

法律声明

□ 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



聚类实践



小象学院
ChinaHadoop.cn

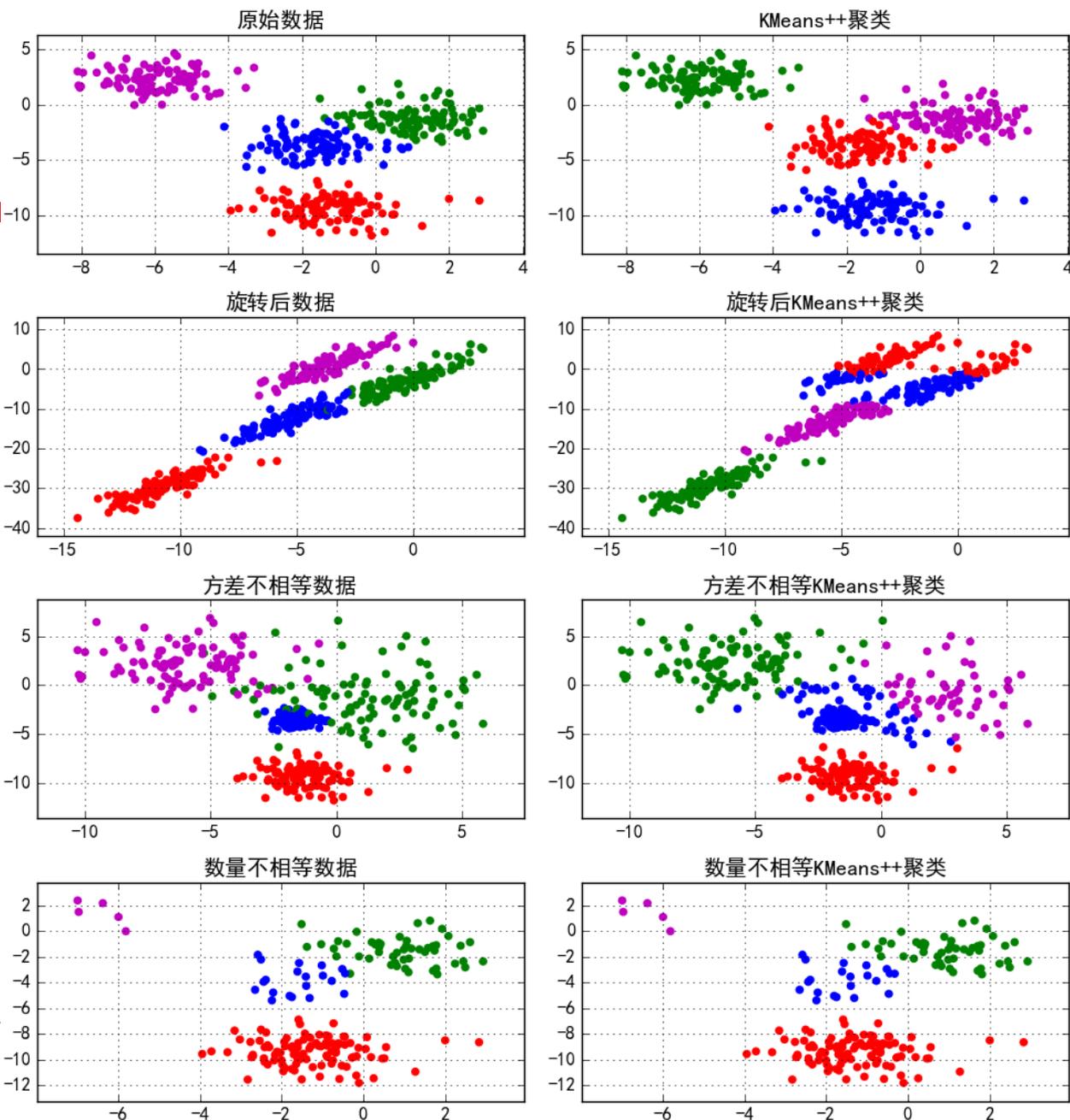
邹博

本次目标

- k-Means 算法及其适用范围
- VQ 算法及其应用
 - 图像压缩
- AP 算法
- MeanShift 算法
- DBSCAN 算法
- 谱聚类 SC

