KAN网络预测分析应用任务定义

1. 概述

本课题服务于软件城规的国家重点研发项目，对于城市环境品质指标的预测是项目课题中一个子任务。在城市规划学中，来自于城市的建筑、道路、用地、设施和街景等地理空间指标表征了一个城市的空间概况，而城市环境品质指标中的声、光、热和感知等指标与城市的宜居程度息息相关。使用城市的空间信息对环境品质指标进行的预测有助于深入了解城市空间指标对于城市环境品质产生影响的内在机理，从而帮助做出更为合理的城市规划。

Kolmogorov–Arnold Networks (KAN) 是一种打破MLP连接主义的全新机器学习网络架构。KAN网络可以看作是AI for Science任务上的语言模型，科学的语言是函数，KAN 由可解释的函数组成，相比常常被看作黑盒的基于MLP的神经网络，KAN在科学问题上会具有更强的可解释性。

对于城市的空间指标和环境品质指标，项目前期已经进行了数据集的获取和整合。

本课题旨在基于KAN网络，利用城市地理空间信息预测城市环境品质指标，同时研究地理空间指标的重要性分析。

2. 子任务1：使用KAN分析自变量中的关键因素

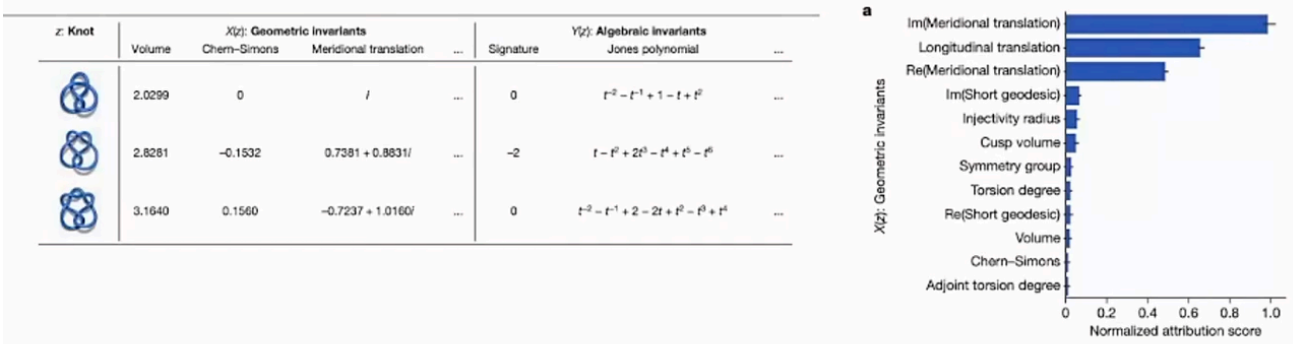
使用KAN的监督或者无监督训练，可以用于**发现自变量之间的潜在关系以及分析影响最终因变量的关键因素**，用于AI4Science任务中可以增强模型的可解释性，同时也可以减少更多无关的自变量输入，减少噪声提高模型精度。

科学的语言是函数。KAN 由可解释的函数组成，因此当人类用户注视 KAN 时，就像使用函数语言与它进行交流一样。事实上，不同领域的科学家对于哪些功能简单或可解释的看法可能不同。因此，科学家更希望拥有一个能够讲科学语言（功能）的人工智能，并能方便地与个别科学家的归纳偏见进行交互，以适应特定的科学领域。

科学家与AI合作：在数学领域的一个例子，Knot Theory问题。

A math example: knot theory

In Deepmind's great paper "Advancing mathematics by guiding human intuition with AI", they showcase how AI + Scientists collaborate together to discover a relation in knot dataset.



(1) They used attribution methods to identify important relevant variables (AI).
(2) They come up with the "slope" concept and prove it to be correct (Human).

Can we rediscover these results with KANs?

科学家们将MLP简单的应用在Knot Theory问题，发现了3个重要的因素，从而定义slope去解决了这个问题。作者希望用KAN去做同样是事情，看看KAN是不是也能够发现这个规律

