**实验：在聚类中使用PCA降维**

[**数据集**[**：segmentation\_data.zip]**](http://facweb.cs.depaul.edu/mobasher/classes/CSC478/Data/segmentation_data.zip)

对于此问题，您将使用**图像分割数据集**进行聚类。你将进行实验使用PCA 方法为降低数据维度和消除数据的噪声。你将使用被提供的图片种类任务作为基本事实去比较使用PCA降维算法前后的聚类的结果。数据集被分为三个文件。文件"segmentation\_data.txt"包含有关图像的数据，每行对应于一个图片。每张图片由 19 个特征表示（这些特征是数据集中的列，对应于文件"segmentation\_names.txt"中的特征名称）。文件"segmentation\_classes.txt"包含数据集文件中每个相对应图片的类标签（图片类型）和分类标识号。

您在这个问题上的任务如下：

1. 加载图片的数据矩阵（行作为图片样本，列作为特征）。并且从分类的文件中加载分类标识号 （允许你将 后缀为txt 文件更改为后缀为 csv 文件）。使用你最喜欢的方法（例如[**，sklearn 模块的min-maxScaler**](http://scikit-learn.org/stable/modules/preprocessing.html#scaling-features-to-a-range)），在数据矩阵上使用min-maxScaler方法，以便将每个特征的数值被缩放到 [0，1] 的区间范围内。
2. 使用 sklearn中的Kmeans 实现**，**在图像数据上使用聚类（ K的取值范围为 （1，9），绘制肘部图形以选择最佳质心数 ，以便随后我们可以将集群与预先分配的图像类进行比较）。
3. 在规范化后图像数据矩阵上使用你最喜欢的分类器。打印输出预测准确度和分类报告。
4. 使用 卡方选择前 10 个特征，然后再次使用分类器。将您的结果与 第c题未经过降维处理的数据获得的结果进行比较。
5. 在规范化图像数据矩阵上使用 PCA降维算法。您可以使用Numpy中线性代数的相关模块，也可以使用sklearn中降维的模块（后者效率更高）。分析主成分确定这个数值，r, 使得降维后的数据的可解释性方差的贡献率为95%。画出可解释性方差的贡献率的曲线图。然后使用这些 **r** 数量的成分作为特征，将数据转换到低维度空间。
6. 再次执行 Kmeans，但这次是在降维过后的数据集上。将您的结果与 第b题 未经过降维处理的数据获得的结果进行比较。
7. 再次使用分类器，但这次是在降维过后的数据集上。将这些结果与第 c 题和 第d题中获得的结果进行比较。