作业5 总结

本次作业主要有这样几个问题

最重要的问题是到底聚类要聚成几类。很多同学说要聚类聚成两类，为什么要聚成两类呢？哦，很多同学说因为稳定和不稳定是两类。但是进行聚类的时候，事实上是无标签的数据，你怎么知道这些数据的标签是一个二分类的呢？实际在进行非监督学习的时候，我们是不知道数据背后的真实标签的，只不过本次作业我们想额外观察，聚类后结果的稳定性是否存在一定的分布规律。所以，要聚成几类不应该根据稳定和不稳定分为两类，而是应该根据聚类本身的一些指标。另外，聚类是非监督学习方法，不是监督学习方法，学出来的东西没有正确答案，所以用精度、召回率等指标评价聚类结果也没有意义。根本的就是，对于一个非监督学习问题，应该用非监督学习的视角去看待这个问题。

在技术层面，还发现有一个小问题。在实际操作的时候，我们往往是先降维再聚类。为什么要这样操作？降维的根本是是希望避免维度灾难的问题，维度灾难的一个体现就是数据的稀疏性：在高维空间里，任意两个点之间的欧式距离都很远。很多聚类方法都是基于欧式距离的，如果不进行降维就进行聚类，聚类方法很可能会失效。

另外有同学提到了这样一个思考，如果对于稳定性判别（监督学习问题），先进行降维是不是更好的方式？如果同样只利用m个特征，那么这m个特征是从原始数据降维得到的，还是通过筛选得到的？有同学发现，降维很可能不如选取更有意义的特征。这给了我们很好的启示，面对一个实际问题，降维只是一般化的处理方式，很可能根据物理意义结合数据驱动方法筛选更重要的特征会取得更好的结果，这就是 no free lunch定理。