作业4总结

本次作业同学们完成较好，主要是有两方面问题需要进一步讨论

数据预处理

不少同学都对各种分类模型进行了超参数调优，但是比较少的同学提及了对数据的预处理。一方面，我们提供的数据中包含全0列，应当适当对这些数据进行清洗；另一方面，我们提供的数据维度较高，是否将全部数据都送给模型进行学习是最佳的选择？我们所提供的数据是系统的运行数据，潮流、电压等特征彼此是相互决定的，这些数据之间本身存在一定的相关性，如果对输入特征先进行筛选或者降维，会不会使得训练的变得更容易，更容易调整超参数呢？

假阴与假阳

首先需要界定的是对于我们本次作业更关心的到底是假阴还是假阳。阳性往往被定义为我们关心的事件，所以对于我们这个例子中，不稳定是我们需要格外关心的事件，所以按照一般的惯例，不稳定应当被定位为阳性。所以，将不稳定预测为稳定应当是假阴性，而不是假阳性。值得说明的是，假阳性和假阴性的定义跟01的赋值无关，很多同学认为我们作业里是 将0预测为1 是更严重的，所以我们关注的是假阳性，这样的理解是有偏差的。

另外如何根据precision和recall来改善模型呢？同学们主要有3种做法

1. 调整样本权重，给不稳定的样本增加更多的权重，模型训练会让不稳定样本的判断结果更加准确
2. 改变阈值，对于稳定的判定阈值提高，对稳定样本的预测概率提升到一定水平，才能认为样本稳定
3. 改变目标函数，模型训练的目标改为F1得分或者recall等直接可以反映假阴性或假阳性的指标

也有同学，根据ROC曲线，判断目前得到的模型的accuracy和recall是否实现了一个比较好的tradeoff