《机器学习》课程实验报告



**网络与信息安全学院**

**班 级：** 2118021

**姓 名：** 夏雨轩

**学 号：** 21009201006

**提交时间：** 12.22

**一、实验名称**

* 监督学习之KNN；
* 无监督学习之K-means聚类、DBSCAN算法实验；

**二、实验目的**

* 掌握KNN的原理，学会利用KNN解决分类问题；
* 理解K-means聚类算法、DBSCAN算法的基本原理；
* 学会用python实现KNN、K-means和DBSCAN算法。

**三、实验工具**

Python集成开发环境(IDE)

(1) Anaconda: https://www.continuum.io/ （推荐）

(2) IDLE: Python解释器默认工具

(3) PyCharm: https://www.jetbrains.com/pycharm/

(4) 实验数据集：Python的scikit-learn库中自带的数据集，可使用datasets.load\_iris()载入。

**四、实验原理**

(1) KNN（K-Nearest Neighbor）算法原理

存在一个样本数据集合，也称为训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每一数据与所属分类对应的关系。输入没有标签的数据后，将新数据中的每个特征与样本集中数据对应的特征进行比较，提取出样本集中特征最相似数据（最近邻）的分类标签。一般来说，我们只选择样本数据集中前k个最相似的数据，这就是k近邻算法中k的出处，通常k是不大于20的整数。最后选择k个最相似数据中出现次数最多的分类作为新数据的分类。

说明：KNN没有显示的训练过程，它是“懒惰学习”的代表，它在训练阶段只是把数据保存下来，训练时间开销为0，等收到测试样本后进行处理。

(2) K-means算法原理

k-means算法是一种聚类算法，所谓聚类，即根据相似性原则，将具有较高相似度的数据对象划分至同一类簇，将具有较高相异度的数据对象划分至不同类簇。聚类与分类最大的区别在于，聚类过程为无监督过程，即待处理数据对象没有任何先验知识，而分类过程为有监督过程，即存在有先验知识的训练数据集。

k-means算法中的k代表类簇个数，means代表类簇内数据对象的均值（这种均值是一种对类簇中心的描述），因此，k-means算法又称为k-均值算法。k-means算法是一种基于划分的聚类算法，以距离作为数据对象间相似性度量的标准，即数据对象间的距离越小，则它们的相似性越高，则它们越有可能在同一个类簇。数据对象间距离的计算有很多种，k-means算法通常采用欧氏距离来计算数据对象间的距离。

(3) DBSCAN算法原理

DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。

这类密度聚类算法一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。通过将紧密相连的样本划为一类，这样就得到了一个聚类类别。通过将所有各组紧密相连的样本划为各个不同的类别，则我们就得到了最终的所有聚类类别结果。

**五、实验步骤**

需要描述清楚算法流程，包括加载数据、数据处理、创建模型，训练，预测，评估模型等。如果必须给出实现代码才能更好地说明问题时，也必须先有相关的文字叙述，然后才是代码，代码只是作为例证。

