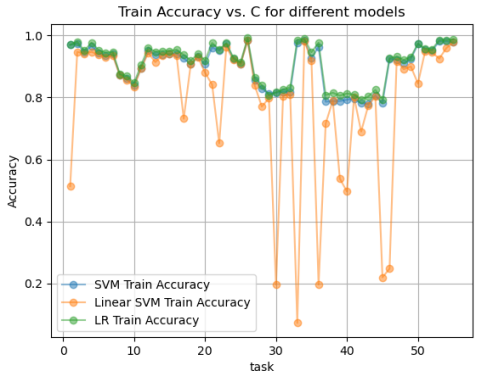
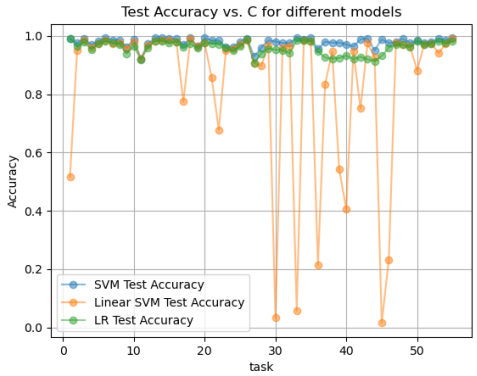
21302010040_叶天逸_PJ1-蛋白质-报告

任务一: 只需要根据cast标签来划分diagram就可以了

任务二

模仿SVM的代码即可
不同model关于task的准确率图: LinearSVM比较动荡



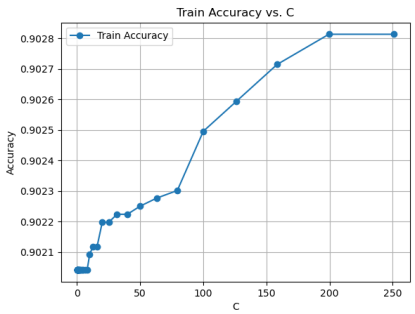


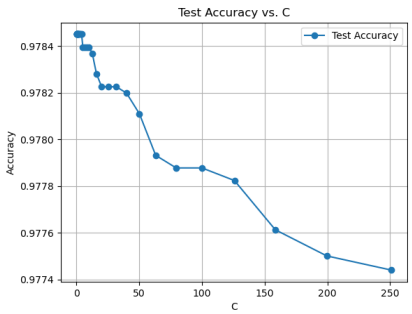
SVM 很快，LR 慢一点，LSVM 慢, 因为太慢, 在特征工程之后分析其性能 特征工程后发现反而LR和LSVM更快了, ==写在任务四后面==

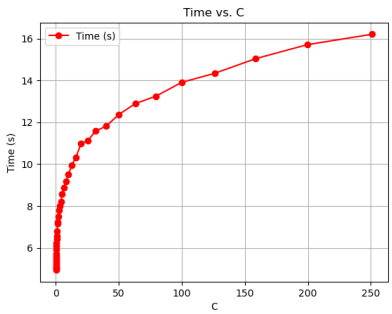
任务三

研究C的影响

C从 10^{-1} 到 10^{2} 指数增长; 训练准确率上升, 而测试准确率下降, 用时变长
解释:
C较高的时候模型可能过拟合: 模型在训练集上表现得很好，但是在未见过的数据上表现不佳。模型过于复杂，以至于能够很好地拟合训练数据中的噪声和细节，但是却无法泛化到测试集中的样本。
当 C 的值较小时，模型的正则化效果较强，会对模型的复杂度进行限制，这样可以防止模型过拟合。随着 C 的增加，正则化效果减弱，模型的复杂度增加，使得模型更容易过拟合。
同时训练时间也会增加。因为模型更复杂，训练过程需要更多的计算资源和时间来优化模型参数。
SVM rbf enf=False



