聚类

Yasaka 陈博

有监督学习

- X，Y
- Model
- xi
- y

无监督学习

- 有X
- 没有Y
- 利用X相似性
- 聚类
- 对大量未标注的数据集，按内在相似性划分为多个类别，类别内相似度大，类之间的相似度小

聚类

- X，很多特征
- 簇
- K个簇
- one-hot编码，假设有6个簇
- X m*n维 变成 X m*6维度
- 降维

相似度

二维空间的公式

0ρ = sqrt((x1-x2)^2+(y1-y2)^2)　|**x**| = √(x2 + y2)

三维空间的公式

0ρ = √((x1-x2)^2+(y1-y2)^2+(z1-z2)^2)　|**x**| = √(x2 + y2 + z2)

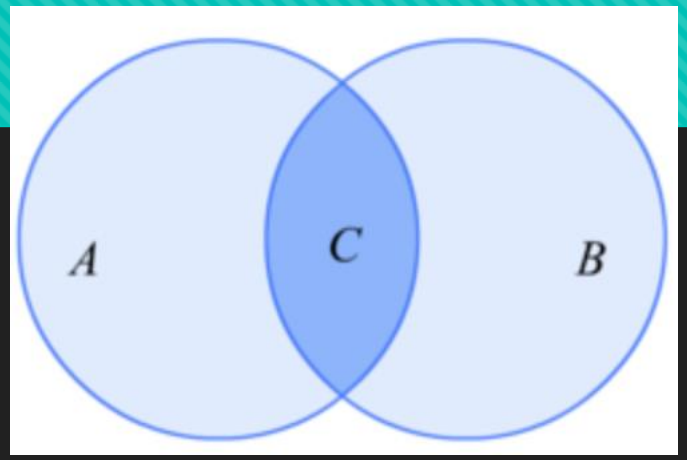
定义式：ρ(**A**，**B**) = [∑(a[i] - b[i])^p]^(1/p) (i = 1，2，…，n)

- 多维空间的向量点之间的距离
- 欧式距离
- 闵可夫斯基距离公式中，当p=2时，即为欧氏距离；当p=1时，即为曼哈顿距离；Block Distance
- 当p→∞时，即为切比雪夫距离，就是哪个维度差值大就是哪个作为距离
- Jaccard相似系数
- （Jaccard similarity coefficient）用于比较有限样本集之间的相似性与差异性
- Jaccard系数值越大，样本相似度越高

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

$$d_j(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}=\frac{A\Delta B}{|A\cup B|}$$

推荐系统



- 用户实际喜欢 [8, 9, 17, 25, 4]
- 给用户的推荐[9, 10, 17, 24, 4, 8]
- [8, 9, 25]
- 哪两个推荐结果更好呢？
- A集合和B集合相交的越多，它的相似性越强，当然要考虑它们并一起的大小，因为集合越大越可能相交的越多，这就有了Jaccard相似系数
- 度量集合，考虑热门商品
- 空间嵌入点的问题，有时候用欧式，有时候用余弦距离，度量文档相似性

相似度

- Jaccard系数
- 网页去重
- 防考试作弊
- 论文抄袭检查

准确率/召回率：

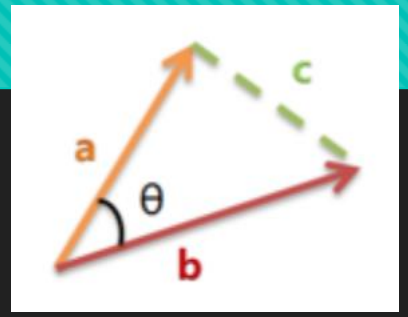
$$\text{Precision}(u) = \frac{R(u) \cap T(u)}{R(u)} \qquad \text{Recall}(u) = \frac{R(u) \cap T(u)}{T(u)}$$

Jaccard 系数：

$$Jaccard(u) = \frac{R(u) \cap T(u)}{R(u) \cup T(u)}$$

$$F_1 = \frac{2 * PR}{P + R}$$

相似度



$$\cos\theta = \frac{a \bullet b}{\|a\|\|b\|}$$

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$\cos\theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}}$$

- 余弦距离，余弦相似度
- 余弦值的范围在[-1,1]之间，值越趋近于1，代表两个向量的方向越接近；
- 越趋近于-1，他们的方向越相反；接近于0，表示两个向量近乎于正交。
- 最常见的应用就是计算文本相似度。将两个文本根据他们词，建立两个向量，
- 计算这两个向量的余弦值，就可以知道两个文本在统计学方法中他们的相似度情况。
- 文档相似度测量SimHash
- 考虑推荐，余弦其实是Jaccard的分母，看重得是相同的部分，如果是欧式距离，看重得是差异

相似度

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}.$$

- Person相关系数
- 两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商
- 相对熵
- P和Q相同，相对熵为0

$$D_{KL}(P\|Q) = -\sum_{x\in X} P(x) log \frac{1}{P(x)} + \sum_{x\in X} P(x) log \frac{1}{Q(x)} = \sum_{x\in X} P(x) log \frac{P(x)}{Q(x)}$$

相似度

- Person
- [3, 7, 9, 2, 4, 14, 9]
- [4, 7, 8, 9, 5, 17, 19]
- 相似相关性

相似度

- 余弦距离和Person的关系
- 相当于，X和Y它们的均值都是0，去做
- 余弦距离相当于把X，Y坐标向量各自平移到原点后的夹角余弦
- 回顾文档间使用夹角余弦，表达了文档去均值后的随机向量间的相关系数
- 人脸图片之前都是人工提取特征，现在是多层卷积层，神经网络深度学习来做

相似度

- 对于高维空间点之间度量
- 对于集合度量
- 对于自然语言处理度量
- 对于函数度量

聚类

- 本质上，N个样本，映射到K个簇中
- 每个簇至少有一个样本
- 一个样本只属于一个簇
- 最基本：
- 先给定一个初始划分，迭代改变样本和簇的隶属关系，每次都比前一次好

K-Means

- 选择K个初始的簇中心，随机的，先验知识给的，拍脑袋给的
- 某一个样本和某一个聚类中心的距离
- 计算所属聚类的样本均值
- 举例子

$$label_i = \arg\min_{1 \le j \le k} \|x_i - \mu_j\|$$

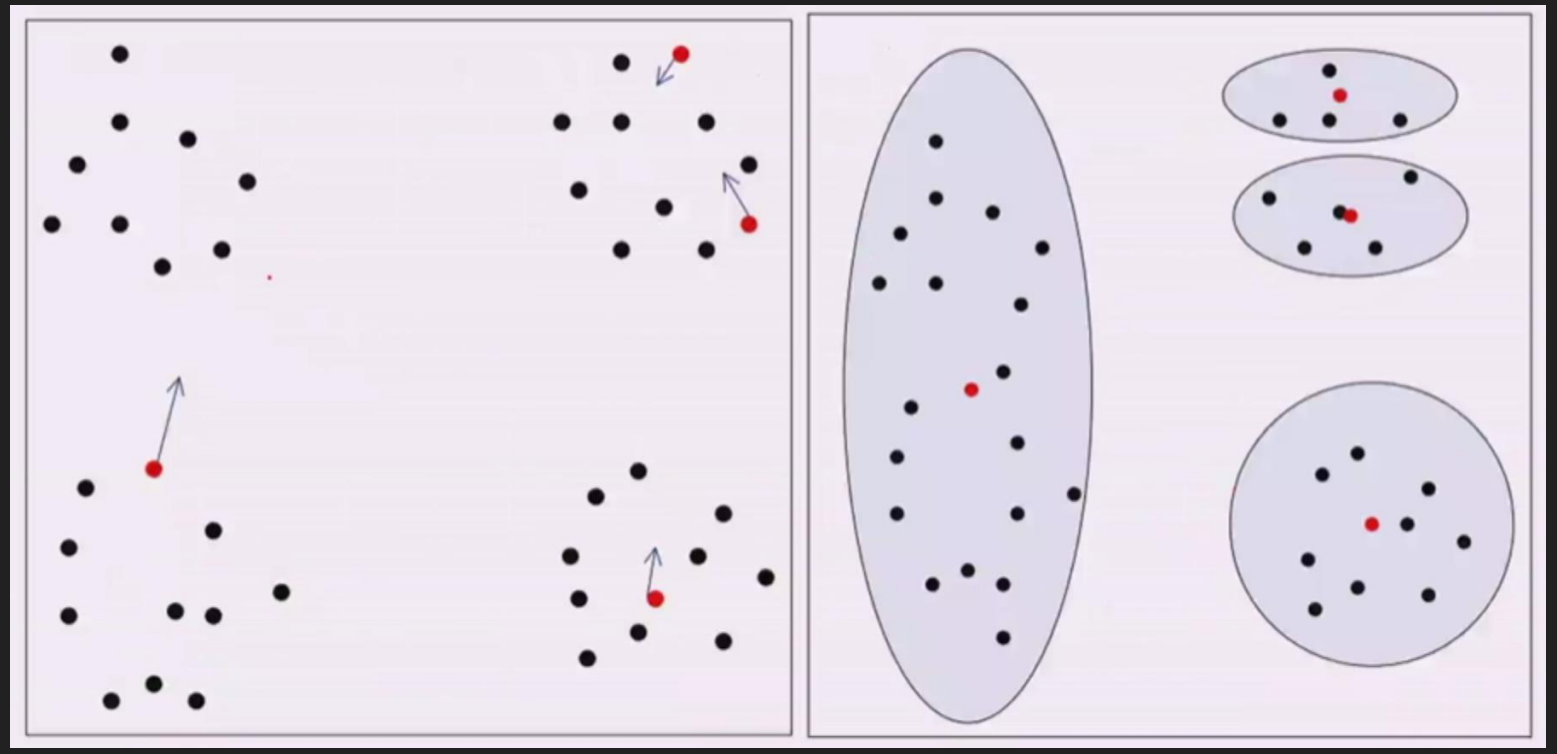
$$\mu_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i$$

