2019/12/20 UPDATE: 评测方法

一般而言，ROC 曲线是针对二分类问题，而对于物理粒子的多分类问题，ROC 曲线是使用 one vs all 方法将多分类问题转换为 4 个二分类问题，并采用 macro 取其算数平均数。具体细节内容可以在网上查找。

背景

宇宙中大多数物质由原子构成，原子又由原子核和电子组成。其中，电子是基本粒子，但原子核又可分为质子和中子，并可进一步分为夸克和胶子。这些夸克和胶子的相互作用非常强烈，以至于只有通过极高能量的质子对撞才能让它们摆脱束缚。在高能碰撞时可以产生包括夸克和中子在内的大量粒子，向某个方向射出，这些粒子团被称为喷注（jet）。

喷注可以分为 1）胶体喷注，2）轻夸克喷注，3）魅夸克喷注，4）美夸克喷注。由于它们的不同内在特性（如质量和色量子数），不同种类的喷射经历不同的衰变过程，其内部结构也在实验中显示出不同的观测值。

尽管在理论物理模拟中可以很容易地识别出喷注的味道，但目前在实验中没有可靠的方法可以对所测量的真实喷注进行分类。因此，**开发一种稳健的算法来识别喷注味道，将让我们可以更直接地比较实验测量和基本粒子理论**。

比赛任务

本次比赛提供**粒子碰撞数据集**，其中包含对撞中产生的喷注信息（质量、能量、方向等），以及相关的碰撞事件信息和喷注中所包含的粒子信息，要求选手根据喷注的性质（如喷注所含的粒子数、喷注能量、喷注质量、喷注方向），以及喷注中所有粒子的特征（方向、质量、能量等）和对应的碰撞事件，把**喷注分成四类中的一类**。
本次比赛分为简单赛道和复杂赛道，依次进行。简单赛道只要求选手根据**喷注属性**的数据集进行分类；复杂赛道在此基础上，又加入**喷注所含粒子的属性**文件和**碰撞事件**文件，数据的体量和维度剧增，难度也相应加大。

简单赛道（2019 年 12 月 9 日至 12 月 25 日）

选手根据喷注的性质（喷注所含的粒子数、喷注能量、喷注质量、喷注方向）进行分类。（简单赛道开放时间较短，建议提前报名参赛）

复杂赛道（2019 年 12 月 26 日至 2020 年 2 月 28 日）

选手根据喷注的性质、喷注中所有粒子的特征、以及喷注所在的碰撞事件进行分类。

粒子碰撞数据集

本数据集共包含 200 多万条喷注信息，分为 **EVENT**、**JET**、**PARTICLE** 三类文件，三者为上下层级关系，一个碰撞事件（EVENT）会产生若干个喷注（JET），而一个喷注中会包含若干个粒子（PARTICLE）。EVENT 文件是对碰撞事件的描述，JET 文件详细说明了喷注的属性，PARTICLE 文件进一步描述了喷注中所含各个粒子的属性。详情请见数据页。（https://www.biendata.com/competition/jet/data/）

参考文献

自 2016 年开始，物理学界开始尝试将深度学习引入喷注分类任务中。在此过程中可以发现，最新的机器学习技术创新可以相当显著地提升模型性能。2016 年，华盛顿大学和加州大学的团队公布了一批 1000 万喷注的数据，并开发出了深度学习算法把这些喷注分类。不过喷注数据颗粒度较粗，没有喷注中每个粒子的数据，而且分类中把轻夸克和胶子作为一类，没有进行进一步分类 [1]。

目前，已有多种机器学习技术已经在相关数据集上得到应用。2017 年，麻省理工学院的研究团队将模拟喷注数据中粒子的密度转化为二维图片，并用卷积神经网络等计算机视觉技术对图片进行分类[2]。同年，多个团队报道利用喷注衰变产生的树状演变结构，可以采用自然语言处理中的 RNN 及 LSTM 网络，显著提升分类的准确率[3][4]。一篇 2019 年发表的论文表明，如果考虑一些物理学家设计的变量作为特征，最高能把胶子-夸克分类的 ROC AUC 数值提升超过 10%左右，达到 0.899。[5]

注：[1]和[3]将轻夸克和胶子作为一类（轻味道粒子），没有进行进一步分类。[3]和[4]区分了夸克和胶子。[5]区分了轻味道粒子中的夸克和胶子。

[1] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban and D. Whiteson, Jet flavor classification in high-energy physics with deep neural networks, Phys. Rev. D 94 (2016) 112002 [arXiv:1607.08633] [INSPIRE].

[2] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, JHEP 01 (2017) 110 [arXiv:1612.01551] [INSPIRE].

[3] S. Egan, W. Fedorko, A. Lister, J. Pearkes and C. Gay, Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC, arXiv:1711.09059 [INSPIRE].

[4] Cheng, T. Recursive Neural Networks in Quark/Gluon Tagging, Comput. Softw. Big Sci. 2 (2018), no. 1 3. arXiv preprint arXiv:1711.02633.

[5] Hui, L., LUO, M., Kai, W., ZHU, G., & Tao, X. Quark jet versus gluon jet: fully-connected neural networks with high-level features. SCIENCE CHINA Physics, Mechanics & Astronomy. (2019)