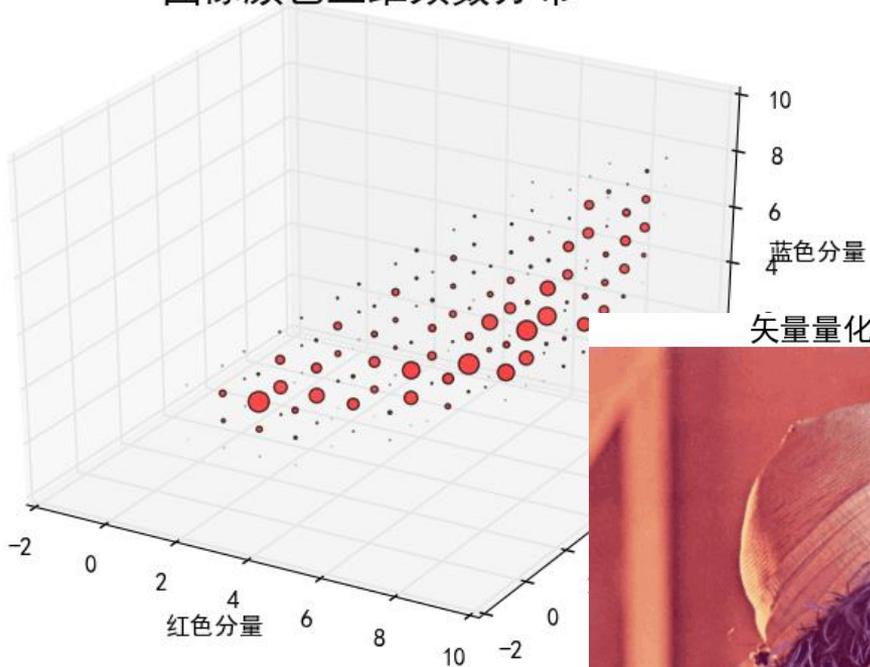
k-Means

数据分布对KMeans聚类的影响

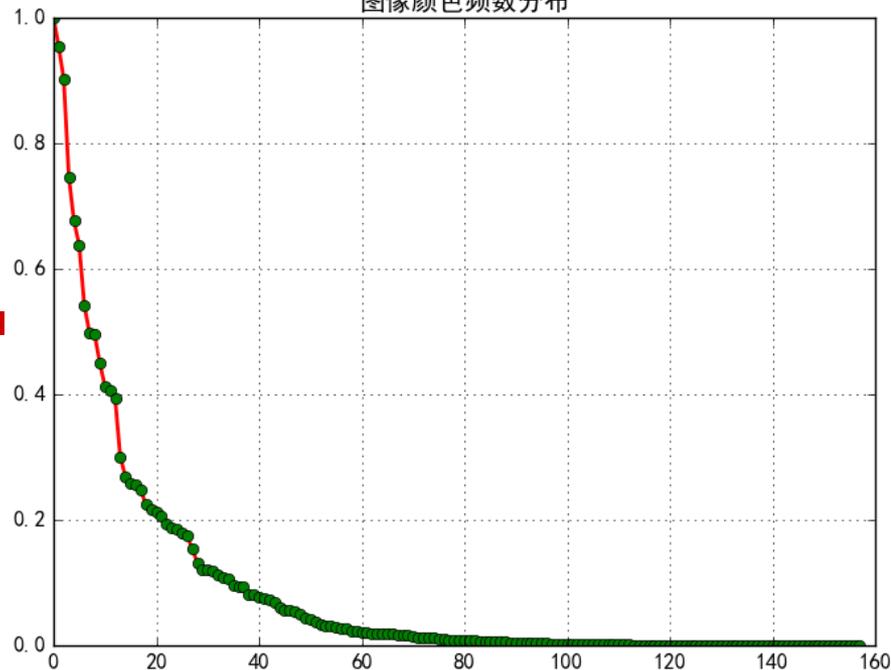


Vector Quantization

图像颜色三维频数分布



图像颜色频数分布



矢量量化后图片：100色

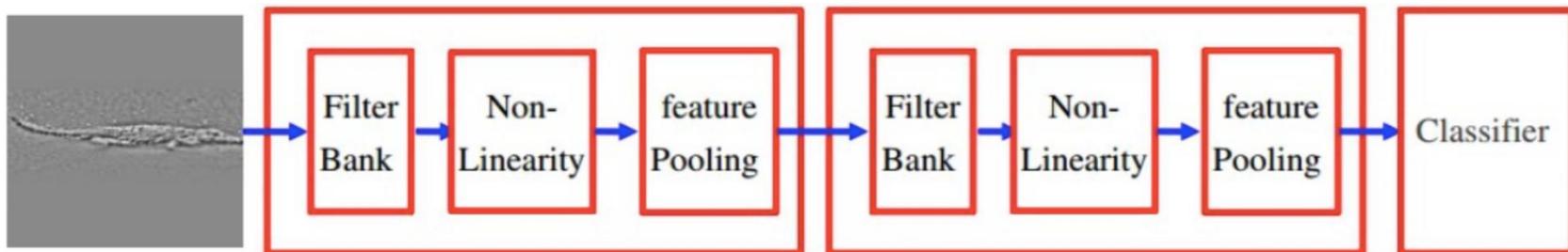
原始图片



传统模型

- Fixed Features + unsupervised mid-level features + simple classifier
 - SIFT + Vector Quantization + Pyramid pooling + SVM
 - [Lazebnik et al. CVPR 2006]
 - SIFT + Local Sparse Coding Macrofeatures + Pyramid pooling + SVM
 - [Boureau et al. ICCV 2011]
 - SIFT + Fisher Vectors + Deformable Parts Pooling + SVM
 - [Perronin et al. 2012]

传统图像分类



Oriented Edges **Winner Takes All** **Histogram (sum)** **K-means** **Pyramid Histogram.** **SVM or Another**
Or **Sparse Coding** **Elastic parts Models,...** **Simple classifier**

SIFT

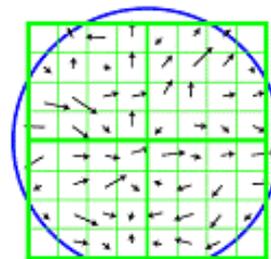
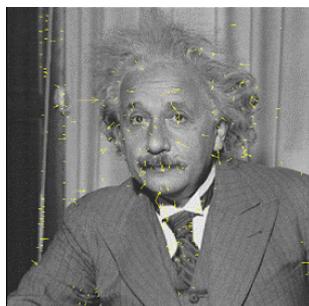
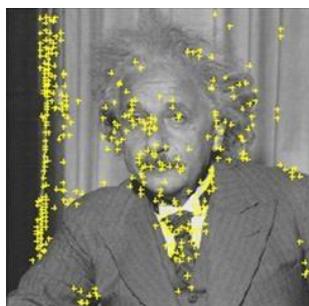
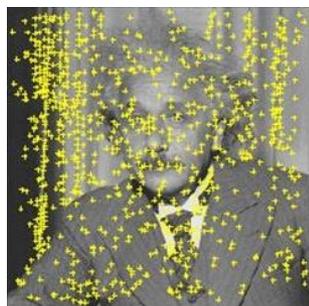
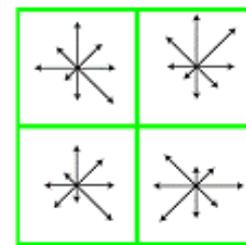


