例，p是特征数量，$val(S)$是对集合S中的特征值的预测，它是在集合S中未包含的特征上进行边缘化：

$$val_x(S) = \int \hat{f}(x_1,\dots,x_p)\,d\mathbb{P}_{x\notin S} - E_X(\hat{f}(X))$$

精确的 Shapley 值必须使用第j个特征和不使用第j个特征的所有可能的集合来估计，当特征数增多时，可能的集合数量呈指数增长，因此可采用 Strumbelj 等人[14]提出的蒙特卡洛抽样方法来近似估计。

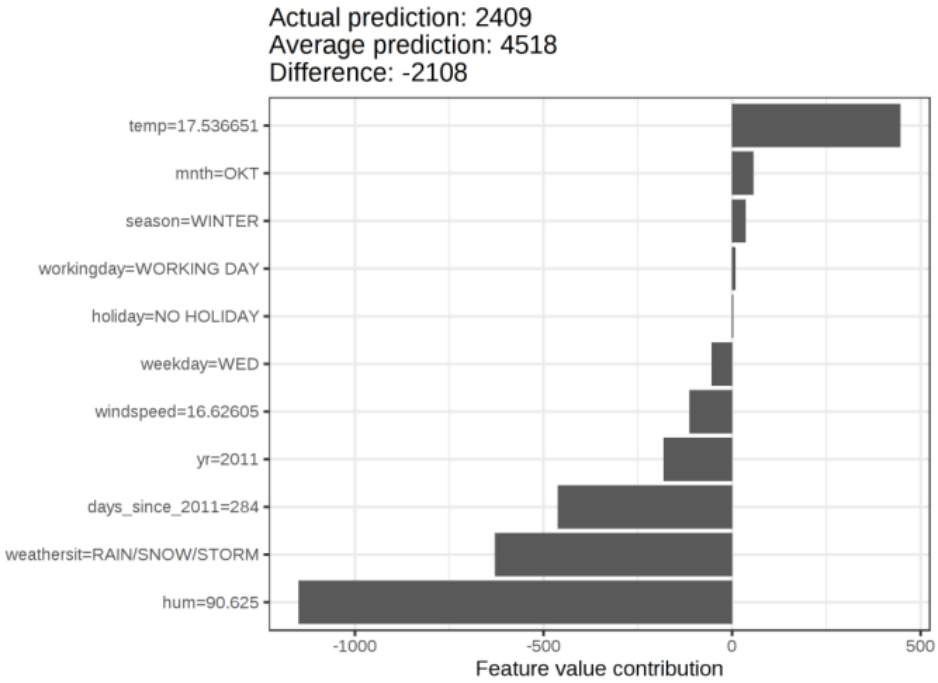

图 3 Shapley 值示例。自行车数量预测任务，某天的预测量为 2409，比平均预测量少 2108。不同特征对自行车数量的影响：温度起到了积极作用，而天气和湿度等产生了负面影响[5]。

SHAP(SHapley Additive ExPlanations)将 Shapley 值解释为一种加性特征归因方法(additive feature attribution method)，并将 LIME 和 SHAP 联系起来[15]。SHAP 的作者还提出了 KernelSHAP 和 TreeSHAP，前者是一种基于核的代理方法，可根据局部代理模型对 Shapley 值进行估算，后者是一种基于树的模型的有效估方法。此外，SHAP 还带有许多基于 Shapley 值聚合的全局解释方法。

基于样本的解释方法(Example-based Explanations)选择数据集的特定实例来解释机器学习模型的行为或解释底层数据分布。基于样本的解释大多与模型无关，因为它们使任何机器学习模型都更具可解释性。与模型无关的方法的不同之处在于，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。这种方法对于图像的解释非常有效，因为它允许我们直接查看可解释的数据实例。反事实解释(Counterfactual Explanations)就是一种基于样本的解释方法，可用于解释实例的预测结果。如图 4 所示，“事件”是实例的预测结果，“原因”是该实例的特定特征值，将其输入到模型并会“引起”某个预测[5]。对一个实例，我们改变其特征值，然后

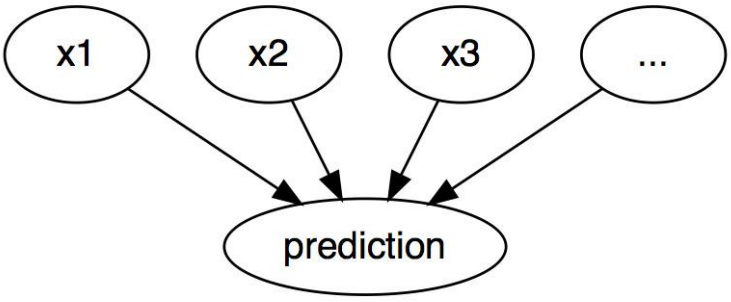

图 4 黑盒机器学习模型输入特征与预测之间的因果关系（不一定是真正的因果关系）。

观察模型预测的变化。我们感兴趣的是预测结果以某种方式变化的情况，比如预测类别发生变化或者预测达到某个阈值。预测的反事实解释描述的就是将预测更改为预定的输出时，特征值的最小变化。最简单的得到特定实例的反事实解释的方法是随机地改变感兴趣的实例特征值，反复试验搜索，在达到预期输出时停止。但是，随机实验方法无法保证“特征值的最小改变”。Wachter 等人[16]提出了另一种产生反事实实例的方法：定义一个损失函数，输入感兴趣的实例、反事实和期望的预测结果，度量反事实的预测和期望预测之间的距离以及反事实与感兴趣的实例之间的距离，采用优化算法优化损失，得到反事实实例。以回归问题为例，其优化目标为：

