**韩淑婷周报**

目录

[一 第一周 2](#_Toc21905069)

[1 论文研究 2](#_Toc21905070)

[2 Linux学习 3](#_Toc21905071)

[二、第二周 4](#_Toc21905072)

[1 论文研究 4](#_Toc21905073)

[三、第三周 6](#_Toc21905074)

[1 论文研究 6](#_Toc21905075)

### 一 第一周

## 1 论文研究

**1.1 文献翻译：**

Assessing Heuristic Machine Learning Explanations with Model Counting

用模型计数评估启发式机器学习解释

摘要

机器学习（ML）模型广泛应用于金融、医学、教育等领域的决策过程。在这些领域，ML的结果对人类有直接的影响，比如，用ML评估一个人是否能获得贷款或者一个人是否能从监狱释放。然而我们不能盲目地依赖黑盒ML模型，我们需要去解释ML做出的决策。这推动了一系列ML解释系统的发展，其中包括LIME和它的后继者ANCHOR。由于现有的工具创造出来的机器学习解释都具有启发性，因此有必要对其进行验证。我们提出了一种基于SAT的方法来评估ANCHOR产生的解释的质量。我们将一个训练过的ML模型和一个给定预测的解释进行编码并用作命题公式。然后通过使用最新的近似模型计数器，我们用求出的解的个数来评估所提供的解释的质量。

介绍

近些年ML的发展在很大程度上说明了AI的影响和社会意义。ML的应用领域快速扩展，包括安全至上的一些设置（如智能汽车）和直接影响人类的一些领域（如经济、医学、教育和司法）。提高ML模型置信度的工作不仅包括验证模型属性还包括在ML模型的操作不能被人类直观解释的情况下，解释模型的预测。可解释的AI（XAI）的一般领域既针对自然可解释的ML模型（包括解释树和解释集）的发展，也针对ML模型被视为黑盒的环境（例如神经网络）中的解释计算、分类器的集合，等等。

最著名的XAI方法是基于启发式的，并在特征值空间没有被完全分析的意义上提供了所谓的局部解释。在最近提出的许多计算局部解释的方法中，LIME和它的后继者ANCHOR事两个成功的实例。然而，由于LIME和ANCHOR都是基于启发式的，那么一个自然的问题就产生了：基于启发式的解释在实践中有多精准？本文重点关注ANCHOR（因为它相对于LIME有改进），并提出了一种新颖的方法来评估用于计算局部解释的启发式方法的质量。对于每个计算出的解释，ANCHOR给出一种质量的度量，或者称为解释的估计精度，即解释适用及预测匹配的实例的百分比。从ML模型的编码、ANCHOR计算得到的解释以及目标预测开始，本文建议使用（近似）模型计数来评估解释的实际精度。具体来说，本文考虑了二值化神经网络（BNN），利用了最近提出的针对BNN的命题编码，并在知名数据集上对ANCHOR的质量进行了评估。正如我们所展示的，ANCHOR的解释质量可能有很大的不同，这表明对有些数据集来说，ANCHOR的解释相当准确，但对于另一些数据集来说，ANCHOR的解释可能相当不准确。我们得到的有点出乎意料的实验结论表明XAI需要更多更正式的研究。

**1.2 个人总结**

由于ML模型是黑盒，结果是否是有意义有待解释，因此产生了ML解释系统。针对ML解释系统质量的高低，本文提出一种基于用模型计数的方法来评估。本实验利用BNN对ANCHOR进行评估，得到的实验结果是：ANCHOR的表现不稳定，对于某些数据集来说，ANCHOR得到的解释很精确，对于另一些数据集而言，ANCHOR得到的解释不精确。这代表XAI领域还需要更多更深入的研究。

**1.3 ANCHOR解读**

Anchors: High-precision model-agnostic explanations

ANCHOR：高精度模型解释器

文中介绍了一种新颖的模型无关的解释复杂机器学习模型的系统，该系统使用高精度的规则进行解释，称它为Anchors。Anchors使用局部的充分条件来解释模型行为。首先，设计了算法来有效地生成这样的规则；然后设计了多组实验，针对各种复杂模型和不同领域，来验证Anchors的可扩展性；最后，通过user study来说明，Anchors能允许用户预测模型的行为，并且用户的预测精确度比通过其他解释模型或无解释模型的情况要高。

