数据格式说明与作业帮助

整个第二阶段的作业是第一阶段的作业相反的一个过程，需要各位同学根据PMT的输出波形信息，推断得到事例的粒子类型。

数据说明

我们将数据储存在HDF5文件中，其中有三个表：ParticleTruth每个事例中粒子的粒子种类；Waveform保存波形信息。每个表都有对应的事例编号或通道编号，如下所示。

Waveform

EventID(int16)	ChannelID(int16)	Waveform(int16[1029])
1	0	
1	1	
1	2	
…	…	

PEGuess

EventID(unit32)	ChannelID(uint32)	PETime(unit16)	Weight(float64)
0	0	365	0.5000
0	0	393	0.6000
0	1	296	0.7000
…	…		…

提供的训练集数据还会在提供ParticleTruth以及PETruth的表以供使用。其中，PETime描述的是光电子引起波形的起始时间，Weight是基于波形分析算法中Lucy迭代器得出的一个估计， 是对 PETime 的补充，补充说明在该PETime下，波形信号由多少个光电子叠加得到。例如Weight=1 表示 PETime 处有 1 个光电子，因此 可以把训练集的 PETruth 表看成有一列省略的 Weight=1 的列。Weight=2 就表示该 PETime 处有 2 个光电子重叠。进一步推广，如果分析算法认为在 PETime 处有 50% 的几率有 1 个光电子，则 可以把该行的 Weight 写为 0.5。

生成的原始数据集中Weight是float型的，而训练数据集中的Weight是uint8型的，这会丢失掉一些迭代信息，为了保证测试数据集的准确性（当然，这前后的数据类型变化体现了我们教学团队的认识变化），我们保持了原始生成数据类型的float型，而来源于`Ghost Hunter2020`的训练数据集不做改变，继续使用uint8型；而最终的测试数据集给同学们提供了原始的float的Weight，你可以根据需要决定是否需要将其在使用时转化为uint8型。

ParticleTruth

EventID(int64)	Alpha(int16)	E(single)	x(single)	y(single)	z(single)
1	1	0.5692	344.63010	-217.0379	444.1518
1	1	15.1537	-74.691422	-413.0363	421.1786
1	1	1.9479	-70.241890	-62.0475	-48.8988
...	...	...	....	....	....

PETruth

EventID(int64)	Channel(int16)	PETime(int16)	PEType(int8)
0	0	245	1
0	0	277	1
0	1	248	1
...	...	...	....

其中，PEType表征了该Channel接收到的光子的类型是切伦科夫光子还是闪烁光光子。

使用HDFView可以打开数据文件，查看文件的大致结构。

提交文件

大作业提交HDF5格式的文件，在文件中写入一个简单的表格，一列为EventID，一列为alpha粒子的概率，

EventID(int16)	alpha（float32)
1	0.0
2	0.5
3	1.0
...	...

alpha列表示在你的鉴别结果中，有多大的概率认为该Event为alpha粒子，如alpha=0.5，则你的结果认为该Event有50%的概率为alpha粒子。

训练与测试用的数据可以从 清华云盘下载，其中 `train.h5` 是训练数据（摘自Ghost Hunter2020训练数据集），`problem.h5` 是最终测试数据，需要提交结果到CrowdAI进行排位评分。`example.h5` 是输出样例。评测数据是使用与训练数据相同的参数生成的。`PMT_Position.txt` 提供了光电倍增管的位置信息，可根据需要自行选用。

`train.h5` 数据集摘自 `Ghost Hunter2020` 的训练数据集。同学们如有需要，可登陆 `CrowdAI` 数据平台后，前往 Ghost Hunter2020决赛的dataset下，下载测试集以训练使用。

作业帮助

对于数据分析过程需要进一步学习的同学，可以到 Ghost Hunter2020赛事主页找到一些可能有帮助的文档或者视频。