# 面向用户的可解释机器学习（论文整理部分）

## 可解释领域对各种解释算法是如何进行分类的？

### 最常见的分类标准是内置可解释（透明模型）与事后可解释[1][2]

**内置可解释（透明模型）**

有些简单模型的最佳解释就是模型本身，这种方法仅限于复杂性较低的模型族，如非0权重较少的线性模型、较浅的决策树和规则模型。在内置可解释中，为了提高模型的可解释性，就需要对模型施加一些限制。比如：约束模型的大小；稀疏性、单调性等性质可以提高可解释性。

内置可解释中更细的分类有

算法透明度：用户理解模型从输入数据产生任何给定输出的过程。模型必须可以通过数学分析和方法被完全探索。

可分解性：解释模型各个组成部分的能力。如输入、参数、计算过程等。并非每个模型都满足。

可模拟性：模型能够被人类模拟，被人类从整体上进行思考、推理。如稀疏的线性模型。可模拟模型同时也是可分解和算法透明的模型。

**事后可解释**

对于已经训练好的且不容易理解（或者说复杂）的模型，通过各种手段来提高其可解释性。有以下技术（方法）来进行事后解释。

文本解释：通过学习生成有助于解释模型结果的文本来解决模型的可解释性问题[3]，该文章使用神经符号主义生成解释文本。

可视化：旨在将模型的行为可视化。如对神经元或者神经网络中间层进行可视化；或者使用t-SNE对高维数据降维，从渲染出的二维可视化效果中可以看出相近的点可能出现在一起。被认为是向不熟悉 ML 建模的用户介绍模型所涉及变量内部复杂交互的最合适方法。

实例分析：提取有代表性的实例进行分析，以更好地理解模型。如可视化的教学系统Rivelo，通过交互式探索一组实例级解释，分析师可以了解二元分类器预测背后的原因。这些解释与模型无关，将模型视为一个黑盒子，帮助分析人员交互式地探索高维二进制数据空间，检测与预测相关的特征。[4] 在强化学习领域，实例就是一个轨迹(St,At,St+1,Rt),如何抽取出重要的轨迹，也有相关文章[5],也有文章研究如何对轨迹进行分析以便于用户理解[6]

代理：建立一个代理模型来对原有的模型进行解释。该方法中比较有名的是LIME[7]及其变体（LIME也可以放到特征归因一类）。LIME 围绕不透明模型的单个预测建立局部的线性模型来解释。还有 G-REX[8],从不透明模型（如集合或神经网络）中抽取规则。针对多个隐藏层的神经网络，有DeepRED[10] 为每个类从输出层 y 到输入层 x 逐层反向提取规则，最后合并所有规则, 得到了根据神经网络的输入来描述输出的规则集 R(x → y)。

特征归因[13]：从输入特征的角度来分析模型的预测, 旨在通过排序或测量每个特征在待解释模型预测输出中的影响、重要性。一个卓有成效的方法是SHAP[9]，SHAP 是对单个预测进行解释，为每个特征分配一个特定预测的重要性值(shap value)。[11] 它提出了一种将网络分类决策分解为输入元素贡献的方法，将每个神经元视为一个可以分解和扩展的对象，然后通过网络聚合和反向传播这些分解，最终形成深度泰勒分解。特别地，针对卷积神经网络模型，有两类经典的方法来得到输入特征对于决策的重要性大小。第一类是[12], 使用反卷积网络DeConvNet 将CNN各隐藏层的特征进行可视化，每层的可视化结果反映了网络的层次化特点，每层可以分别学习到图像的轮廓、颜色和纹理等。另外，通过遮挡输入图像的不同区域并观察输出结果的变化，找到对结果影响最大的特征。另一类是CAM及其变体。[14]。卷积神经网络的最后一层卷积层包含了最丰富的空间和语义信息，于是CAM充分利用了最后一层卷积的特征，并将后面的全连接层替换成了GAP层（全局平均池化），通过可视化的热力图将模型认为最显著的结果显示出来，因此可用于解释模型预测的结果。但是添加了GAP改变了网络模型，还需要重新训练模型，限制了应用场景。于是就有了Grad-CAM[15]，它使用梯度的全局平均来计算权重，不需要修改模型。

事后可解释中还有一种分类方法是从是否与模型有关来分，模型无关的解释方法可以适用于任何模型，另一种是专为某些ML模型进行设计的。比如LIME、SHAP是模型无关的，CAM就只适合卷积神经网络。

## 用户关注的是什么？

1.可信度：即使深度学习模型在一些测试集上表现出了良好的性能，能保证在现实环境的应用场景里还能表现得很好吗？这就要求模型能给出做出决策的原因或者做出决策依据的特征，从而使得用户能够判断该模型的推理是否符合逻辑，是否值得信任。

2.Fairness：用户需要确保模型做出的决策符合伦理与法律的要求，不会出现不公平或者不道德的行为，这同样要求模型做出解释。

3.

对于仅仅想要使用该模型的普通用户，满足可信度与伦理、法律的要求就够了。但对于专家用户、想要改进模型的用户、想从中学习到知识的用户来说，他们对可解释还有更多的要求。

1. Interactivity：能否让用户与模型交互，从而让用户更好地理解模型或者能够融入一些专家知识。

2. teaching:模型能否根据做决策依据的特征进行归纳与提炼，从而帮助用户来学习知识。

3.

## 针对不同的问题，各有什么解决方法？

## 参考论文

[1] Explainable Artificial Intelligence : A Survey (综述)

[2] Explainable Artificial Intelligence (XAI): Concepts, taxonomies,opportunities and challenges toward responsible AI (综述)

[3] Towards Explainable Neural-Symbolic Visual Reasoning (文本解释)

[4] Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations (实例分析)

[5]HIGHLIGHTS: Summarizing Agent Behavior to People (实例分析)

[6] Interestingness elements for explainable reinforcement learning: Understanding agents’ capabilities and limitations (实例分析)

[7] “Why Should I Trust You?” Explaining the Predictions of Any Classifier (代理模型、局部解释、模型无关、特征归因)

[8] G-REX: A Versatile Framework for Evolutionary Data Mining (代理模型)

[9] A Unified Approach to Interpreting Model Predictions (特征归因、局部解释)

[10] DeepRED – Rule Extraction from Deep Neural Networks (代理模型)

[11] Explaining nonlinear classification decisions with deep Taylor decomposition (特征归因)

[12] Visualizing and understanding convolutional networks (特征归因)

[13] 深度学习可解释性研究综述 (综述)

[14]Learning Deep Features for Discriminative Localization (特征归因)

[15] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (特征归因)