复杂的机器学习模型确实带来了高准确率，但是也使得模型对于用户来说是一个黑盒，而用户对理解模型行为的需求越来越关注，使得可解释的机器学习开始盛行。可解释的机器学习分为全局可解释的模型和局部可解释的模型。全局可解释的模型一般是特别设计的，局部可解释模型一般是与模型无关的。

**1.4 BNN解读**

Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or −1

二值化神经网络：训练权重和激活度限制为+1或-1的神经网络

通过研究论文对BNN（二值化神经网络）进行理解。论文中提出一种方法训练BNN，在运行时权重和激活值均二值化，训练时进行梯度计算，用Torch7 和 Theano框架做了实验，证明了可以在MNIST, CIFAR-10 和SVHN数据集上训练BNN，效果不错。文中还证明了在前向传播过程，BNN极大地减少了内存消耗（大小和访问次数），用位移操作代替多数计算操作。而且二值化CNN导致卷积核的重复。文献认为在专用的硬件上可以减少60%的时间复杂度。最后，文中写了一个二进制矩阵乘法GPU核，预计在运行MNISTBNN时比未经优化的核快七倍，并且保证其分类精度。决定式的二值化有两种方法。第一种是：正数是1，负数是-1。第二种是：随机式的二值化，对x计算一个概率p，当p大于一个阙值（计算机随机产生）时为+1，否则为-1。第二种方法虽然看起来比第一种更合理，但是在实现时却有一个问题，那就是每次生成随机数会非常耗时，所以除了在文献试验中训练时的一些激活值一般使用第一种方法。

## 2 Linux学习

我参考《鸟哥的Linux私房菜》一书对Linux系统进行学习，在实体机上安装了Linux发行版Ubuntu 19.04。主要学习了在Linux系统下，对于文件的处理；文件权限和用户权限的管理；软件包的管理——包括从源码到apt（在虚拟机上安装CentOS使用yum进行软件包管理）的在线管理；学习了Linux下文件系统的管理，包括分区和挂载等；学习了在Linux发行版下一些Shell的基础。

### 二、第二周

## 1 论文研究

**1.1 基数约束的CNF**

Towards an Optimal CNF Encoding of Boolean Cardinality Constraints

布尔基数约束的最优CNF编码

基数约束的CNF是指对CNF中命题变量的数量作出一些限制。常见的形式有：n个变量的CNF中，至少或者至多有k个变量可以被赋值为真。文中提出了一个统一的框架来解决这类问题。文中提出两种改进现有结果的新编码，一个只需要7n子句和2n个附加变量，另一个要求O(n\*k)个子句。其优点是：仅通过单位传播就可以在线性时间内检测到不一致性。此外文中还证明了任何此类编码所需子句数的线性下界。

1.2 **ML解释：基于启发式的白盒方法**

Deep inside convolutional networks: Visualising image classification models and saliency maps

本文提出了图像分类的可视化模型，这一模型使用了深度卷积网络（Convnet）。基于分别计算分类得分的梯度，我们考虑了两种可视化技术：第一种生成了一张图可以最大化图像分类的得分，由Convnet捕获。第二种对于给定的图片和类，计算其显著性图。我们的研究表明，这种显著图可以用在弱监督目标分割上（使用卷积网络）。最后，我们建立了基于梯度的卷积网络与反卷积神经网络的联系。

这是一种基于启发式的白盒方法，通常用在计算机视觉领域。

**1.3 LIME解读**

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

我为什么信任你？解释任意分类器的预测

尽管被广泛采用，机器学习模型仍然大多是黑匣子。然而，了解预测背后的原因对于评估模型中的信任非常重要。如果一个人计划基于预测采取行动，或者在选择是否部署新模型时，信任是至关重要的。这种理解进一步提供了对模型的洞察，可以用来将不可信的模型或预测转化为可信的模型或预测。本文中，我们提出了一种新的解释技术LIME，它通过在预测周围局部地学习一个可解释的模型，以可解释和可靠的方式解释任何分类器的预测。我们进一步提出了一种解释模型的方法，通过以非冗余的方式呈现具有代表性的个体预测及其解释，将任务定义为子模块优化问题。我们通过解释文本（如随机森林）和图像分类（如神经网络）的不同模型来展示这些方法的灵活性。解释的有效性通过新的实验来证明，无论是模拟实验还是人类实验。我们的解释在各种需要信任的场景中增强了用户的能力：决定是否应该信任一个预测，在模型之间进行选择，改进一个不可信的分类器，以及检测为什么不应该信任一个分类器。