K-Mediods

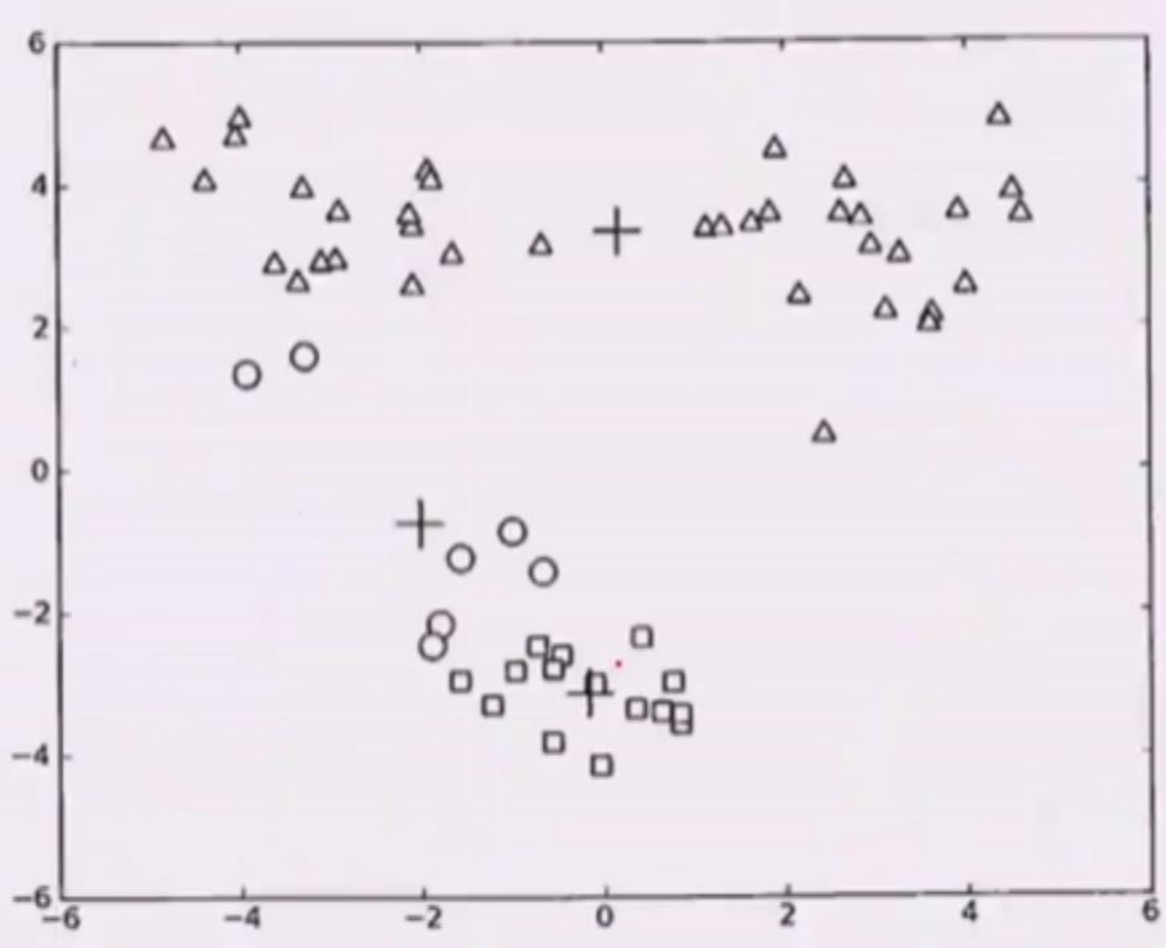
- 数组1，2，3，4，100的均值为22，其实求均值的话，离里面大多数的值还是比较远的
- 取中位数的话是3，更好一些，因为100可能是噪声

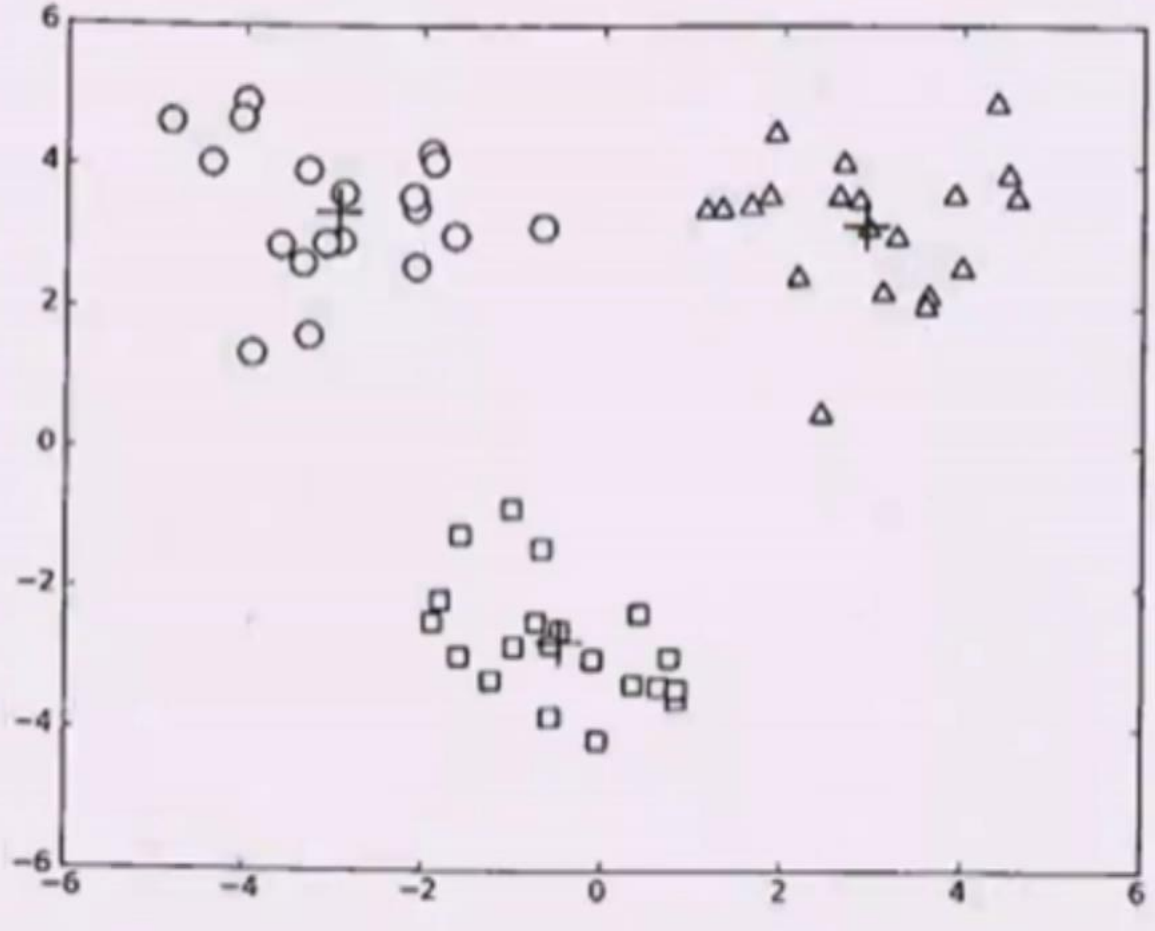
二分K-Means

- 选择的初始值会对聚类结果有影响吗？如何调整？
- 那么首先回答损失函数是什么？
- MSE1，MSE2，MSE3，MSE4
- 两个簇里面的样本数量都很小
- 两个簇中心很近
- 两个MSE很小
- 合并
- 簇中心离的很远
- MSE很大
- 分开



