作业3 总结

本次作业大家完成的比较好，主要有以下这些问题

输入输出

LSTM是一种处理时序数据的神经网络，所以他的输入天然地包含了多个时间步，每个时间步都输入了一系列的特征。因此对于本次作业，输入特征的维度是 48x20，一个样本就是一个矩阵。且对于本次作业的代码，是滚动预测的，也就是说每次都是用前48h的数据来预测后24h的数据，训练样本每次向后平移1h。有同学只观察了输入的excel文件，说训练的特征是w1-w25，但是事实上，如果仔细观察代码， 这些特征并没有被输入。有同学对这些特征进行了相关性的分析，发现这些量测相关性较高，实际上没有必要全部输入给神经网络，这些观察很好。同学们可以继续思考的问题，为什么时间特征使用了独热编码，而不是采用1-7/1-12这样的编码方式？既然是滚动预测的，那么同一个时间点理论上可以被预测48次左右，那么最终做预测的时候，应该选择哪一个预测值作为结果呢？

损失函数

quantile这个词本身是分位数的意思，对应的是分位数损失。有同学将其称为量化误差，助教不知道这个名词是怎么样翻译过来的。

我们课堂上使用的损失函数是MSE loss，是最小化均方误差。quantile loss更多用于概率预测或者区间预测，取不同的quantile值是对不同分位数进行预测，相当于会得到一个置信区间。quantile=0.5时，相当于在对中位数进行预测。很多同学说发现预测结果整体偏大或偏小，因此尝试调小或调大quantile这个参数，这样的尝试有有意义的，但是助教建议：此时还是发生了欠拟合，所以应该优先调整网络结构等超参数，调整quantile虽然看起来会更好，但是其实不是在做中位数的预测，可能缺少理论依据。

测试

大家都顺利完成了在测试集上测试模型的任务。但是还有两点需要讨论

有些同学把测试集的测试结果也加入了学习曲线。正常来说学习曲线包含的内容是训练集和验证集。测试集一般只做测试，不利用测试集对神经网络进行超参数的调整。所以测试集数据放入学习曲线本身并没有很大的意义。为了保证测试集的独立性，在调整参数时也不应该参考测试集上的结果。

一些同学选取了task2中的数据作为测试集，这一点可能要小心。我们原本的代码中是利用task1的数据进行训练，如果希望利用task2的数据进行测试，就需要保证task1和task2的数据是同分布的，否则测试结果是没有意义的。比如你建立了北京负荷的预测模型，但是用广东/新疆的负荷数据做测试，这样大概率不会得到令人满意的预测结果。