Image gradients



Keypoint descriptor

词袋模型bag of words



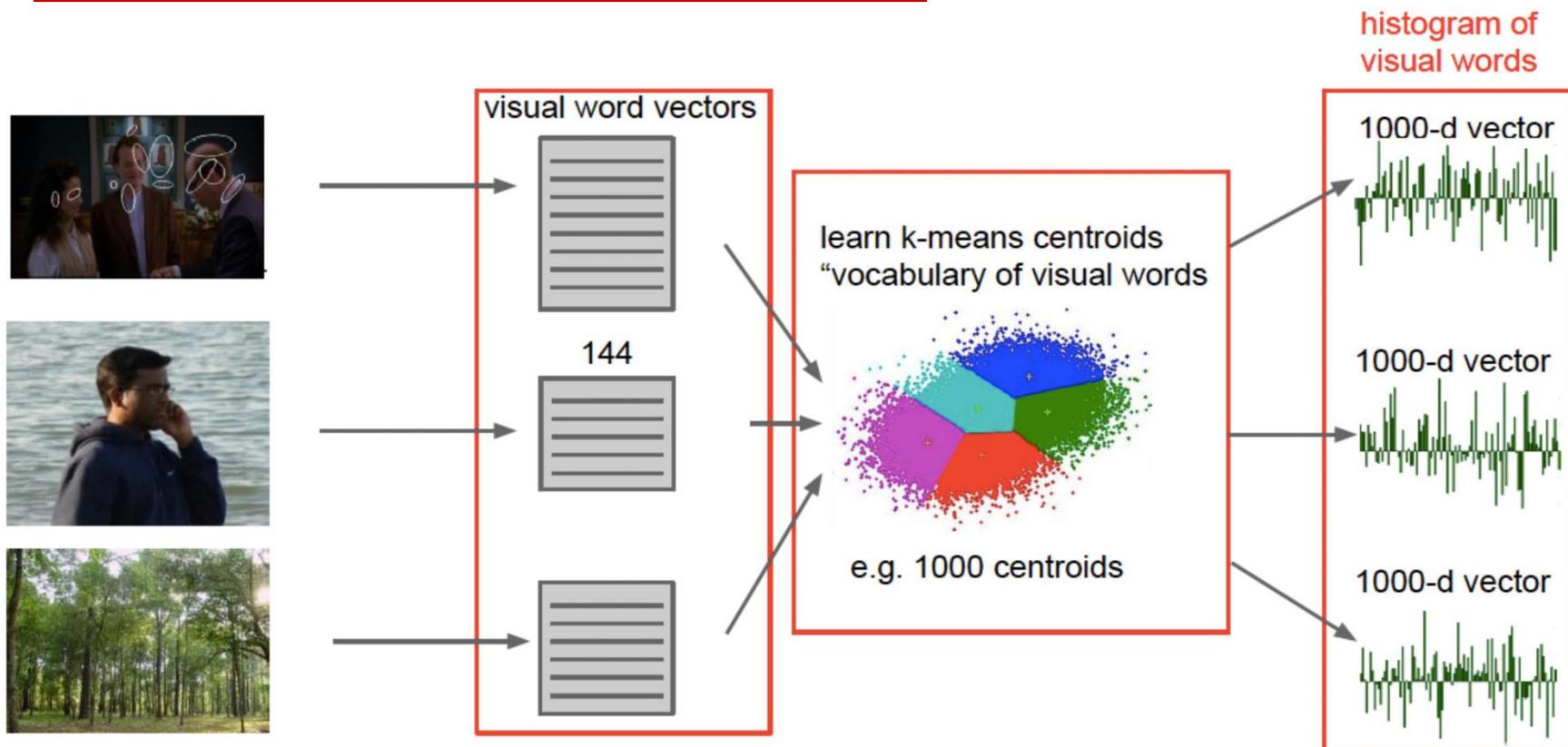
1. Resize patch to a fixed size (e.g. 32x32 pixels)
2. Extract HOG on the patch (get 144 numbers)

