大数据分析第五次作业

【使用方法以及理解】

1. Kmeans：要将鸢尾花数据集分为三组，则需要产生三个质心，先随机产生三个质心，然后分别计算每个点距离质心的距离，归入对应的组，然后修正质心为所有值的平均值，直到所有数据均被使用，无法更新质心，此时即可按照距离三个质心的距离将数据集分为三个组。(自己模仿sklearn实现发现正确率有点低)
2. 决策树：根据基尼系数的计算，确定决策树每一个节点的属性，然后根据对应属性是否满足将数据集分为三份。（使用sklearn库）
3. SVM：使得决策边界满足离其最近的点到其距离最大化。（使用sklearn库）
4. BPNN：输入部接收外来的输入训练集X，由训练部进行神经网络的权系数W调整，然后由输出部输出结果（正向传播）。在这个过程中，期望的输出信号可以作为训练信号输入，由该信号与实际输出进行比较，产生的误差去控制修改权系数W（反向传播），从而实现对应的测试集数据的预测。（使用Tensorflow神经网络）

【数据集划分方法】

使用留出法进行数据集的划分，将一部分作为训练集而另外一部分作为训练集，而另外一部分作为测试集，由于训练集与测试集均要满足类似分层而代替全体样本的条件，所以使用将鸢尾花数据集全体打乱的方法，利用np中的种子相同的方法对数据集进行划分产生测试集与训练集。（我使用了sklearn中留出法划分函数train\_test\_split，不过准确率不如上面的打乱划分）