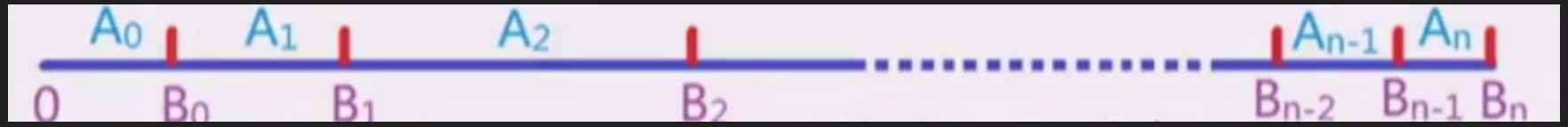
二分K-Means





K-Means++

- 选择初始化簇中心稍微远一点
- 随机选择第一个
- 算每个样本到第一个样本距离，样本距离可以算成概率
- 概率化选择

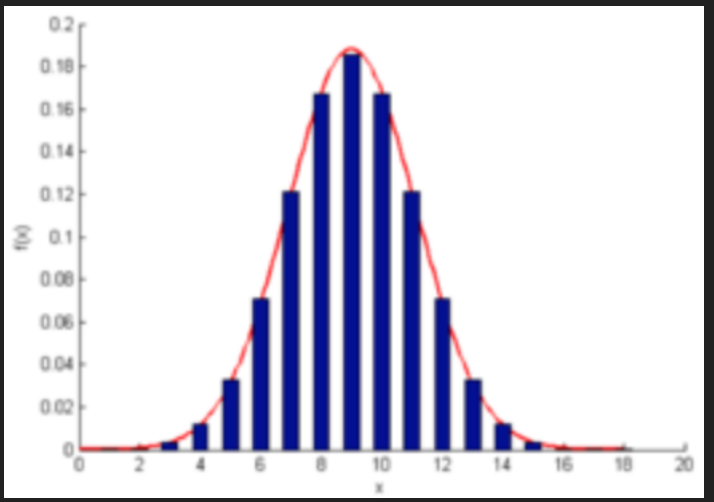


K均值损失函数

- 假定数据点分布符合高斯分布
- K个高斯分布混合得到的样本数据分布
- 最大似然估计！
- 似然函数取最大值
- 概率密度相乘 再 各个族似然相乘
- 找到那些个miu可以使得取最大
- 这个就是K均值的损失函数，从机器学习角度重新看待，平方误差

$$X \sim N\left(\mu,\sigma^2\right), Y = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

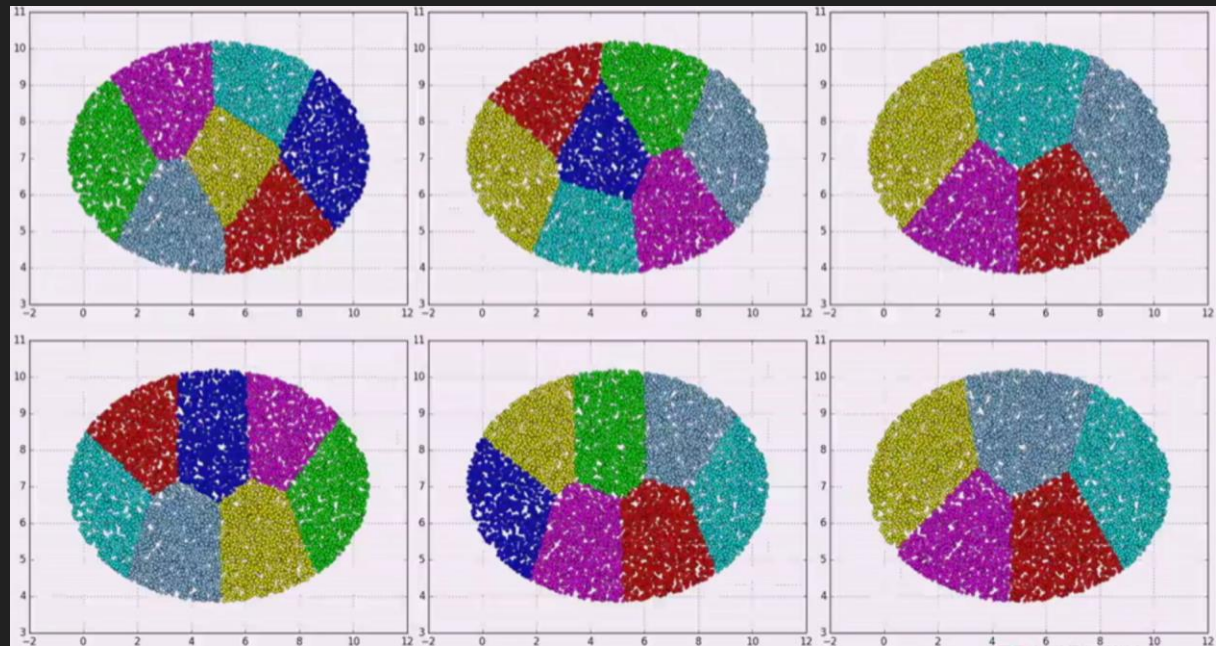


$$J\left(\mu_1,\mu_2,\cdots\mu_k\right) = \frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{N_j}\left(x_i-\mu_j\right)^2$$

求偏导

$$\frac{\partial J}{\partial \mu_j} = -\sum_{i=1}^{N_j}\left(x_i - \mu_j\right) \xrightarrow{令} 0 \Rightarrow \mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i$$

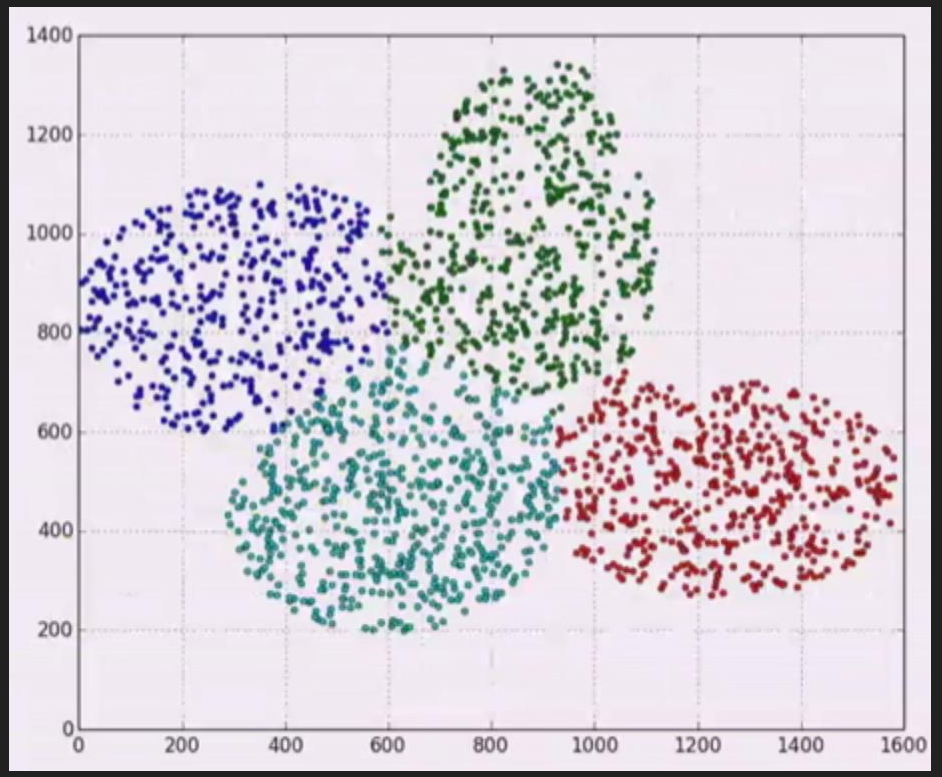
- 公式化解释K均值
- 所以K均值假设了高斯混合模型，GMM，并且假设了方差sigma是一样的
- K均值是在给定损失函数的情况下，梯度下降的一个应用
- 高斯混合分布不是线性回归凸函数，有多个极小值
- K-Means++或者多算几次
- 淬火法或遗传算法来计算全局最优解
- n_init