repeat for each detected feature

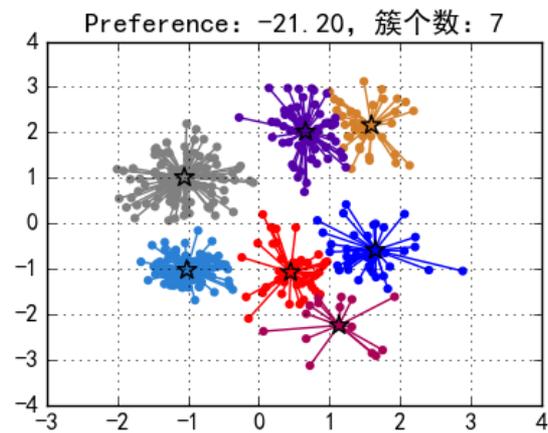
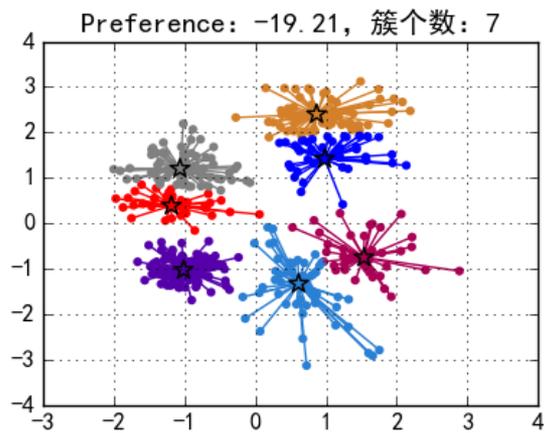
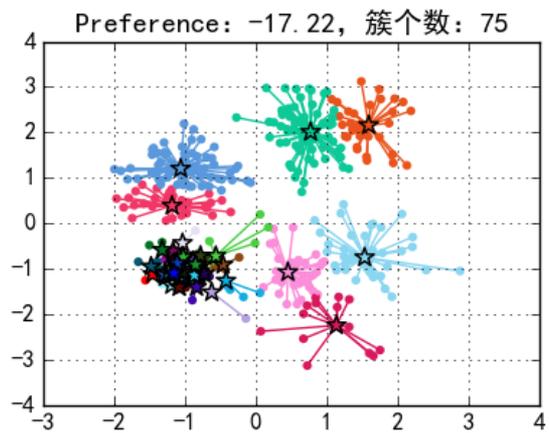
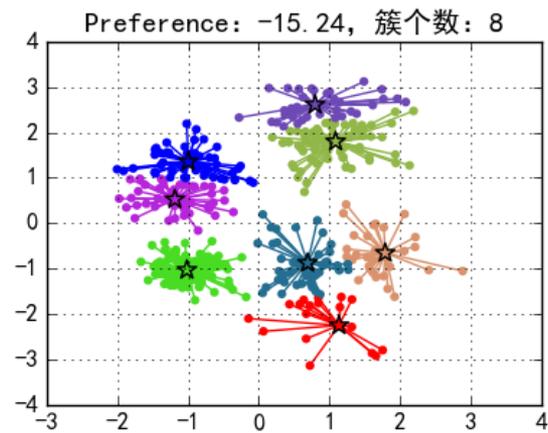
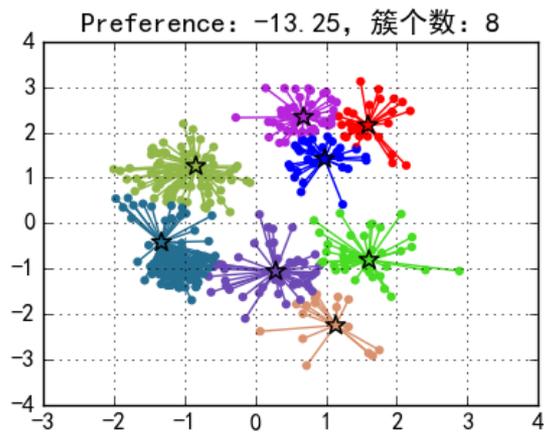
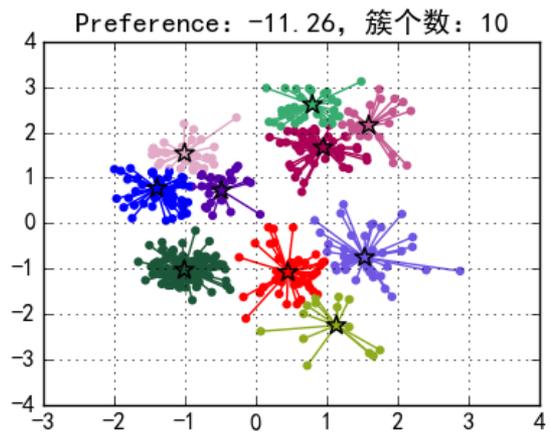
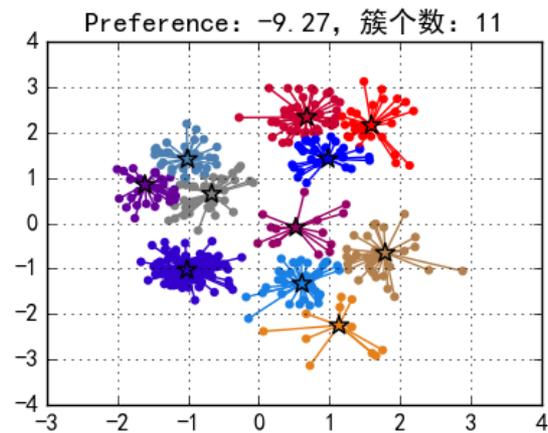
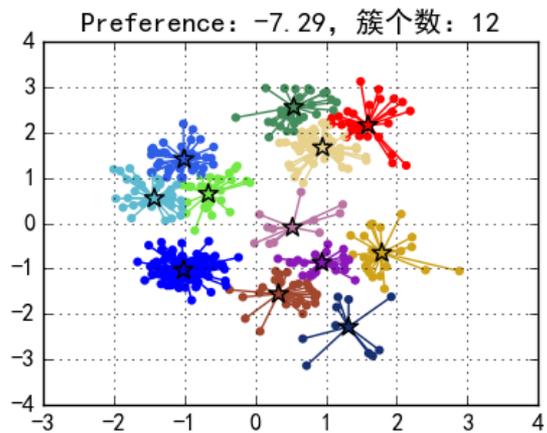
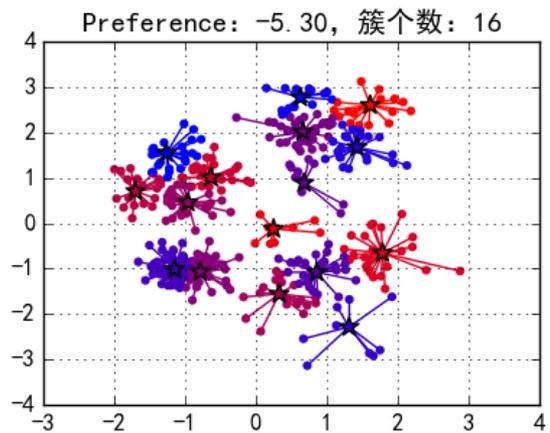


gives a matrix of size
[number_of_features x 144]