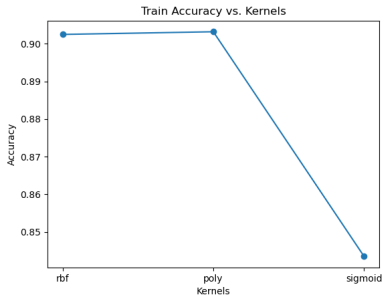


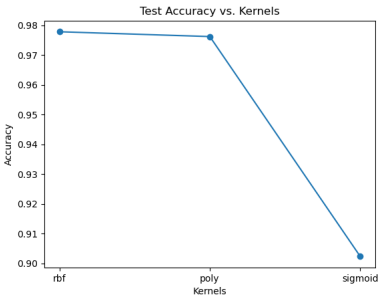


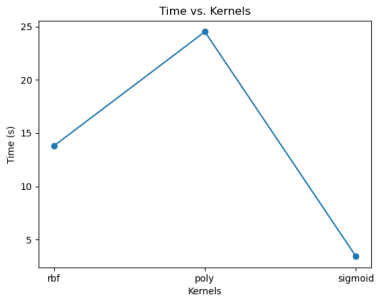
研究核函数的影响

研究了C 为0.01和100的情况,发现图像都是下面的形态
poly 比 rbf 训练准确率高, 测试准确率低, 也就是说, poly 比 rbf 更加过拟合
sigmoid的表现远远不如其他两个, linear更是收敛不了
poly 用时最长, 在25s左右, rbf 14s, sigmoid极快, 3s，linear 时间无穷....
在使用特征工程后，明显加快，linear也收敛了
解释:

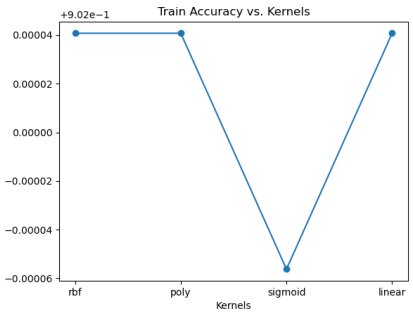
1. 模型可能是线性不可分的, 或者是数据维度太多, 所以linear用不了, 等特征工程
2. poly 通常更灵活，能够在训练集上更好地拟合数据，因此训练准确率更高。然而，由于其灵活性，poly 也更容易过拟合，导致在测试集上的泛化能力较差，测试准确率较低。
3. sigmoid 在某些情况下可能表现不佳，特别是当数据分布与其假设不匹配时。sigmoid 通常用于处理二元分类问题，并且在其他情况下可能不太适用，因此性能可能不如poly 和 rbf。

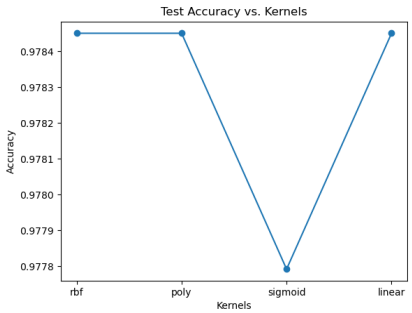


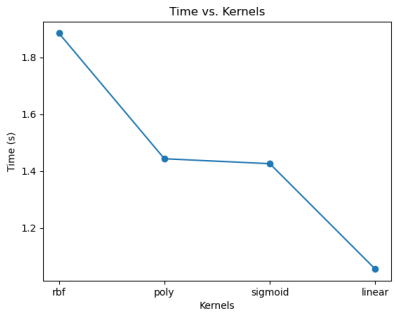




使用Chain的特征工程后进行K测量





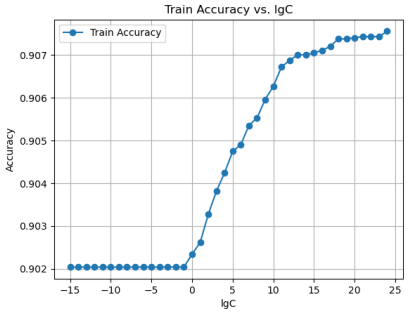


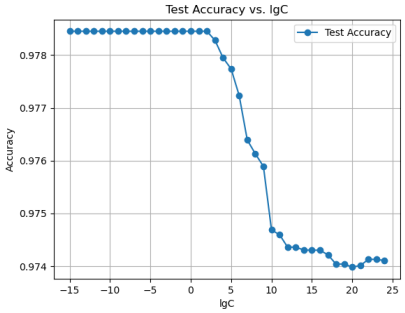
任务四

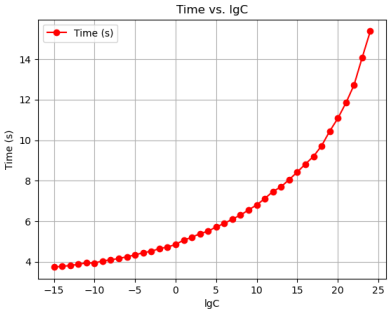
以svm为模型, 以不同的特征作为特征向量:

原子名 Atom

进行特征工程-原子名后, 进行C测量.
当特征维度减少后, 准确率对C的增大更加敏感了, 也就是, 训练准确率快速上升, 测试准确率快速下降
解释: 特征维度减少后，模型更容易过拟合训练数据，因为模型可以更灵活地适应训练数据的特征。因此，增大正则化参数C可能会导致模型在训练数据上的准确率迅速上升，但在测试数据上的表现下降.
在这种情况下, C不能取大
为了更好看到, 采用C的对数: 发现准确率能呈现逻辑斯谛

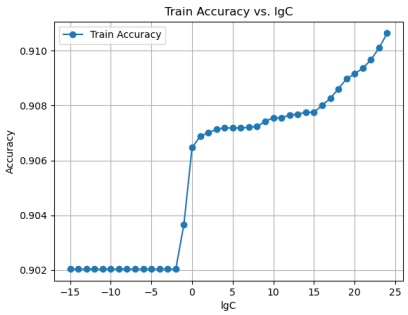


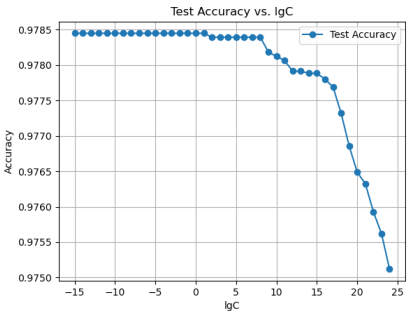


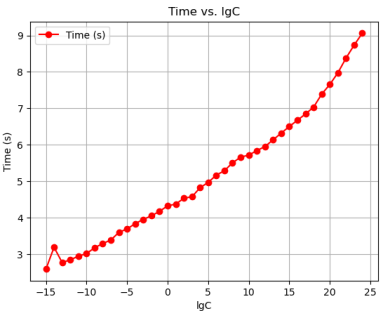


残基 Residue

我们发现这里时间基本上关于lgC线性了, 准确率也没有明显下降.