我们在建立模型的时候，经常会思考我们的模型是不是够稳定，会不会出现样本偏差效应， p>>N时候会不会过拟合？我们检查模型稳定，我们进行一些cross-validation来看看各项评估指标方差大不大。 可是如果样本一开始因为采样偏差导致样本有偏，导致模型和实际情况有差异，这个就不太好评估了。同样，p>>N也会有类似的问题，尤其在文本挖掘领域。一般情况，如果特征不是很多的话，尤其像logistic regression这样的模型，我们会把模型权重给打印出来看看，看看训练出的模型结果，是否和人的经验吻合。下面是LIME文章中提到一个文本分类的例子，预测一段文本是无神论相关的还是基督徒相关的。文中分类器预测结果这篇文本是无神论相关的，可是主要区分特征却与人的经验十分不吻合的，这样的模型是不能让人信服的，当我们把这几个特征删除后，预测结果又反向了。我们可以通过人工构建一些由这些特征组成的文本来加入到预测实验中，会大大降低模型性能。

LIME是Local Interpretable Model-Agnostic Explanations的缩写。LIME的目的是试图解释模型在预测样本上的行为，这种解释是可被理解的，并且这种解释是模型无关的，不需要深入到模型内部。作者提出的方法一种局部方法，非全局的，在每个预测样本附近随机采样产生一些样本。选取K个特征，我们可以在采样的样本以及这K个特征上，做加权回归模型。回归模型输出的K个特征以及权重，就是分类器对预测样本的解释。

**1.4 ANCHOR解读**

Anchors: High-precision model-agnostic explanations

ANCHOR：高精度模型解释器

可解释的核心是：用户能足够理解模型的行为，且能精确地预测模型对于样本的预测结果。大多数的局部可解释模型，都是使用一个线性模型去拟合模型的局部行为，这样线性模型能给出样本中不同特征的相对重要性。但是，由于线性模型拟合的是局部的结果，对于一个未知样本，不能确定线性模型的结果是否适用于该样本（即不确定该样本是否在局部范围内），这也就是上文提到的“覆盖度”，线性模型的覆盖度是不确定的。这样就会导致低用户精确度（用户预测模型行为的精确度）。本文提出新的模型无关的可解释模型，基于if-then的规则，我们称它为Anchors。

Anchors给出的解释是模型在局部行为的充分条件，也就是说，若模型满足该条件，则模型一定（大概率）会给出某种分类。局部可解释模型的思想为：当原始模型很复杂以至于难以给出简洁的解释时，聚焦于单个样本的预测来做出解释是可行的。接下来我们在不同的数据集（表格，文本，图片等）及不同的模型（情感分类，词性分类，文本生成）上做了实验来验证Anchors, 并进行了user study来说明Anchors对用户理解模型行为的贡献。

### 三、第三周

## 1 论文研究

本周我主要对研讨班上学习的论文进行整理，第一篇论文的研究内容主要与机器学习在SAT问题中的应用有关，其余论文是关于强化学习和深度学习算法的改进。

CrystalBall: Gazing in the Black Box of SAT Solving

现代SAT求解器通过复杂的启发式方法实现了可伸缩性和鲁棒性，这些启发式方法难以理解和解释。 因此，新算法洞察力的发展主要局限于专家直觉，而对新洞察力的评估则仅限于根据求解程序的运行时间或代理求解程序的运行时间的性能度量。 在这种情况下，是否有可能开发一种框架，以提供对SAT求解器执行的白盒访问，从而可以帮助SAT求解器开发人员和用户综合现代SAT求解器的算法启发式算法？