词袋模型 bag of words

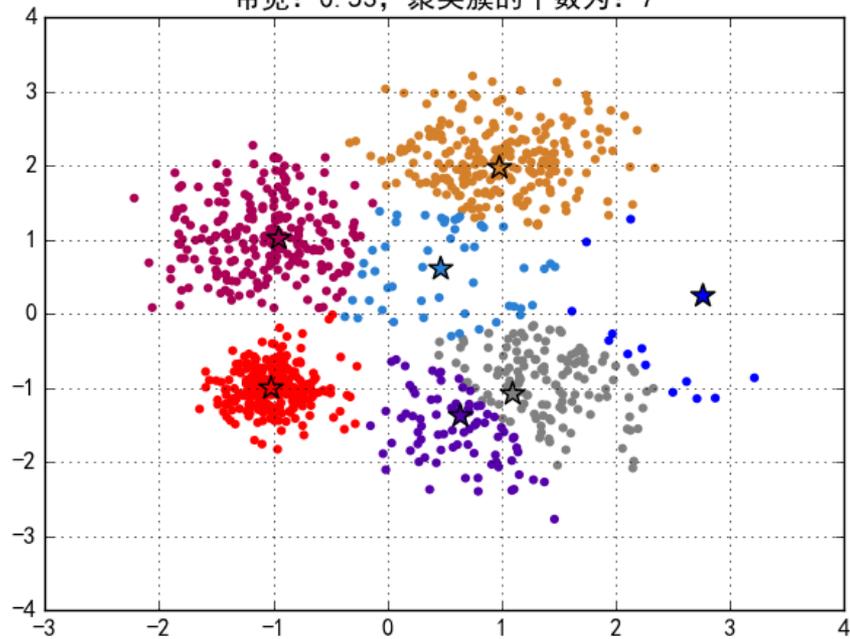


AP聚类

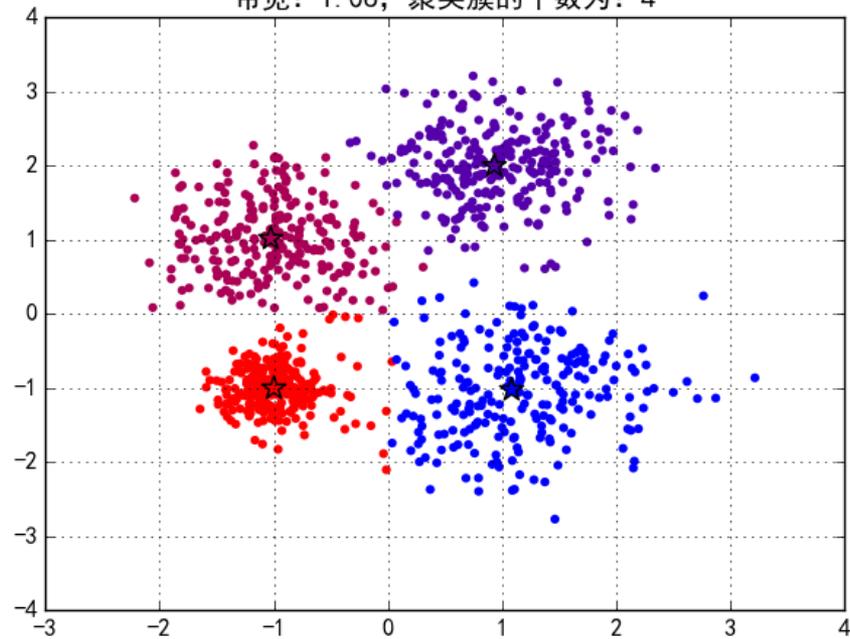


MeanShift聚类

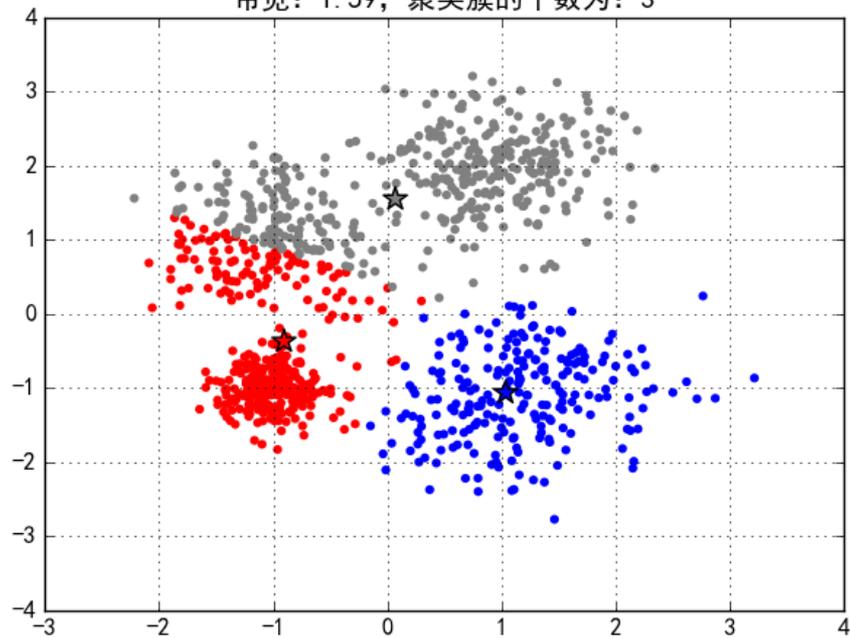
带宽: 0.53, 聚类簇的个数为: 7



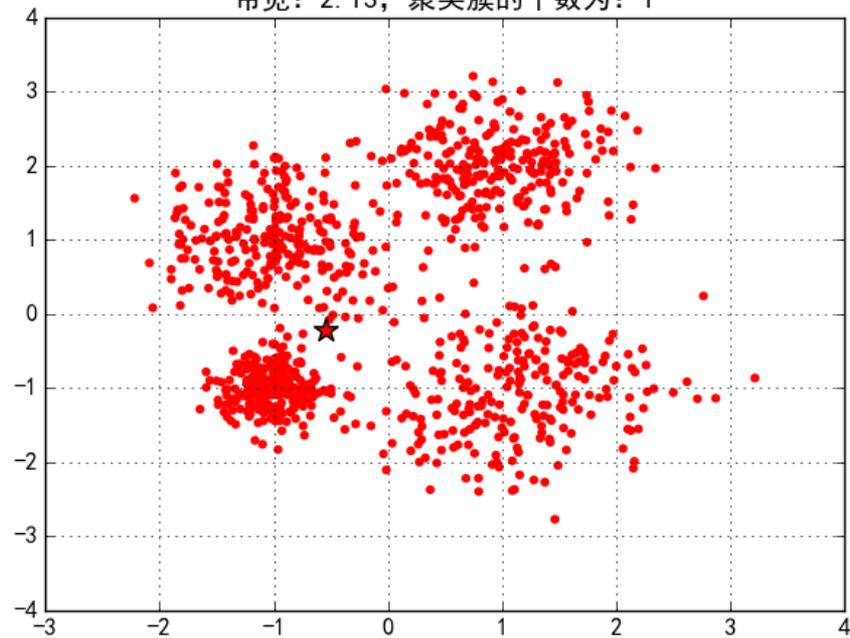
带宽: 1.06, 聚类簇的个数为: 4



带宽: 1.59, 聚类簇的个数为: 3

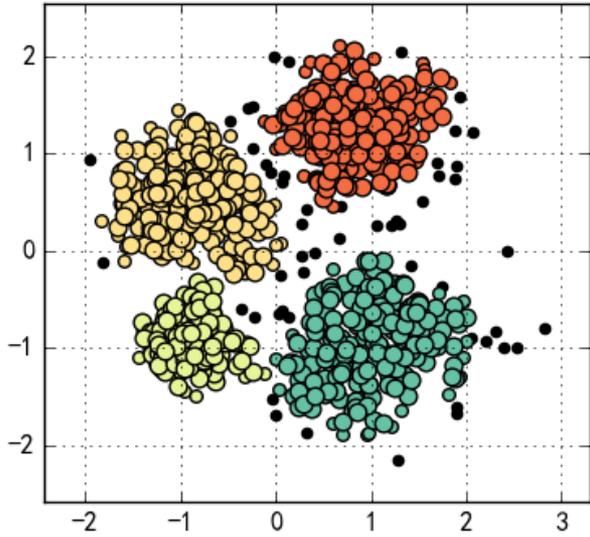


带宽: 2.13, 聚类簇的个数为: 1

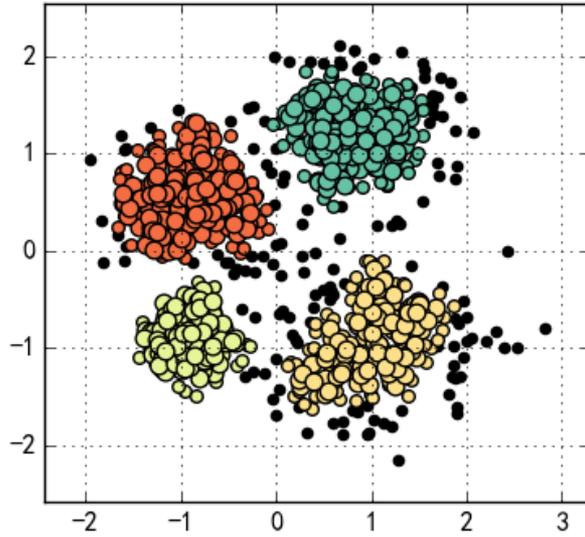


DBSCAN聚类

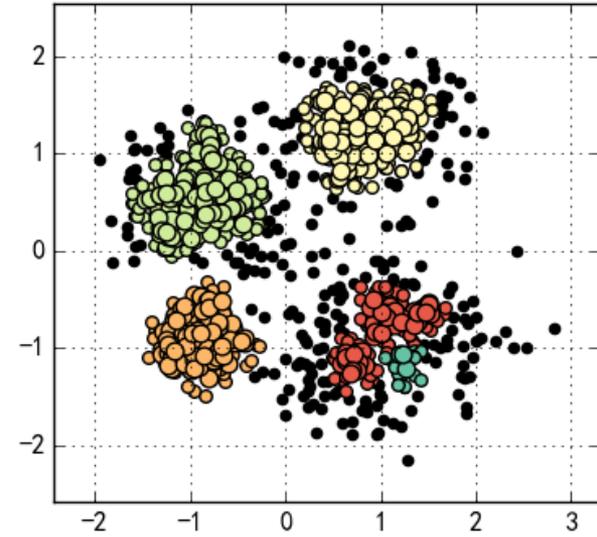
$\epsilon = 0.2$ $m = 5$, 聚类数目: 4



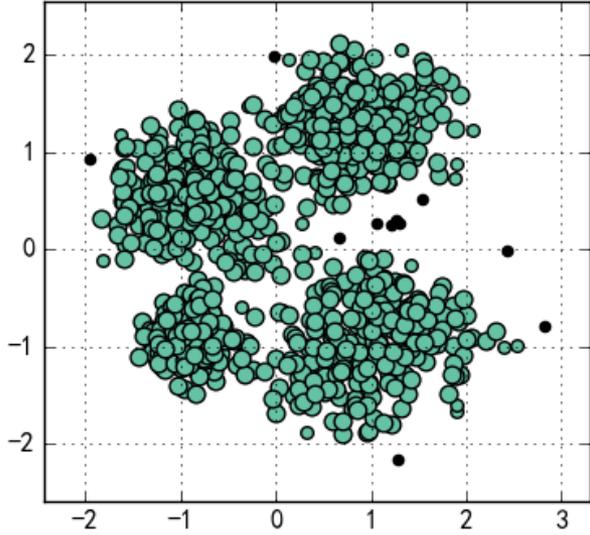
$\epsilon = 0.2$ $m = 10$, 聚类数目: 4



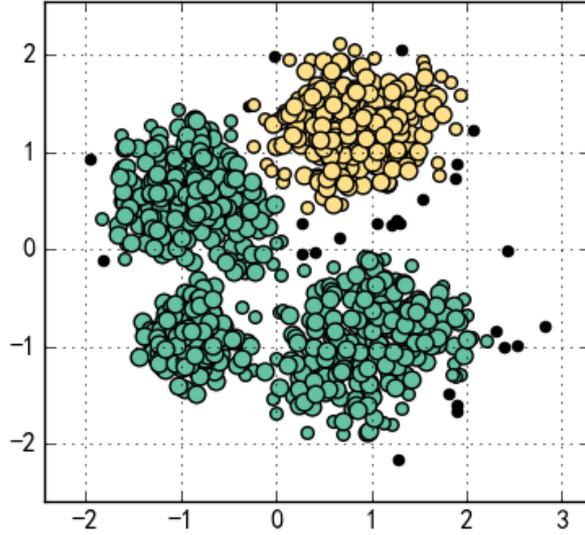
$\epsilon = 0.2$ $m = 15$, 聚类数目: 5



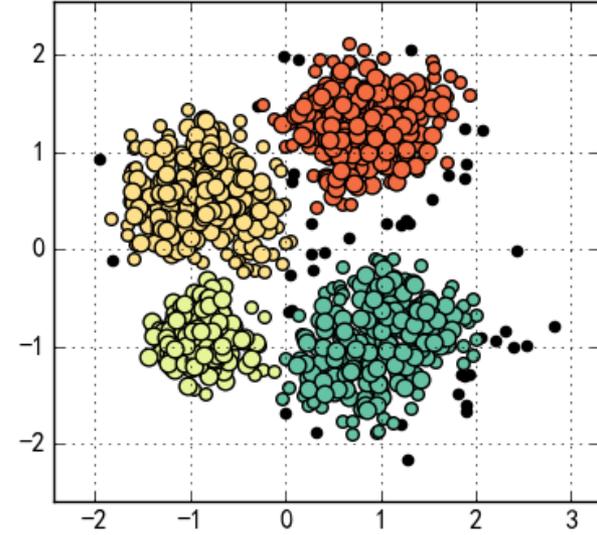
$\epsilon = 0.3$ $m = 5$, 聚类数目: 1



$\epsilon = 0.3$ $m = 10$, 聚类数目: 2

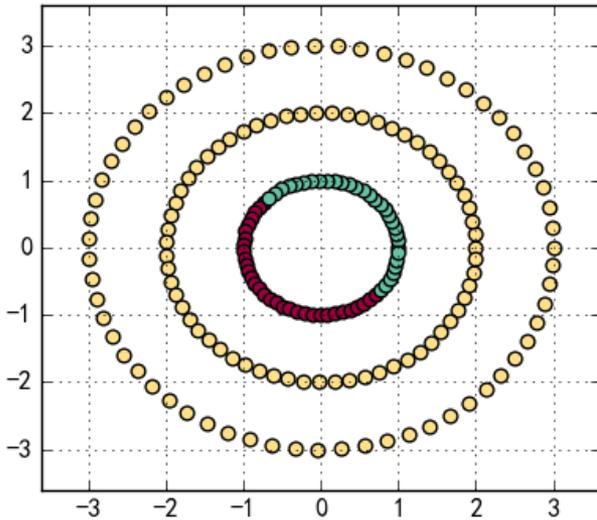


