# 实验三：PCA降维分析和聚类实验

任务：

1. 掌握PCA降维分析和可视化数据方法，选做t-SNE
2. 掌握系统树图和K-means聚类， 进而熟悉GBDT方法；
3. 掌握Grab CUT等原理和应用。

## Ⅰ 方法原理介绍

* **主成分分析（PCA）**

Principle Component Analysis 顾名思义，是通过分析向量空间的主成分，将主成分提取出来，不重要的成分略去，从而达到降维压缩信息的目的。那什么才是主成分呢？大家应该知道，一个空间会有自己的一组基向量，空间中的任何一个向量都可以通过基向量的组合来表示。

具体的原理请大家参见吴恩达（Andrew Ng）的网页介绍：<http://deeplearning.stanford.edu/wiki/index.php/%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90> ，这里不赘述。

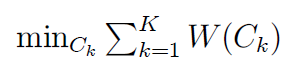
一般步骤是这样的：先对原始数据零均值化，然后求协方差矩阵，接着对协方差矩阵求特征向量和特征值，这些特征向量组成了新的特征空间。

* **聚类（Clustering）**

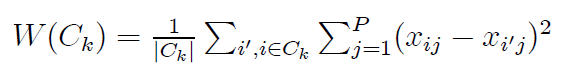
聚类分析是一种无监督学习方法，是一种在数据集中寻找子群或类的技术，应用非常广泛。本次实验主要涉及K 均值聚类(K-means clustering)和层次聚类（hierarchical clustering）。

**K 均值聚类**

进行K 均值聚类前，首先要确定要分类的数量。然后将每个观测值分配到K 个类中。分类的思想是最小化所有类的内间相异度：



其中*W (Ck)*是衡量内间相异度的指标，一般可以是平方欧式距离，



要精确的解决上面的问题(全局最优解)，对于感测数据量比较大时，计算量将是非常大的。而K 均值算法就是求解上述问题局部最优解的一种方法，

*  确定分类数K，为每个观测随机分配一个类标签；
*  重复，如果没有达到停止条件
  + 计算K 个类的类中心；
  + 将观测分配到距离最近的类中心所在的类。

由于K 均值聚类方法得到的解并非全局最优，所以随机初始会影响最终的收敛结果，解决方法一般是进行多次随机初始，在得到的结果中选取最优解。

K 均值算法简单有效，但也存在一些问题。如聚类数必须事先人为确定，分类初始化影响较大，对于某些数据集会失效等。对应这些也有提出一些对应的解决方法，可以查阅相关资料。

**层次聚类**

系统聚类相对与K 均值聚类的一个优点，就是不需要事先确定分类的数量，一旦系统树图(dendrogram)计算完成，可以得到任意类数(不大于观测数据量)的分类。



图 系统树图

在系统树图中，最下面的每个叶节点表示一个观测数据，叶节点根据相似度一次向上进行汇合，越早汇合表示越相近。因此纵轴上枝条初次汇合的高度表示两个观测的差异度，纵轴越大差异越大。和K 均值聚类一样，需要找到衡量差异度的指标，和K 均值聚类不同的是，在层次聚类中，还需要衡量集合和单个观测，集合与集合之间的差异度，常用的有3 种度量：最大距离、最小距离和平均距离，最大距离是选择两集合中差异度最大的作为集合间的差异度，相反，最小距离是去其中最小的差异度衡量集合的差异度，同样平局距离是对所有差异度去平均。差异度的衡量可用欧式距离。

**谱聚类（Spectral Clustering）**

谱聚类算法是一种基于图论的聚类算法。它建立在谱图理论的基础上，其本质是将聚类问题转化为图的最优划分问题。与传统的聚类算法相比，它具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点，而且可以获得在放松了的连续域中的全局最优解。

大家可以通过这个笔记进一步了解其应用和发展现状，条理比较清楚。 相关的数学原理需要借助图论相关知识。 https://www.zybuluo.com/hainingwyx/note/593818

## Ⅱ 实验任务

### 主成分分析（PCA）

#### 学习http://scikit-learn.org/stable/modules/decomposition.html#pca

1. 生成3簇（每簇200个数据）数据集，可参考下属代码

import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

#随机生成一个实数，范围在（0.5,1.5）之间

cluster1=np.random.uniform(0.5,1.5,(2,10))

cluster2=np.random.uniform(3.5,4.5,(2,10))

#hstack拼接操作

X=np.hstack((cluster1,cluster2)).T

plt.figure()

plt.axis([0,5,0,5])

plt.grid(True)

plt.plot(X[:,0],X[:,1],'k.')

1. 对上述生成数据，运用函数库中numpy的mean方法来求均值
2. 用numpy中的cov函数用于求协方差矩阵， 了解rowvar参数意义；
3. 调用numpy中的线性代数模块linalg中的eig函数，可以直接由covMat求得特征值和特征向量；
4. 得到了特征值向量eigVals，假设里面有m个特征值，我们可以对其排序，排在前面的n个特征值所对应的特征向量就是我们要保留的，它们组成了新的特征空间的一组基n\_eigVect。将零均值化后的数据乘以n\_eigVect就可以得到降维后的数据。
5. 写出对应代码。
6. **请参考：**

<http://scikit-learn.org/stable/auto_examples/applications/plot_face_recognition.html#sphx-glr-auto-examples-applications-plot-face-recognition-py>**，**

如果对特征值排序，选择前80%和98%对应特征向量，请重构出相应的主成分所构成的特征脸，请给出代码。 （有关函数的解读请参考

http://blog.csdn.net/u012162613/article/details/42192293 ）

### 聚类技术分析

#### Kmeans聚类

1. 了解kmeans的聚类过程：

<http://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html#sphx-glr-auto-examples-cluster-plot-mini-batch-kmeans-py>

1. 针对上一个实验的数据，写一段代码演示不同K值的影响，并可视化之。并说明合理K值是多少？为什么。

#### 层次聚类

1）、学习不同链接选项对层次聚类影响：

<http://scikit-learn.org/stable/auto_examples/cluster/plot_digits_linkage.html#sphx-glr-auto-examples-cluster-plot-digits-linkage-py>

2）、对于iris数据 （from sklearn import datasets），用AgglomerativeClustering算法聚成3类， 画出聚类结果并打印出各个簇的样本数目，最后给出confusion矩阵评估结果。

#### 谱聚类

1）、熟悉谱聚类的原理，

[http://scikit-learn.org/stable/auto\_examples/cluster/plot\_segmentation\_toy.html#](http://scikit-learn.org/stable/auto_examples/cluster/plot_segmentation_toy.html)

2）、实现并解读代码

<https://scikit-learn.org/stable/auto_examples/cluster/plot_coin_segmentation.html#sphx-glr-auto-examples-cluster-plot-coin-segmentation-py>