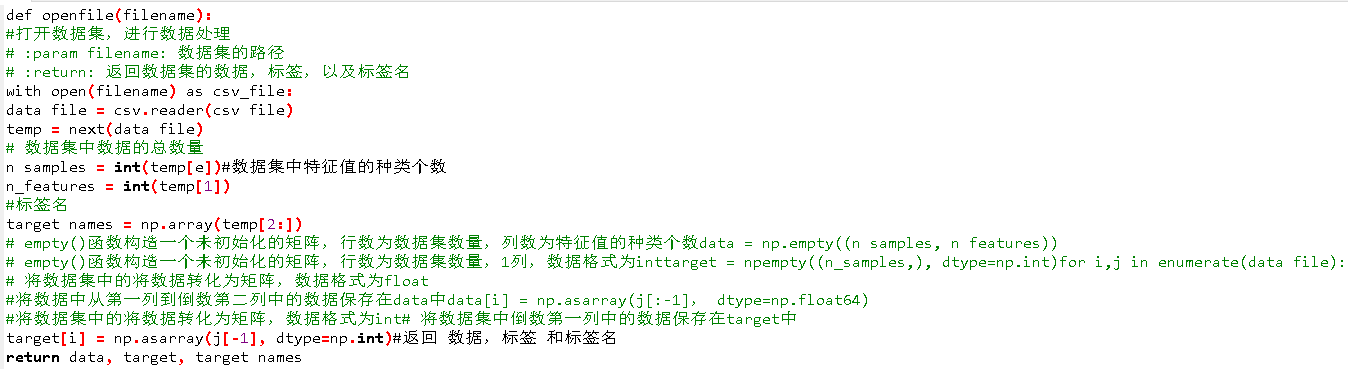
1. KNN

如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最近邻的一个或者多个样本的类别来决定待分类样本所属的类别。

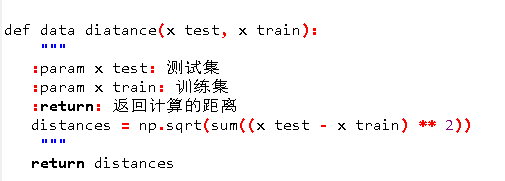
由于KNN算法在做类别决策时，只与极少量的相邻样本有关，因此采用该算法能够比较好的避免样本的不平衡问题。另外，由于KNN算法主要依靠周围有限的临近样本，而不是靠判别类域的方法来确定所属类别，因此对于类域的交叉或者重叠较多的待分类样本集来说，KNN算法较其他算法更合适。