$\epsilon = 0.3$ $m = 15$, 聚类数目: 4

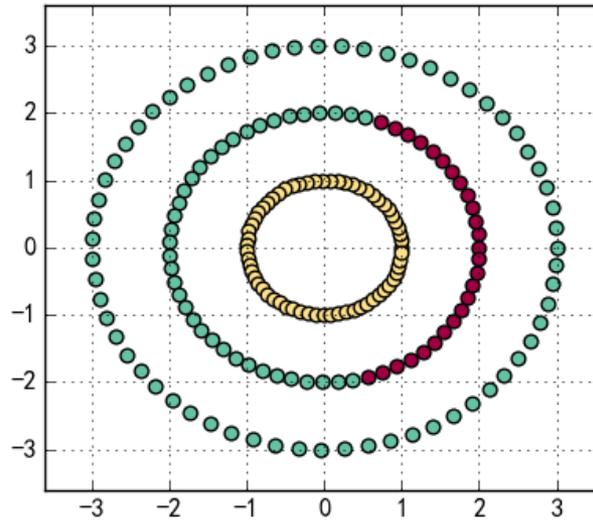


谱聚类

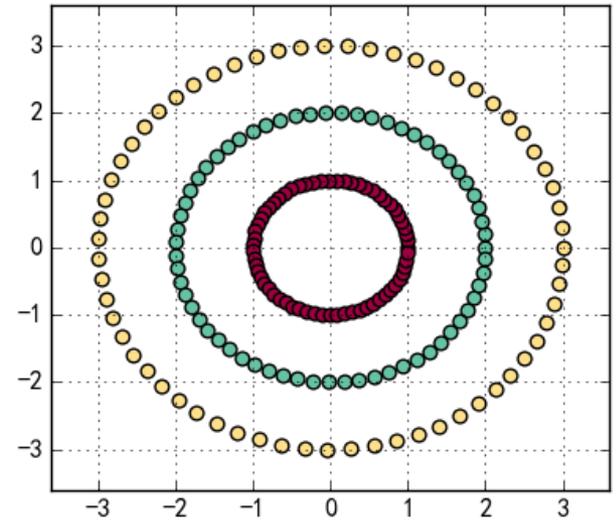
$\sigma = 0.01$



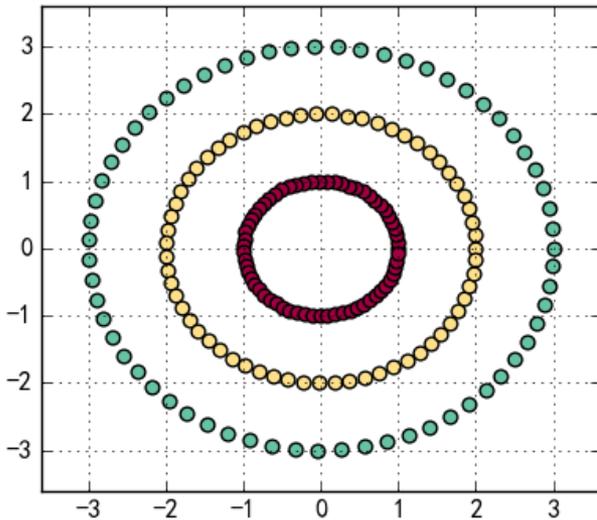
$\sigma = 0.03$



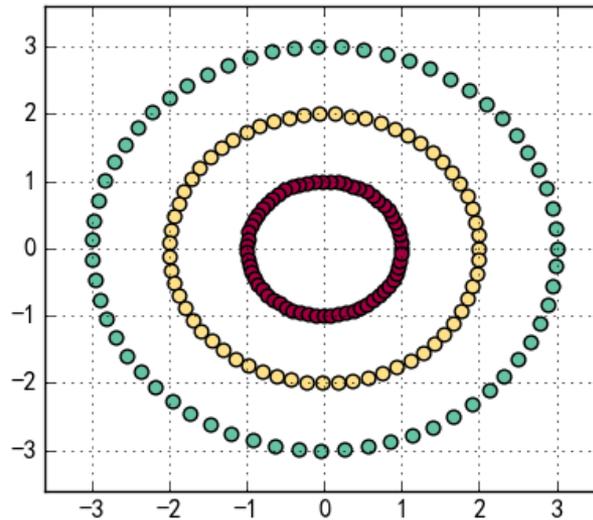
$\sigma = 0.06$



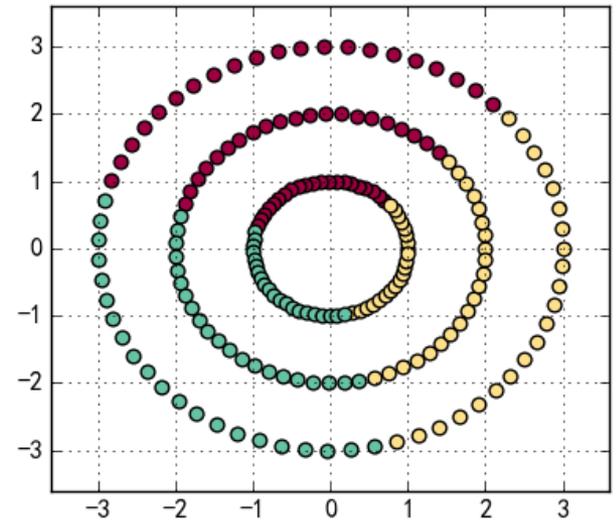
$\sigma = 0.16$



$\sigma = 0.40$

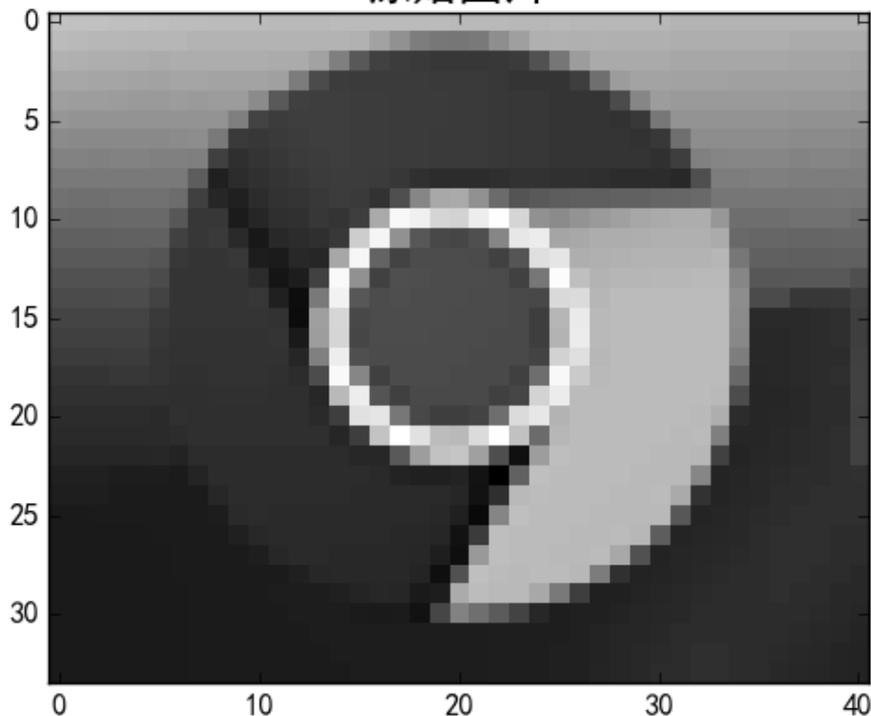


$\sigma = 1.00$

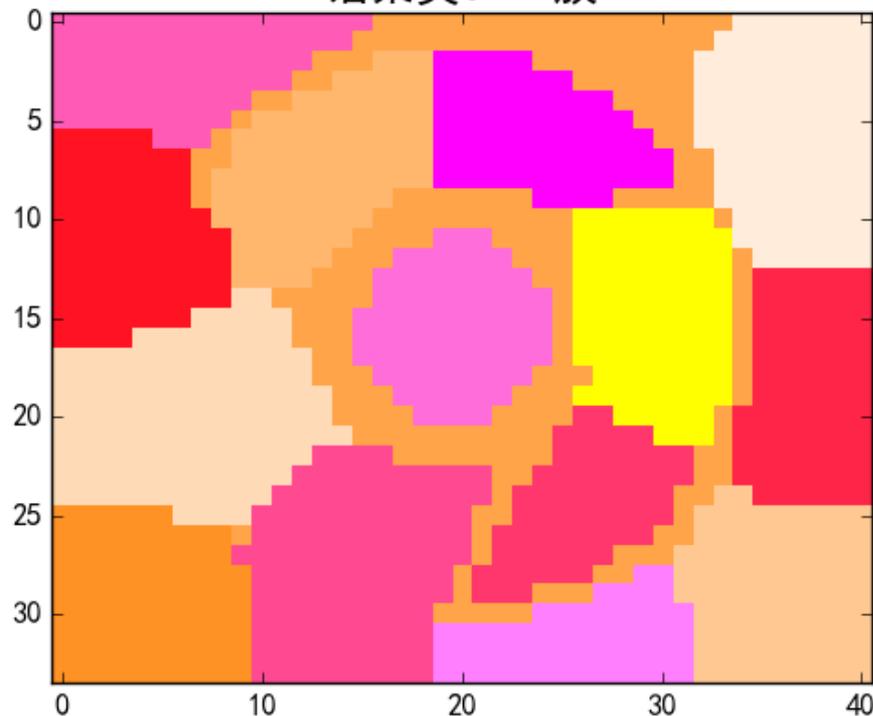


谱聚类与图像切割

原始图片



谱聚类：15簇





graphviz has no attribute 'write'

邹博 回复了问题 · 2 人关注 · 1 个回复 · 3 次浏览 · 2017-04-09 15:48

贡献



sklearn中如何理解Pipeline机制

数据分析与数据挖掘 邹博 回复了问题 · 2 人关注 · 1 个回复 · 28 次浏览 · 2017-04-09 15:39

贡献



关于9.Logistic回归的ppt中第9页的对数线性函数

机器学习 邹博 回复了问题 · 3 人关注 · 3 个回复 · 39 次浏览 · 2017-04-09 15:35

贡献