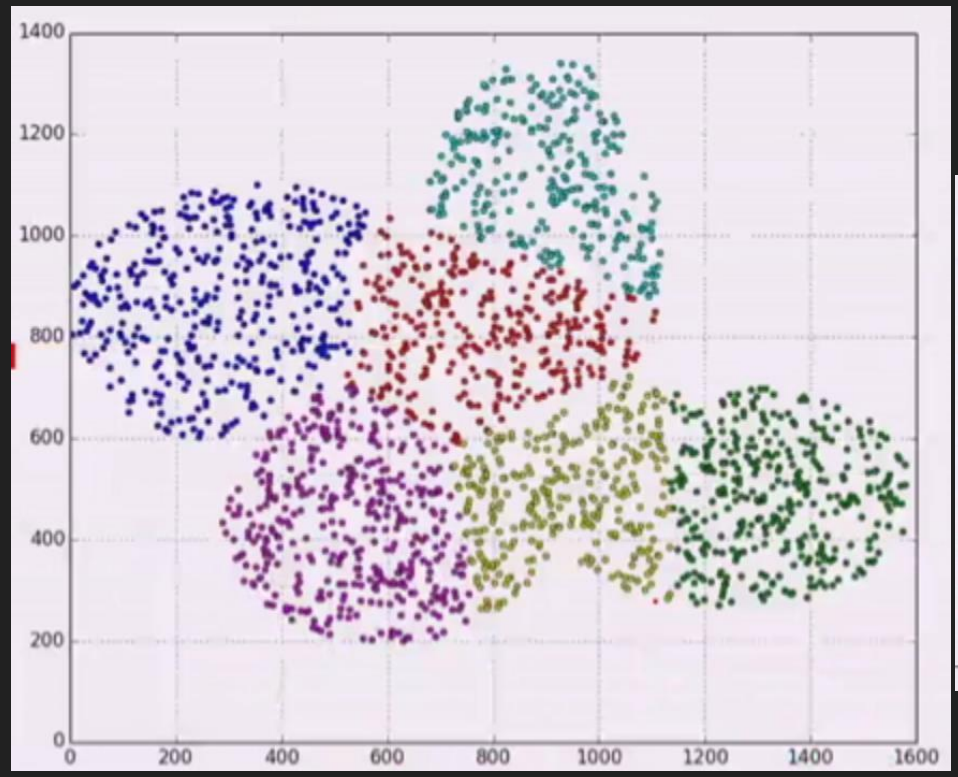


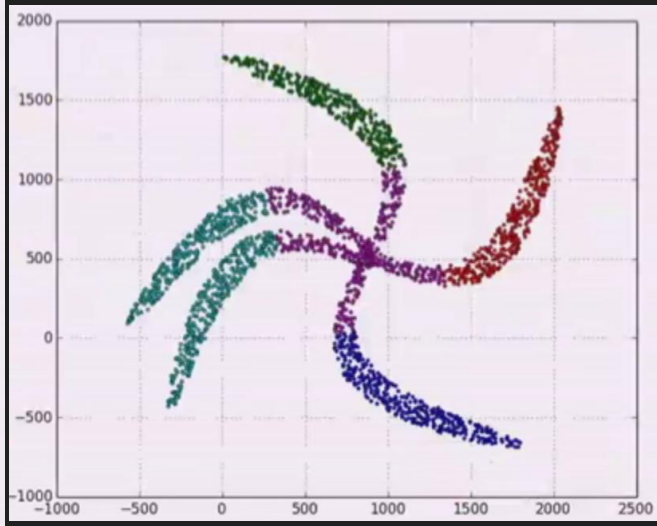
K的选择

- K=N，MSE为0
- K=1，MSE就是原始数据的方差
- 选择一开始下降速度快，后来下降速度慢的
- elbow method，不止于K均值

不同的K


1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600

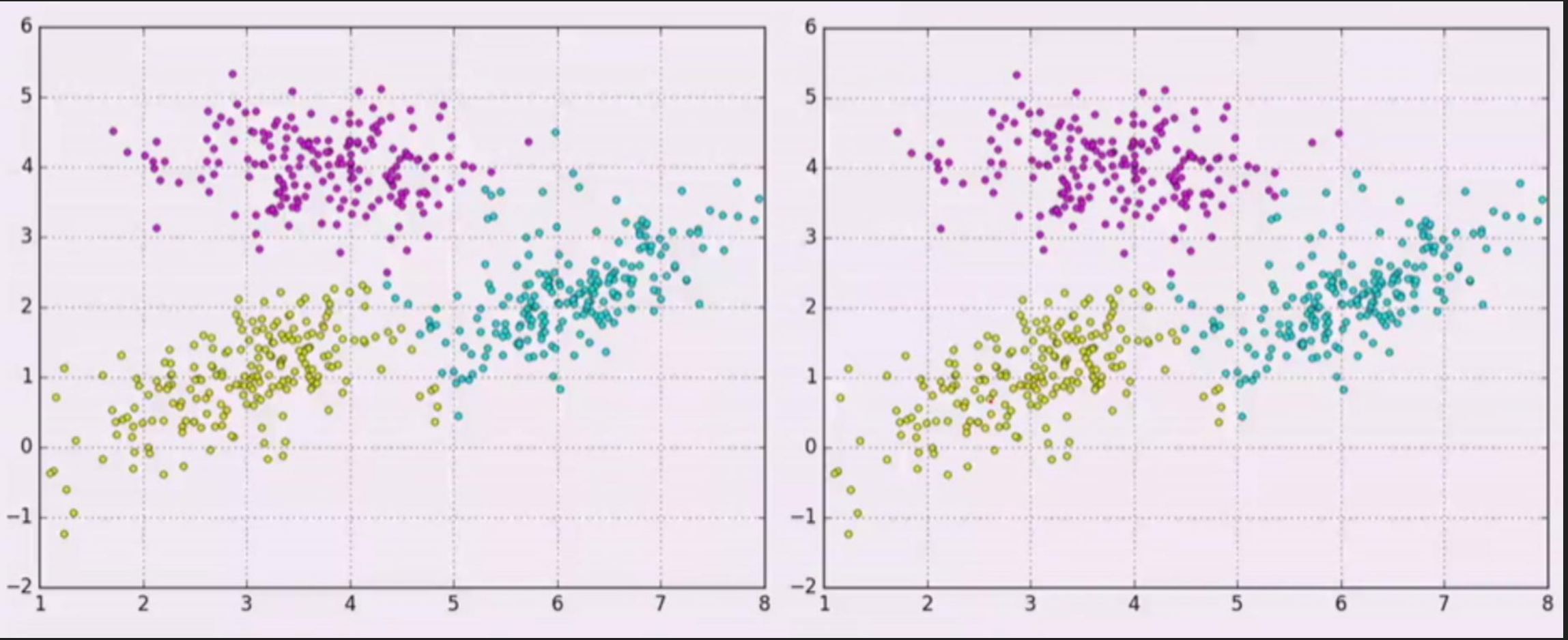

1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600


2000
1500
1000
500
0
−500
−1000
−1000
−500
0
500
1000
1500
2000
2500

Mini Batch K-Means

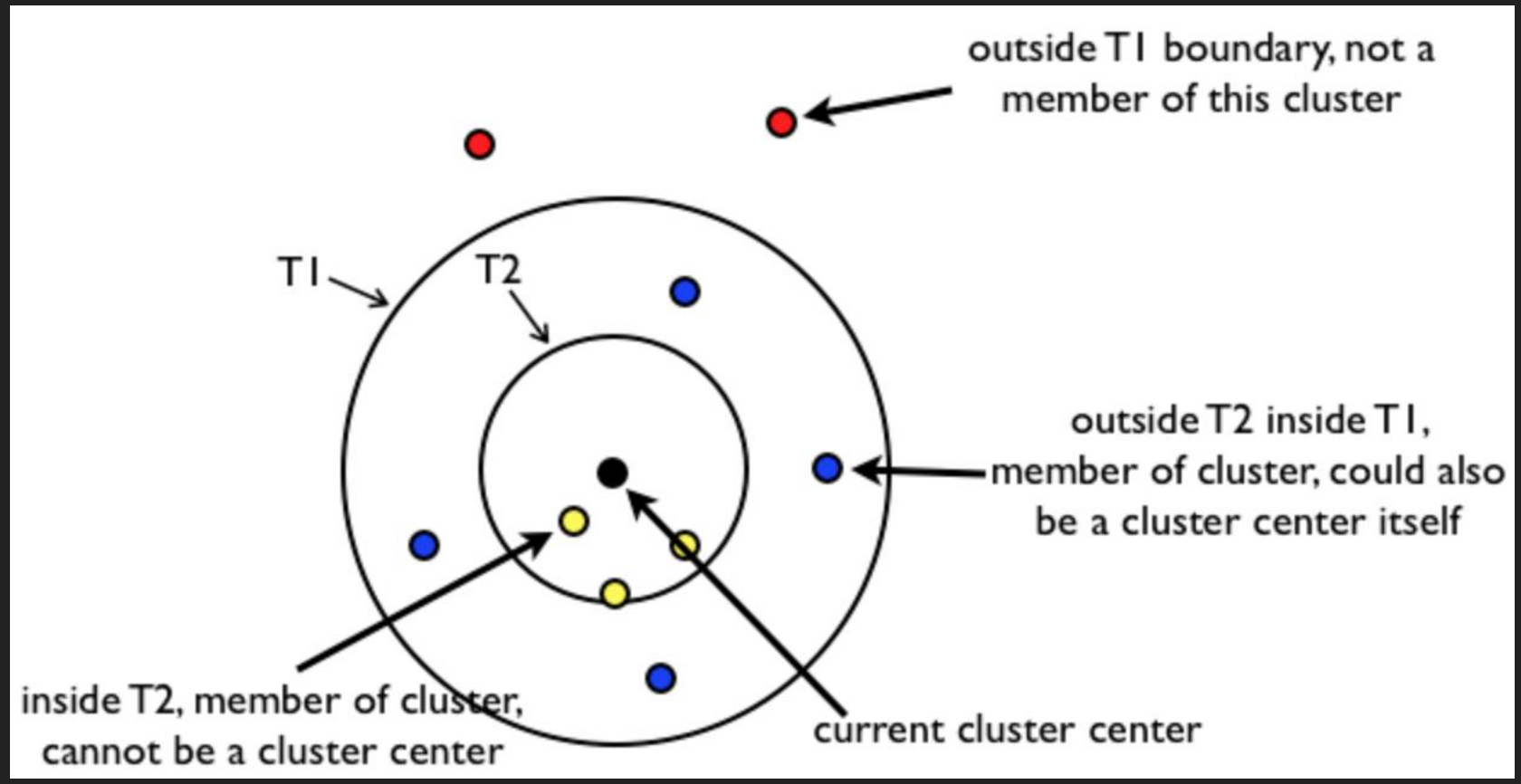