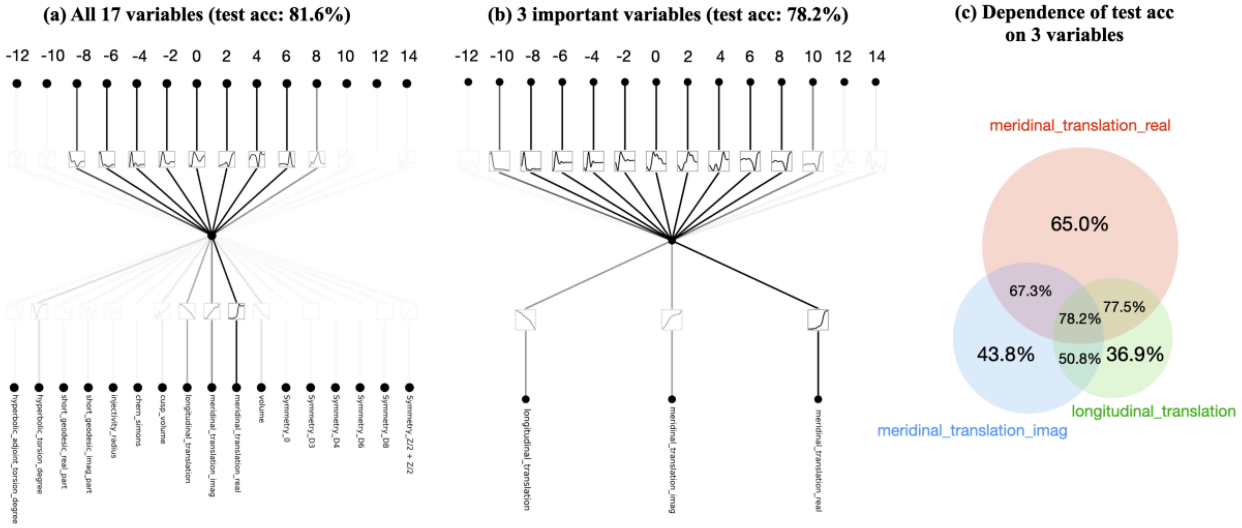


Figure 4.3: Knot dataset, supervised mode. With KANs, we rediscover Deepmind's results that signature is mainly dependent on meridinal translation (real and imaginary parts).

实验结构是KAN也能发现这三个重要因素，并且会比MLP去做更加直观

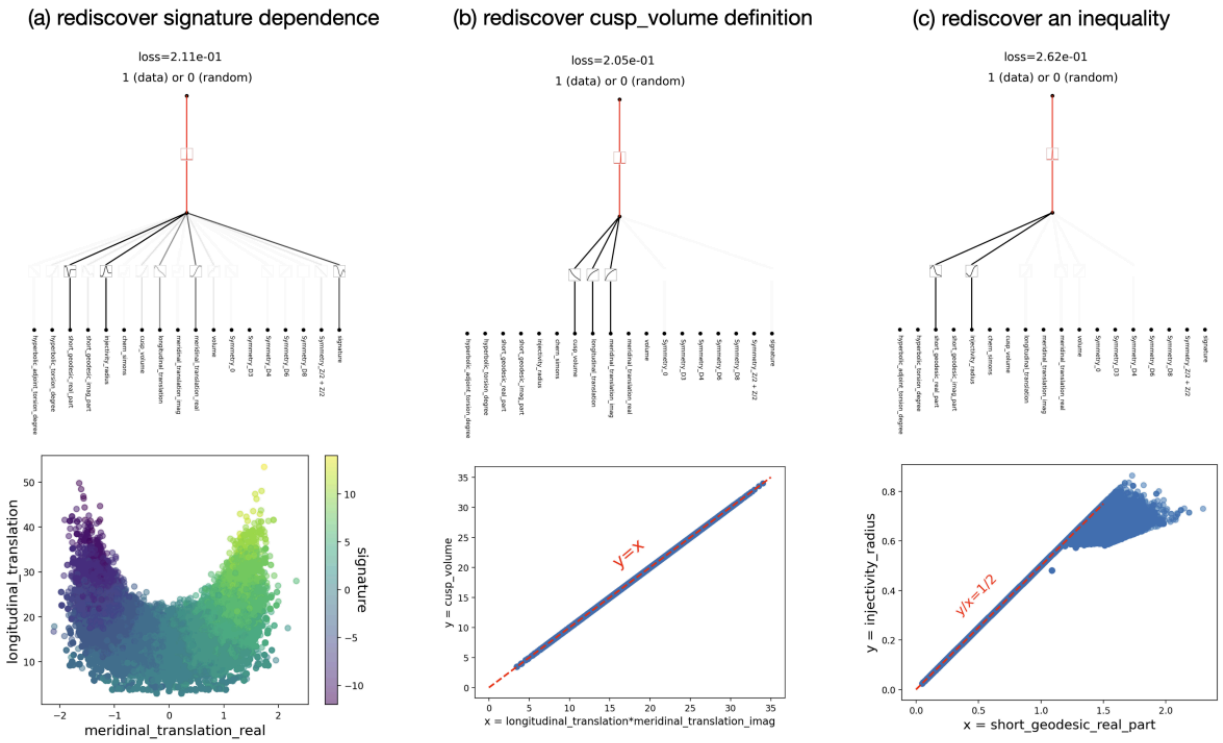


Figure 4.4: Knot dataset, unsupervised mode. With KANs, we rediscover three mathematical relations in the knot dataset.