**算法步骤**

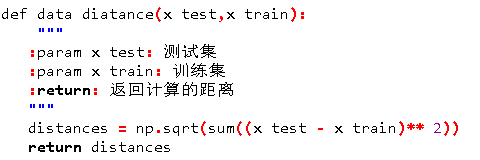
1.对训练数据进行处理，提出每一个样本数据和其对应的标签



2.对训练数据和测试数据进行归一化处理



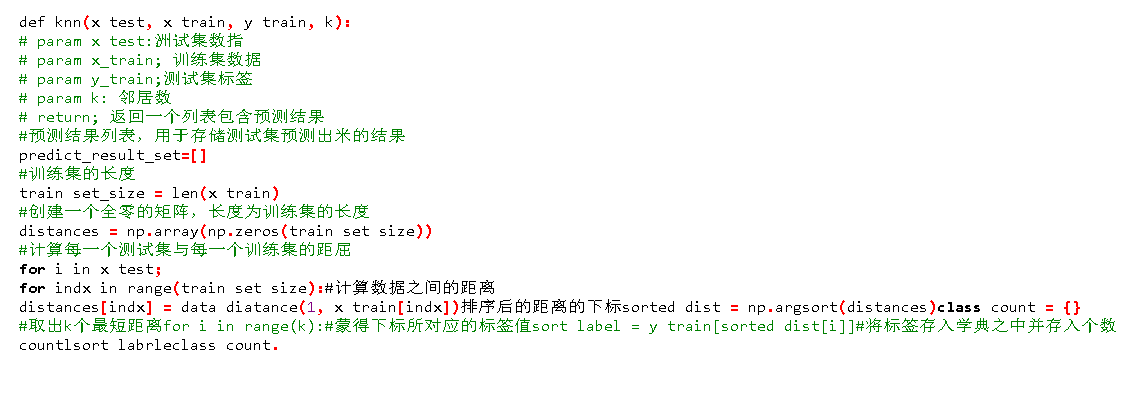
3.计算并存储每一个测试样本到所有训练样本的欧几里德距离

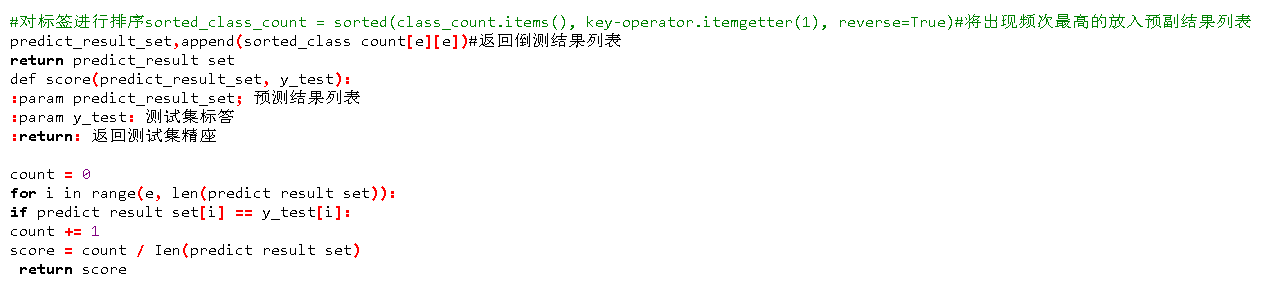


4.对得到的距离集合进行排序

5.选择距离最小的k个样本

6.找出k个样本中最公共的类别，即为该测试样本所属的类别



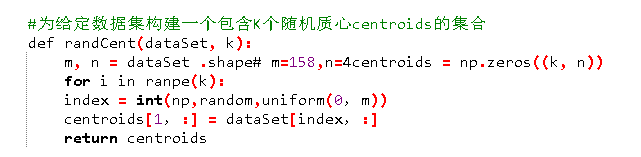


(2) K-means算法

[K-means](https://so.csdn.net/so/search?q=K-means&spm=1001.2101.3001.7020" \o "K-means)算法也称为K\_均值算法，用于聚类算法。聚类是一种无监督学习，他将相似的对象归于一个簇中，簇中心通过簇中所有点的均值来计算。聚类算法与分类算法的主要区别就是分类的目标类别已知，而聚类的目标类别未知。

**算法步骤**

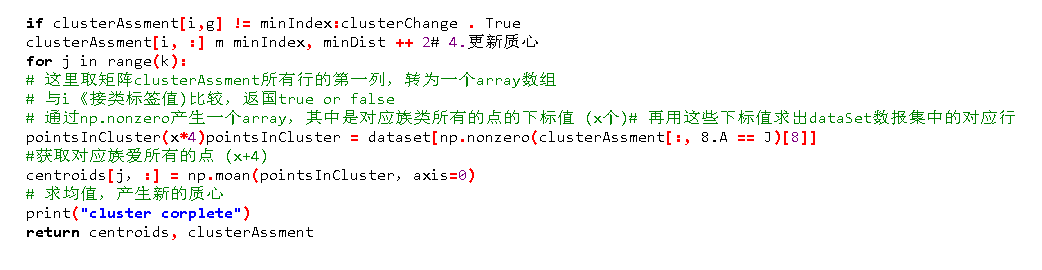
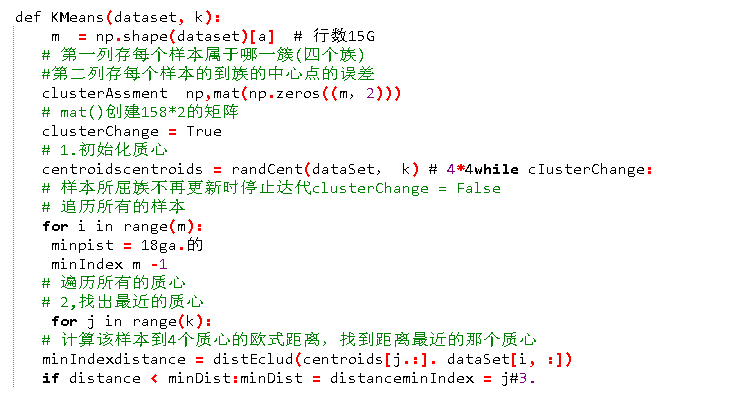
1.随机选取k个初始中心点；



