# 面向用户的可解释机器学习

## 可解释领域常见的分类方法有哪些？

常见的有以下分类

### 内置可解释（透明模型）与事后可解释

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。对于复杂的模型，为了提高可解释性，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。

事后可解释

从训练好的模型中提取出信息。

代理：建立一个代理模型来对原有的模型进行解释。比如从神经网络集合中提取出符号规则。

特征归因：从输入特征的角度来分析模型的预测。比如Saliency Map、shapley value。

可视化：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。

实例分析：如可视化的教学系统 Rivelo，通过实例分析解释分类器。用户可以交互式地探索实例级解释集，以创建心智模型。

## 全局解释与局部解释

全局解释

## 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，从而是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？

1.直接抽取出模型的规则。如：用决策树这种模型将网络分解，得到网络的整体规则。（CRED、DeepRED)