关于“贝叶斯估计中，最大后验概率估计就是结构化风险最小化的例子：当模型是条件概率分布，损失函数为对数损失函数，模型的复杂度由模型的先验概率表示，结构化风险最小化就等价于最大后验概率估计”

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 26 次浏览 · 2017-04-09 15:27

贡献



关于连续值的预测

咨询 邹博 回复了问题 · 2 人关注 · 1 个回复 · 31 次浏览 · 2017-04-09 15:24

贡献



拉格朗日对偶函数为什么一定是凸函数

数据科学 邹博 回复了问题 · 2 人关注 · 2 个回复 · 26 次浏览 · 2017-04-09 15:20

贡献



梯度下降公式中的斯堪J是

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 29 次浏览 · 2017-04-09 15:17

贡献



深度学习适合做预测吗？

深度学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 27 次浏览 · 2017-04-09 15:15

贡献



关于6.4PCA_FeatureSelection.py中plt.legend的参数疑问

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 28 次浏览 · 2017-04-09 15:04

贡献



@邹博 有哪些可以下载数据源的网站？

数据分析与数据挖掘 邹博 回复了问题 · 4 人关注 · 1 个回复 · 31 次浏览 · 2017-04-09 14:53

贡献



LDA主题模型

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 29 次浏览 · 2017-04-09 14:45

贡献



代码10.6bagging_ridged老师提到了采样率设为0.2能够使峰值部分的数据被体现出来。这是为什么呢？

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 22 次浏览 · 2017-04-09 14:26

贡献



GraphViz's executables not found

机器学习 邹博 回复了问题 · 3 人关注 · 2 个回复 · 23 次浏览 · 2017-04-09 13:47

贡献



决策树中关于feature_importances代码的问题

机器学习 邹博 回复了问题 · 2 人关注 · 1 个回复 · 6 次浏览 · 2017-04-09 13:11

贡献



机器学习

907 个问题, 230 人关注



spark

387 个问题, 172 人关注



hadoop

1059 个问题, 155 人关注



python数据分析

171 个问题, 28 人关注



数据分析与数据挖掘

54 个问题, 111 人关注



小心巴

14 个问题, 0 次赞同



又又V

45 个问题, 22 次赞同



铁甲无声

10 个问题, 0 次赞同



带刀锦衣卫

13 个问题, 0 次赞同

我们在这里

□ <http://wenda.ChinaHadoop.cn>

■ 视频/课程/社区

□ 微博

■ @ChinaHadoop

■ @邹博_机器学习

□ 微信公众号

■ 小象学院

■ 大数据分析挖掘

感谢大家!

恳请大家批评指正!