即使在非监督式学习下，KAN也能找到这些潜在的变量关系

3. 子任务2：利用基于KAN的模型提高预测精度

对于夜间灯光强度以及地标温度和热岛强度等因变量指标，现在能够提供3种基线模型分别进行单任务的预测训练，并形成了一套十折嵌套交叉验证的评估体系。

其中，按照标准的十折嵌套交叉验证理论，在训练和测试模型时，采用这十折的函数进行嵌套交叉验证，其中会分为内循环和外循环两个部分：

外循环（Outer Loop）：

- 将数据划分为十折，每次选择一折作为测试集，其余九折作为训练+验证集。
- 外循环在十折上轮流进行，确保每个折都能作为一次测试集。

内循环（Inner Loop）：

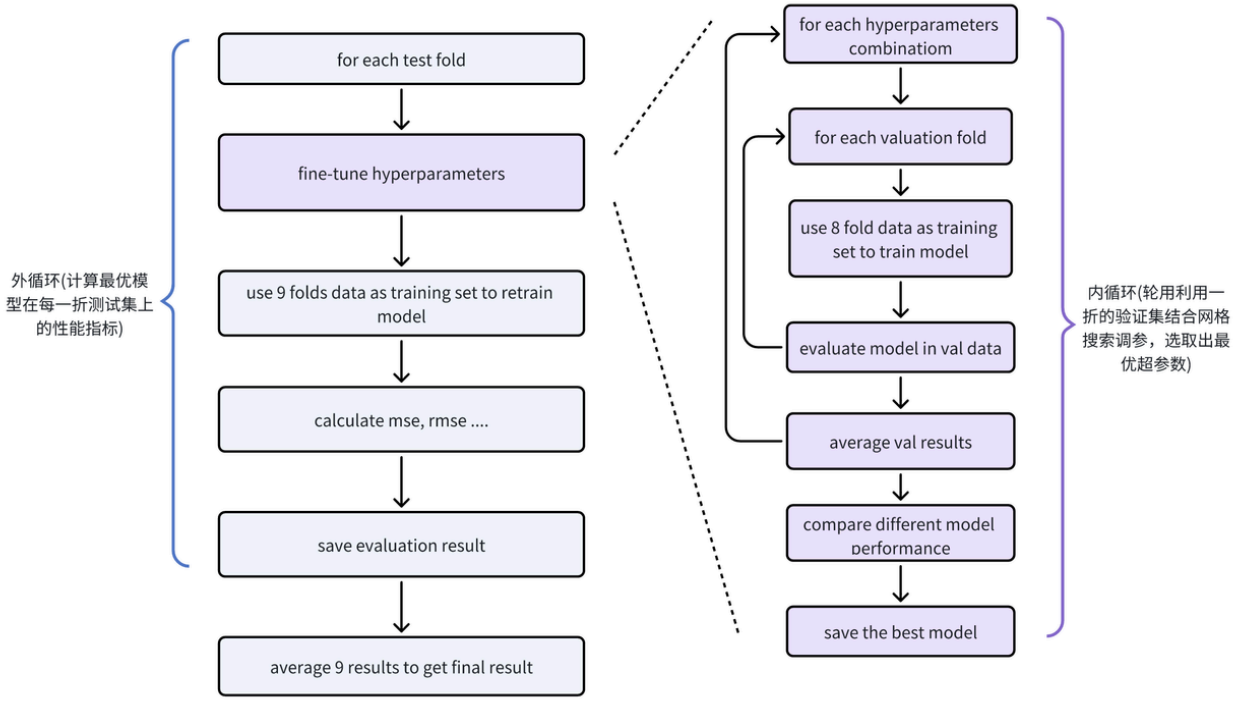
- 在每次内循环开始前，需要确定一组要验证的超参数组合，这里采用网格搜索的方式枚举不同的超参数组合。
- 在外循环中剩下的九折数据中，进一步做交叉验证，用于超参数选择。
- 每次内循环中，选择其中的一折作为验证集，其余八折作为训练集。
- 内循环在八折训练集+一折验证集上轮流进行，从而在不同的超参数组合下获得验证的回归指标，用九次训练的指标平均值代表模型的性能。

超参数选择后重新训练并测试：

- 网格搜索所有超参数并评估后对比所有结果，选出最佳的超参数组合。
- 使用这个超参数组合，在外循环的九折训练集（即之前的训练+验证数据）上重新训练模型。
- 在外循环的测试集（当前轮的测试折）上评估模型，记录下该测试集上的模型性能。

最终结果:

- 外循环完成十次后，得到十组测试集上的性能指标。
- 将这十组性能指标取平均，作为模型的最终性能评估。



解释一下超参数调整中的网格搜索算法，网格搜索算法是一种常用的深度学习超参数调整方法。它通过**遍历给定的超参数组合，找到最优的超参数配置**。具体来说，网格搜索算法将每个超参数的可能取值列举出来，形成网格状排列。然后，通过交叉验证的方式，对每一组超参数组合进行训练和评估，最终选出在验证集上表现最好的超参数组合。例如，对于某个模型要调整学习率和模型层数两个超参数，其中学习率取值为[0.1, 0.01]，对于模型层数取值为[1,2,3]，按照排列组合需要评估2×3=6组超参数对应的模型性能，选出这6组超参数中最优的一组。

基于该评估体系构建新的基于KAN的模型进行预测并评估结果，结合自变量关键因素的分析**提高模型的预测精度**。