- 本质上，我们之前用的是BGD，用SGD呢？
- 那就是从每个簇中不选择全部求均值，而是选择一部分求均值
- 速度快，数据量大可以选择

效果

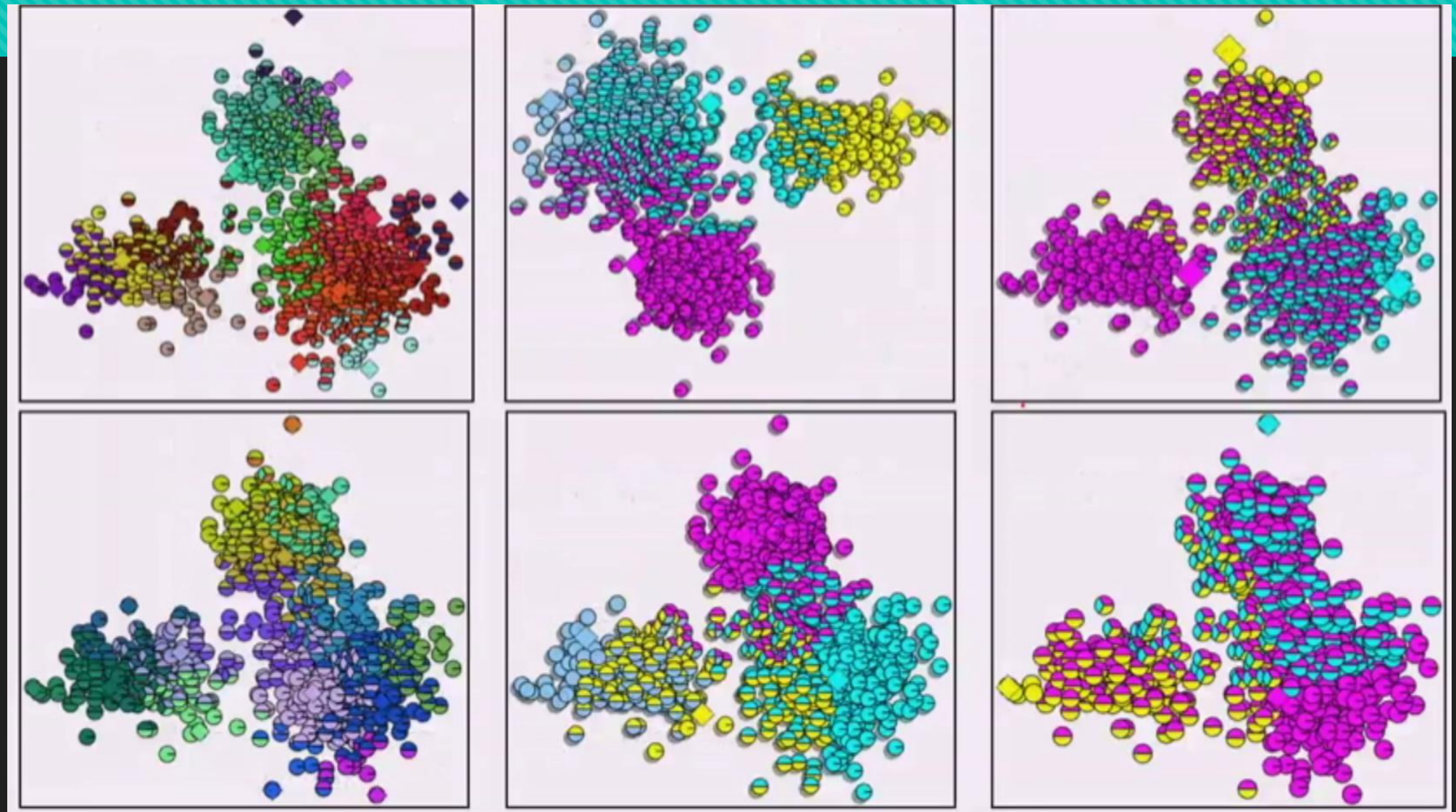


Canopy聚类

- T1和T2不同聚类结果



Canopy算法



Given Label

- 均一性，一个簇中只包含一个类别样本，Precision
- 完整性，同类别样本被归到同一个簇中，Recall
- TradeOff
- V-Measure，F-Measure
- 二者加权平均

$$v_\beta = \frac{(1+\beta)\cdot h \cdot c}{\beta \cdot h + c}$$

ARI评估

C	Y_1	Y_2	$\cdots$	Y_s	sum
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sum	b_1	b_2	$\cdots$	b_s	N