2.针对数据集中的每个样本点，计算样本点与k个中心点的距离，将样本点划分到离它最近的中心点所对应的类别中；

3.类别划分完成后，重新确定类别的中心点，将类别中所有样本各特征的均值作为新的中心点对应特征的取值，即该类中所有样本的质心；

4.重复上面的2 3步骤，直到达到某个终止条件（迭代次数、每一个簇中的点不再变等）



(3) DBSCAN算法

DBSCAN是一种基于密度的空间聚类算法，是一种无监督的ML聚类算法。它可以替代[KMeans](https://so.csdn.net/so/search?q=KMeans&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/wxj_Mq/article/details/_blank)和层次聚类等流行的聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。

**算法步骤**

1.确定两个参数：

epsilon:在一个点周围邻近区域的半径

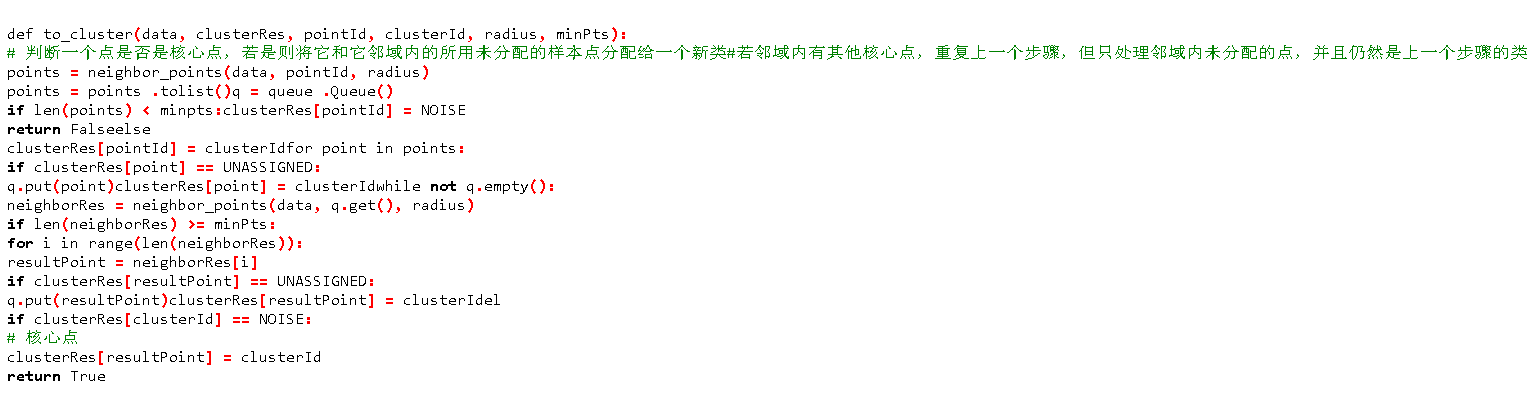
minPts:邻近区域内至少包含点的个数

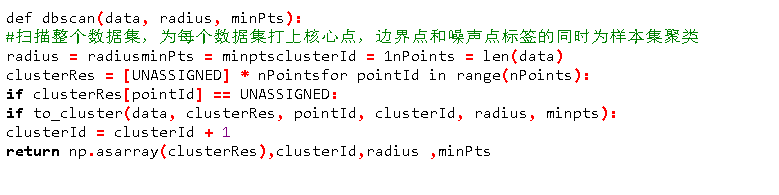
1. 任意选择一个点（既没有指定到一个类也没有特定为外围点），计算它的NBHD(p,epsilon)判断是否为核点。如果是，在该点周围建立一个类，否则，设定为外围点。