$$arg\min_{x'}\max_{\lambda}\lambda(f_w(x')-y')^2+d(x,x')$$

其中，w为模型参数，第一项表示反事实x'的预测$f_w(x')$与期望预测y'之间的距离，第二项$d(x,x')$表示反事实x'和感兴趣的实例x之间的距离，参数λ用来平衡预测距离和特征值距离。对给定的λ，保持w不变并优化损失，使得得到的反事实x'尽可能接近x，其预测结果尽可能接近期望的预测。可以采用任何适合的优化算法来优化该损失。如果该机器学习模型的梯度是可见的，则可以采用基于梯度的优化方法，如 ADAM 等。表 1 展示了一个反事实解释的例子[5]：通过学生的平均绩点(GPA)、种族(Race)和入学考试分数(LSAT)预测该学生第一年的平均成绩（Score）。对每一个学生找到反事实的解释，即如何改变该学生的特征，才能使得预测分数大于 0（分数已进行归一化，0 代表所有学生的平均水平）。前两个学

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

表 1 反事实示例。列 1 是预测，列 2-4 是实例特征，列 5-7 是反事实。

生的分数高于平均分，其对应的反事实表明，可以通过降低他们的入学考试分数来将分数降低至平均分。后三个学生的分数低于平均分，其对应的反事实表明，可以通过将种族从 1（黑色）改为 0（白色），或者提高入学考试分数来将其分数提高至平均分。

基于样本的可解释方法还包括对抗样本（Adversarial Examples）、原型和批评（Prototypes and Criticisms）、有影响力的实例（Influential Instances）等。对抗样本是一种反事实实例，通过在原始实例上作出微小的改变而使得机器学习模型做出错误的预测，其旨在欺骗模型[17]，而非解释模型。在原型和批评方法中，一个原型是一个能够很好地代表数据的实例，而一个批评是不能由一组原型很好地代表的数据实例。原型和批评可以模型无关为机器学习模型提供解释。任何能够返回数据中聚类中心的聚类算法都可以用于寻找数据中的原型，如 K-medoids[18]。但这类方法大多无法找到批评。一种同时寻找原型和批评的方法是 MMD-critic[1]。有影响力的实例方法旨在找到对模型的训练和预测有巨大影响的数据实例。通过识别有影响力的实例，我们可以对模型进行诊断，并更好地解释模型的行为和预测。有两种识别有影响力的实例的方法：删除诊断(Deletion Diagnostics)和影响函数(Influence Functions)[19]，两者都基于稳健统计(Robust Statistics)，受异常值或违反模型假设的影响较小。

参考文献

[1] KIM B, KHANNA R, KOYEJO O O. Examples are not enough, learn to criticize! criticism for interpretability [J]. Advances in neural information processing systems, 2016, 29.

[2] 北京智源人工智能研究院. 机器学习的可解释性 [Z]. 2020

[3] 腾讯技术工程. 机器学习模型可解释性的详尽介绍 [Z]. 2019

[4] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning [J]. arXiv preprint arXiv:170208608, 2017.

[5] MOLNAR C. Interpretable machine learning [M]. Lulu. com, 2020.

[6] BREIMAN L. Random forests [J]. Machine learning, 2001, 45(1): 5-32.

[7] FISHER A, RUDIN C, DOMINICI F. Model class reliance: Variable importance measures for any machine learning model class, from the “Rashomon” perspective [J]. arXiv preprint arXiv:180101489, 2018, 68.

[8] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The Elements of Statistical Learning. Volume 1 Springer; New York, NY, USA: 2001 [Z]. Springer series in statistics).[Google Scholar]

[9] RIBEIRO M T, SINGH S, GUESTRIN C. Model-agnostic interpretability of machine learning [J]. arXiv preprint arXiv:160605386, 2016.

[10] GOLDSTEIN A, KAPELNER A, BLEICH J, et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation [J]. journal of Computational and Graphical Statistics, 2015, 24(1): 44-65.

[11] APLEY D, ZHU J. Visualizing the effects of predictor variables in black box supervised learning models. arXiv 2016 [J]. arXiv preprint arXiv:161208468.

[12] RIBEIRO M T, SINGH S, GUESTRIN C. " Why should i trust you?" Explaining the predictions of any classifier; proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, F, 2016 [C].

[13] SHAPLEY L S. 17. A value for n-person games [M]. Princeton University Press, 2016.

[14] ŠTRUMBELJ E, KONONENKO I. Explaining prediction models and individual predictions with feature contributions [J]. Knowledge and information systems, 2014, 41(3): 647-65.

[15] LUNDBERG S M, LEE S-I. A unified approach to interpreting model predictions; proceedings of the Proceedings of the 31st international conference on neural information processing systems, F, 2017 [C].

[16] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR [J]. Harv JL & Tech, 2017, 31: 841.

[17] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:13126199, 2013.

[18] KAUFMAN L, ROUSSEEUW P J. Partitioning around medoids (program pam) [J]. Finding groups in data: an introduction to cluster analysis, 1990, 344: 68-125.

[19] KOH P W, LIANG P. Understanding black-box predictions via influence functions; proceedings of the International Conference on Machine Learning, F, 2017 [C]. PMLR.