- 已知类别的情况下，看看聚类算法是否对这样的数据集有效
- 评判聚类结果Y和实际结果X相关性
- n11是共同的，a1是X1簇中的样本数量，b1是Y1簇中样本个数
- Rand Index Adjusted Rand index(调整兰德指数)(ARI)
- 表示数据集中可以组成的对数，RI取值范围为[0,1]，值越大意味着聚类结果与真实情况越吻合
- ARI取值范围为[−1,1]，值越大意味着聚类结果与真实情况越吻合。从广义的角度来讲，ARI衡量的是两个数据分布的吻合程度
- **任意取两个是属于某一个类别的概率一样**

$$\mathrm{RI} = \frac{a+b}{C_2^{n_{\text{samples}}}}$$

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

$$\frac{\sum_{i,j} C_{n_{ij}}^2 - \left[\left(\sum_i C_{a_i}^2\right)\cdot\left(\sum_j C_{b_j}^2\right)\right] \Big/ C_n^2}{\frac{1}{2}\left[\left(\sum_i C_{a_i}^2\right) + \left(\sum_j C_{b_j}^2\right)\right] - \left[\left(\sum_i C_{a_i}^2\right)\cdot\left(\sum_j C_{b_j}^2\right)\right] \Big/ C_n^2}$$

AMI

- 利用互信息

$$MI(X,Y)=\sum_{i=1}^{r}\sum_{j=1}^{s}P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$$

$$NMI(X,Y)=\frac{MI(X,Y)}{\sqrt{H(X)H(Y)}}$$

$$AMI(X,Y)=\frac{MI(X,Y)-E[MI(X,Y)]}{\max\{H(X),H(Y)\}-E[MI(X,Y)]}$$

轮廓系数

- 计算同簇内每一个样本到同簇内样本的平均距离，可以度量这个样本和其他同簇样本的相似性
- 计算一个簇内每一个样本到不同簇内所有样本的距离，不同簇的那些样本距离求平均，然后求最小的那个距离，是不相似性
- 第一个值很小，第二个值很大，那就簇内很典型性的样本
- 如果相反，按道理应该属于另外一个簇了
- Si接近1说明样本i聚类合理，Si接近-1说明样本更应该分到其他簇
- Si接近0说明在簇分界上

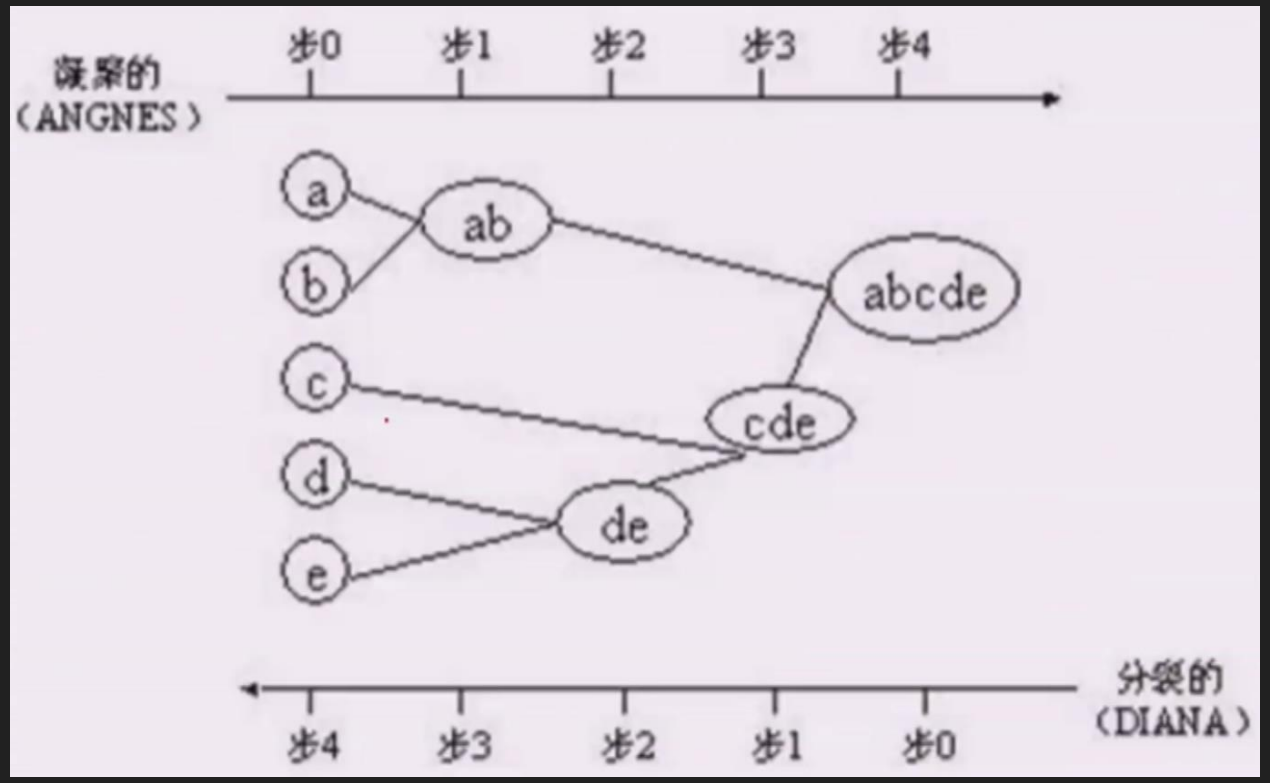
$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

层次聚类

- 理解起来有点像无监督的决策树
- 分裂的层次聚类：DIANA
- 把原始数据集去不断的分裂，然后去计算每个子数据集里面的相似性，然后不断的分裂，把数据集分为很多的类别
- 凝聚的层次聚类：AGNES
- 把一个个样本，不断的自底向上的聚类，然后一层一层的来聚，最后聚成一个完整的数据集，这种用的更多一些

层次聚类

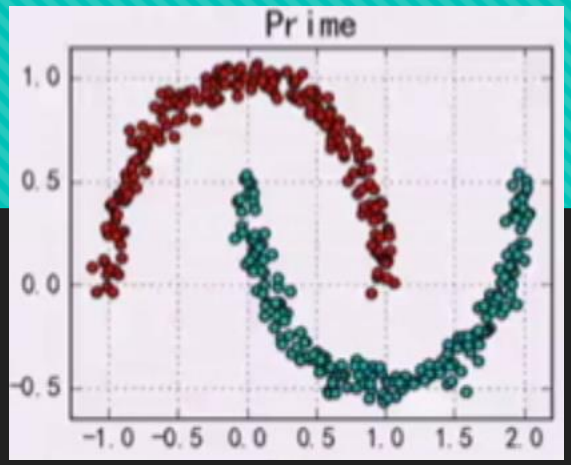
- 如果我们只关系结果的话，那么在某一个时刻是要停止聚类的，在一些数据集中做层次聚类是合适的，包含层次的数据集！地域！



如何凝聚？

- 如果两个样本，可以很好的度量距离，如果已经聚了一层，如何度量簇之间的相似性
- 最小距离：两个簇中，最接近样本的距离，城市和城市边界最短距离，成链状一条线了
- 最大距离：两个簇中，最远的样本的距离，某一个簇存在异常值就很麻烦，簇本身比较狭长
- 平均距离：
- 两两样本距离的平均
- 两两样本距离的平方和