3.遍历其他点，直到建立一个类。把directly-reachable的点加入到类中，接着把density-reachable的点也加进来。如果标记为外围的点被加进来，修改状态为边缘点。

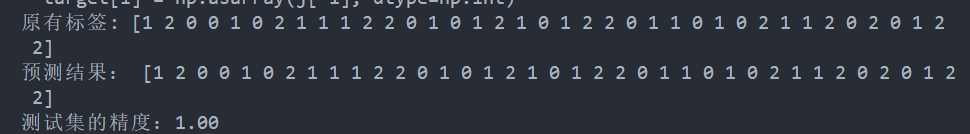
4.重复步骤1和2，直到所有的点满足在类中（核点或边缘点）或者为外围点。

5.利用轮廓函数对算法进行评估，得出最优参数

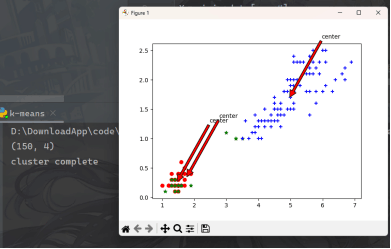
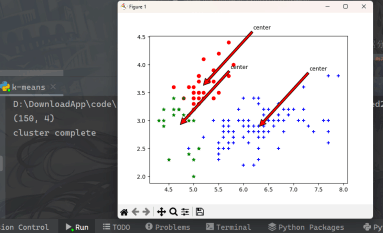




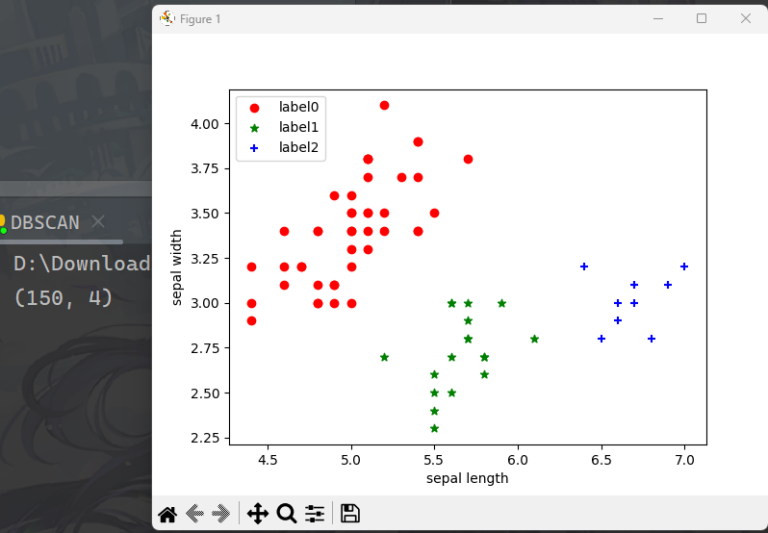
1. **实验数据及处理结果**

**(1) KNN**

**（2）K-means算法**

1. ****

**(3) DBSCAN算法**

****

1. **实验总结及心得**

这次机器学习实验对我来说是一次独立操作的学习体验。通过自己调整模型参数、分析数据、选择特征等实际操作，我更加深刻地理解了机器学习的实际应用过程。

在实验中，我成功地构建了有效的模型，这不仅让我对算法和模型有了更深入的认识，也增强了我解决实际问题的信心。同时，这种独立操作也锻炼了我的问题解决能力，让我学会在面对挑战时保持冷静，一步步找出解决方案。

整个过程中，我还不断调整和优化模型，这使我更加熟悉不同参数和方法的影响。这种实践让我更加深入地了解了机器学习领域的细节，并对理论知识有了更为具体的实际应用。

总的来说，这次独立完成的机器学习实验让我获益匪浅。通过动手操作，我不仅加深了对机器学习实际运用的理解，还培养了独立解决问题的能力。这对我个人的学术和职业发展都将产生积极的影响。