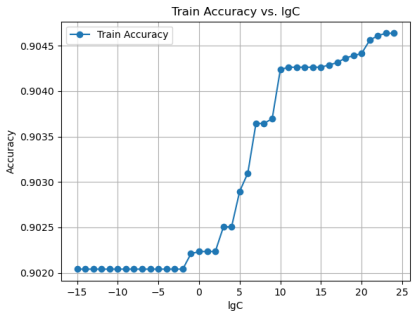


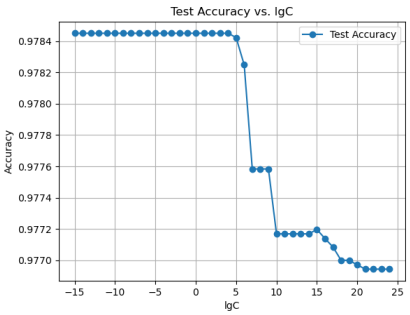


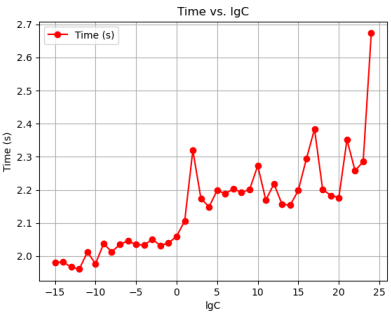


链 Chain

更快了, 并且准确率也没有下降





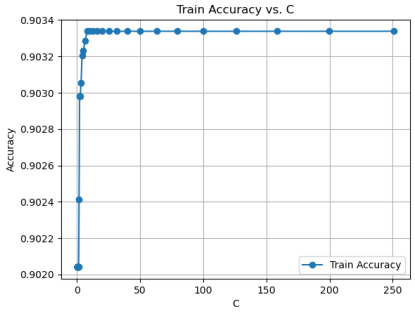


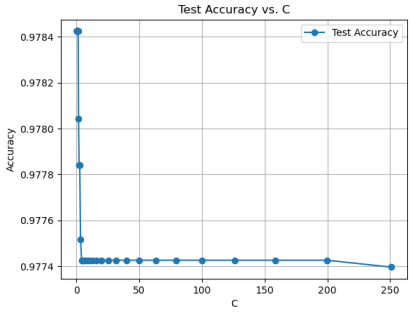
任务二的延续

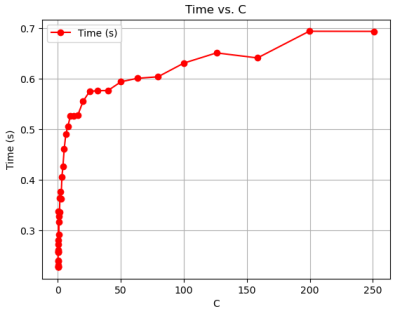
这里已经完成了任务三和四，选择Chain特征， 开始展示LR和LSVM（变快很多）的关于C的图像

LR

当 C=0.0316， Test Acc 0.9784，Time 0.228 s

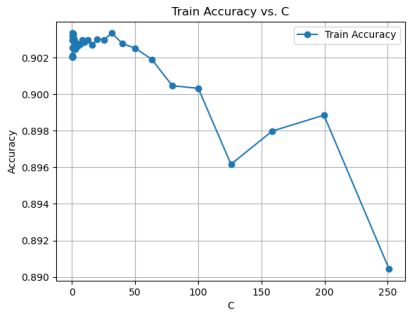


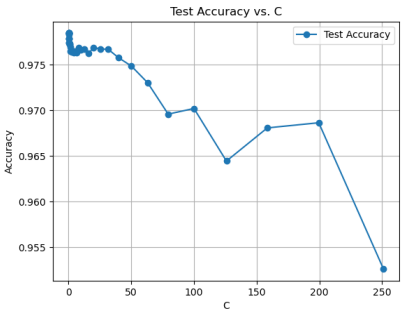


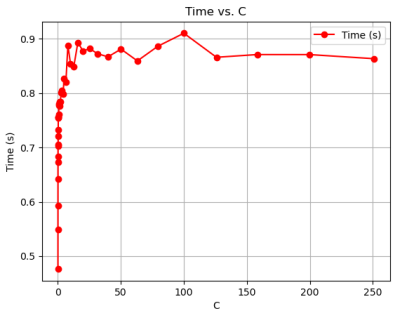


LinearSVM

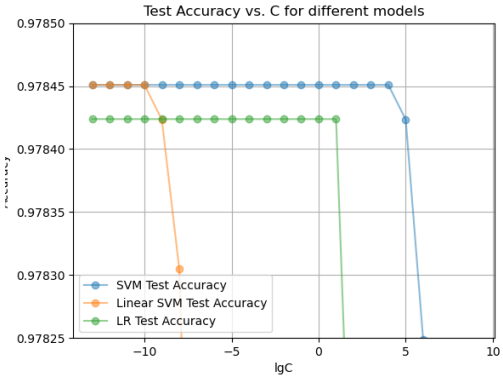
在C取0.0316 时 有 Test Acc = 0.978451， 用时 0.476166 s

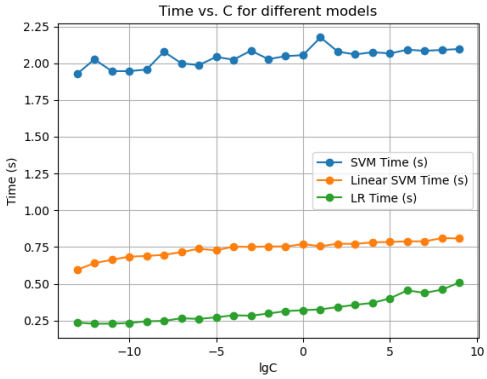






对不同模型横向对比





问题和点

1. 正负样本不平衡，召回率低
2. 采用了单独的代码体来持久化中间结果，详见代码，代码解释详见README
   main 是主函数
   fea 实现了多种特征工程
   CMeasure 是将C(或lgC)作为参数测量正则化对Acc, Time的影响
   KMeasure 是将K作为参数测量核函数对Acc, Time的影响
   fig 是所有输出的图片, out 记录了所有task的结果, data 记录了特征工程的结果npy