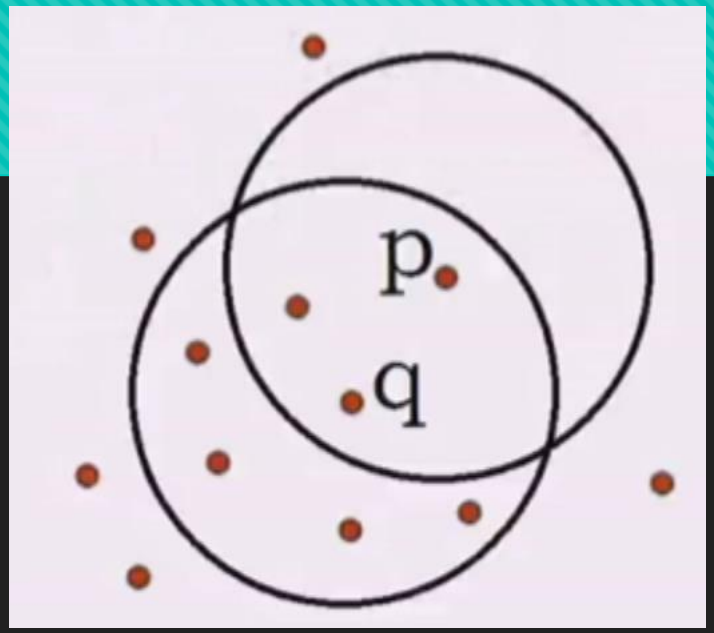
密度聚类

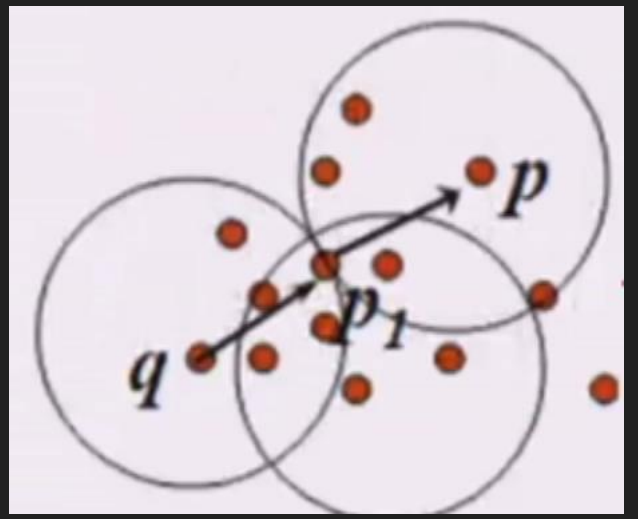


- 统计样本周边的密度，把密度给定一个阈值，不断的把样本添加到最近的簇
- 人口密度，根据密度，聚类出城市
- 解决类似圆形的K-Means聚类的缺点，密度聚类缺点计算复杂度大，空间索引来降低计算时间，降低查找速度

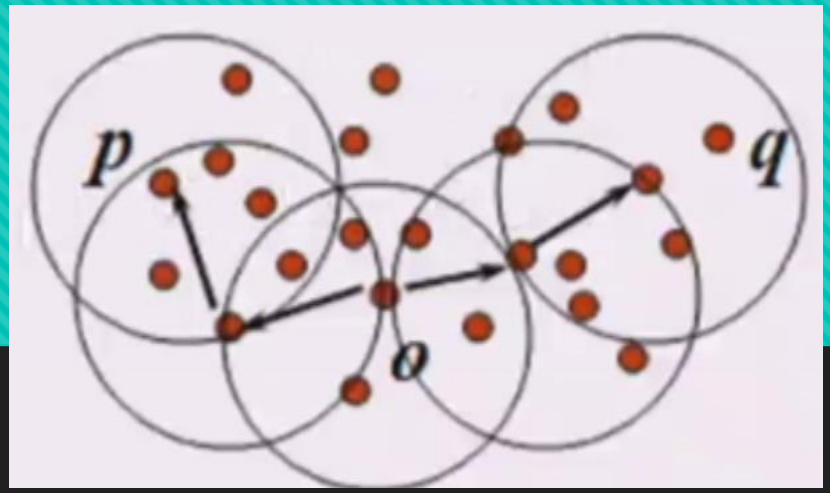
DBSCAN

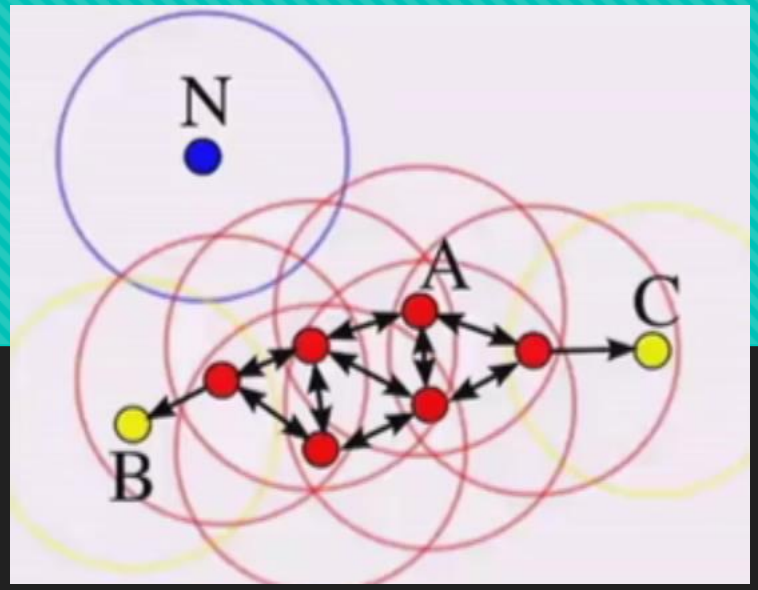
- Density Based Spatial Clustering of Applications with Noise
- 对象邻域：给定对象在半径内的区域
- 如果给定5为阈值，那么q是6，p是3，那么q是核心对象
- 而p是在q这个范围(制定一个半径)内的，那么说q到p是核心密度可达
- q密度可达p1，p1密度可达p，那么q到p是密度可达



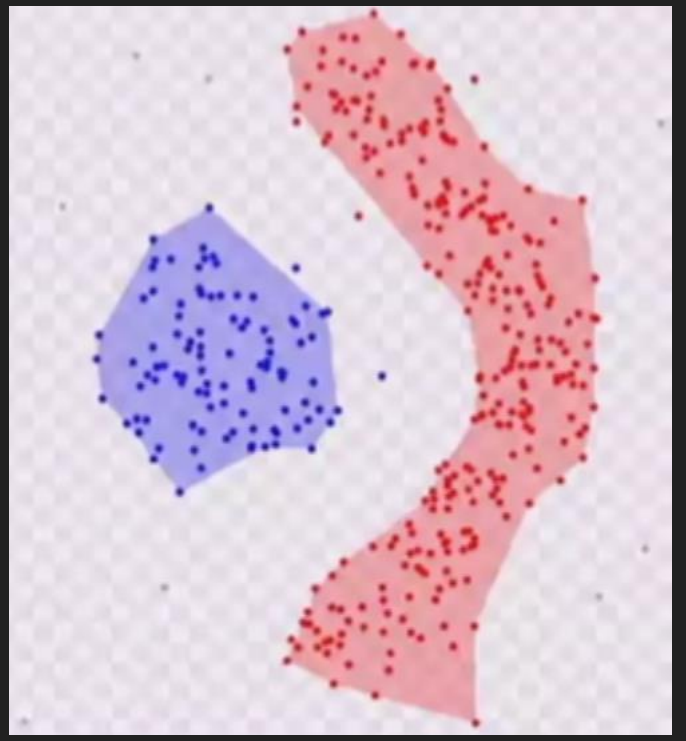


密度可达





- 从o点能密度可达q，也能密度可达p
- p和q叫密度相连
- 簇就是密度相连的最大的点的集合
- 即最大的密度相连构成的集合就是簇
- 如何一个点不是核心对象，也不能被别的点密度可达，就是噪声