本文就设计了一个这样的框架称为CrystalBall。 将现代冲突驱动子句学习（CDCL）求解器视为针对不同任务（例如分支，子句内存管理和重新启动）的分类器和回归器的组合，专注于派生分类器以保留或丢弃学习的子句。与基于机器学习的最新技术不同，CrystalBall采用监督学习，并使用从单个SAT求解器运行中提取的大量千兆字节数据进行预测分析。我们将现代CDCL求解器视为针对不同任务的分类器和回归器的组合，例如分支（分支到哪个变量），子句内存管理（哪些学习的子句保留在内存中以及哪些子句抛出），重新启动（终止时） 分支并重新启动等等。为了更深入地了解基础分类器，第一步，我们建立了一个框架，以在求解阶段为SAT求解器提供白盒访问。 我们设想，这样的框架将允许最终用户对他们的试探法的性能获得深入的数据驱动的理解，并帮助他们设计更好的试探法。 我们并不是要取代专家的直觉，而是提出一种“在环专家”的方法，在此方法中，CrystalBall会为专家提供统计上合理的可解释分类器。

因此本文是利用机器学习对SAT算法进行改进，具体做法是将sat算法中分支、二元传播、矛盾分析等处理过程中的一些数据作为特征保存在数据库中，根据这些保存的特征作为输入进行模型训练，并利用得到的结果生成新的学习子句加入到CNF子句集中。

Deep Exploration via Bootstrapped DQN

科学探索仍然是强化学习（RL）的主要挑战。 诸如“贪婪”之类的常见探索抖动策略不会进行时间扩展（或深层）探索； 这可能导致数据需求成倍增长。 但是，在复杂的环境中，大多数用于统计有效的RL的算法在计算上都不易处理。 随机值函数提供了一种有前景的方法来进行广义的科学探索，但是现有算法与非线性参数化值函数不兼容。 作为解决此类情况的第一步，本文开发了自举DQN。 我们证明自举DQN可以将深度探索与深度神经网络相结合，从而比任何抖动策略都以指数级的速度更快地学习。 在Arcade学习环境中，自举DQN可以大大提高大多数游戏的学习速度和累积性能。

TD OR NOT TD: ANALYZING THE ROLE OF TEMPORAL DIFFERENCING IN DEEP REINFORCEMENT LEARNING

我们对强化学习（RL）的理解是由几十年前使用表格表示法和线性函数逼近器获得的理论和经验结果形成的。 这些结果表明，使用时间差分（TD）的RL方法优于直接蒙特卡洛估计（MC）。 这些结果如何在处理感知上复杂的环境和深度非线性模型的深度RL中保持呢？ 在本文中，我们使用专门设计的环境来控制TD在现代深度RL中的作用，这些环境控制着影响绩效的特定因素，例如奖励稀疏性，奖励延迟和任务的感知复杂性。 将TD与无限水平MC进行比较时，我们能够在现代环境中重现经典结果。 但是我们也发现，即使奖励稀少或延迟，有限水平MC也不比TD低。 这使得MC在深度RL中成为TD的可行替代方案。

NOISY NETWORKS FOR EXPLORATION

我们介绍了NoisyNet，这是一种深度强化学习代理，其权重添加了参数噪声，并表明代理策略的诱发随机性可用于辅助高效探索。 通过梯度下降以及其余网络权重来学习噪声参数。 NoisyNet易于实现并且几乎没有计算开销。 我们发现，用NoisyNet代替A3C，DQN和Dueling代理人（分别为熵奖励和-贪婪）的常规探索启发法，可以在许多Atari游戏中获得更高的分数，在某些情况下，它可以将代理人从亚人类提升为超人类性能。

Evolution Strategies as a Scalable Alternative to Reinforcement Learning

我们探索使用进化策略（ES）（一种黑盒优化算法）来替代流行的基于MDP的RL技术（例如Qlearning和Policy Gradients）。 在MuJoCo和Atari上进行的实验表明，ES是一种可行的解决方案策略，可以很好地扩展可用CPU的数量：通过使用基于公共随机数的新颖通信策略，我们的ES实施仅需要通信标量，因此可以扩展到一千多名并行工作者。 这使我们能够在10分钟的训练中解决10分钟内的3D人形行走问题，并在大多数Atari游戏中获得竞争性结果。 此外，我们着重介绍了ES作为黑盒优化技术的一些优点：它对动作频率和延迟的奖励是不变的，可以忍受极长的视野，并且不需要时间折现或价值函数逼近。