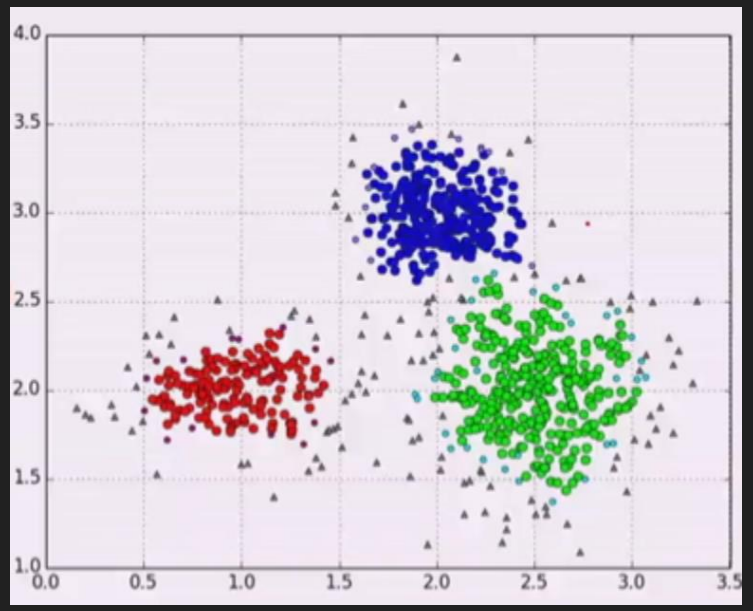


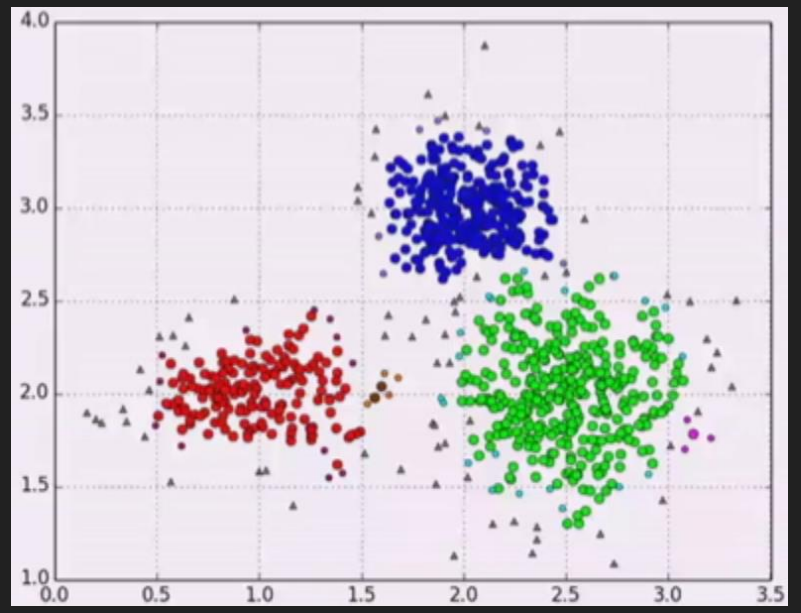
DBSCAN

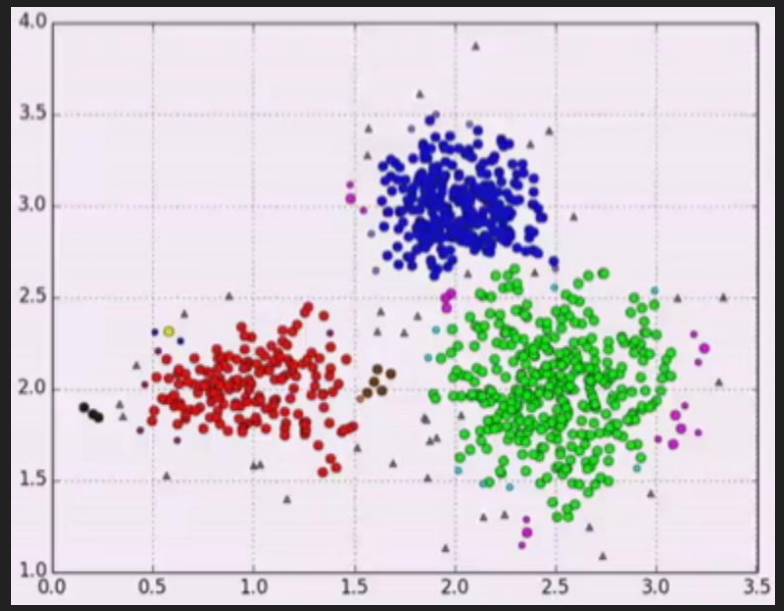
- 举例子，层次聚类
- 如果P密度可达A，B，C，那就把它们连接在一起
- A可达E，F，B不是核心对象，C可达G
- 所有点都要进行是否是核心对象的判定
- 那么这些点同属于一个簇，最后没有更多样本可以加进来，这个时候扫描结束
- 不位于簇中的点就是噪声
- K不需要给定，只给m个阈值，和半径r

r=0.1 m=5,3,2

- 给定的m个数不够，簇会变多，一般发生在边缘处

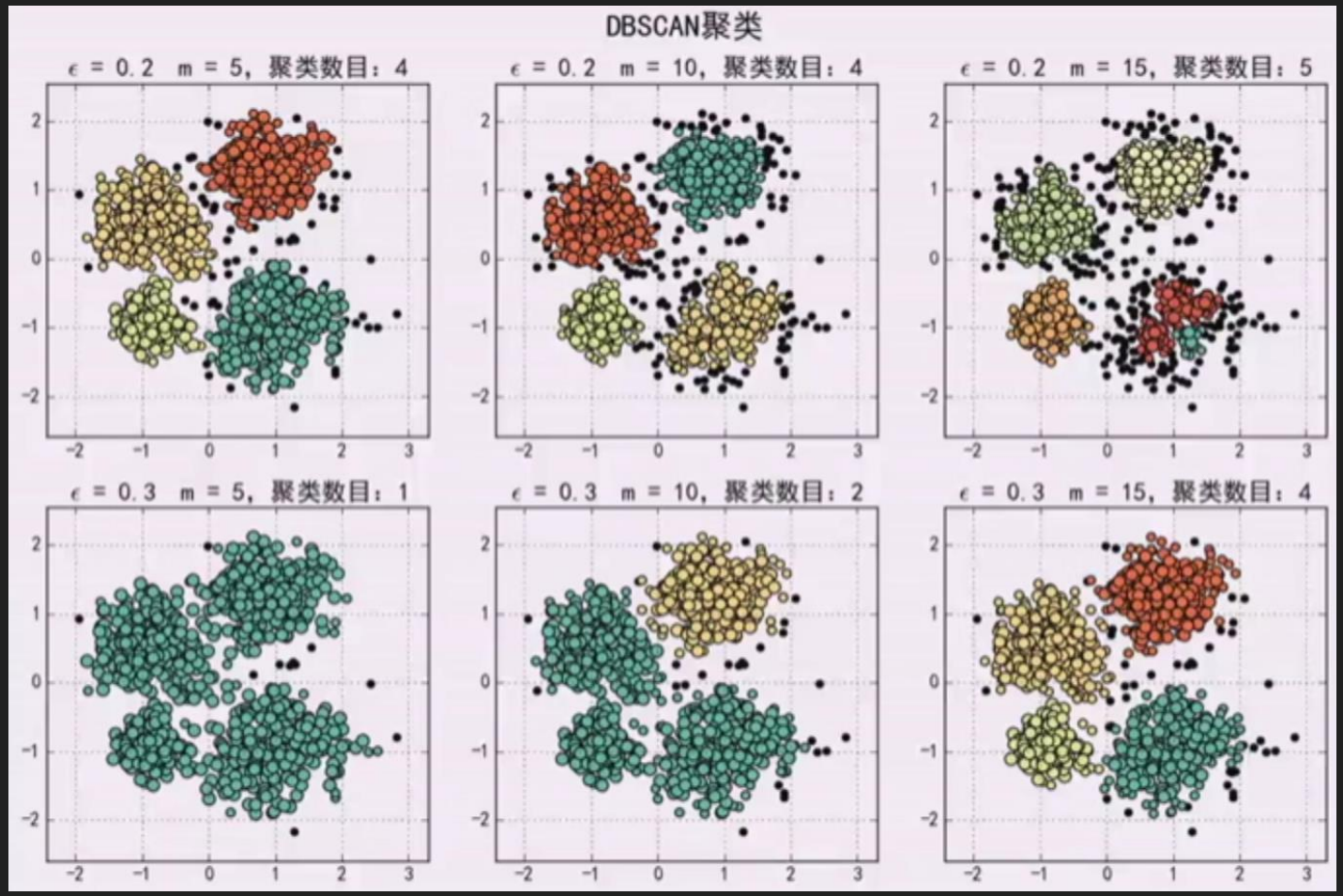






不同的r半径

- 太大就会都聚类到一起



谱和谱聚类

- 谱：Y=A*X，矩阵X乘以A等于对矩阵X做了空间线性变换，那么Y=map(X)，A就是map这个线性算子，它的所有特征值的全体，称之为方阵的谱
- 方阵的谱半径为最大的特征值
- 谱聚类是一种基于图论的聚类方法，通过对样本数据的拉普拉斯矩阵的特征向量进行聚类，从而达到对样本数据进行聚类的目的

谱聚类

- 解决区域重叠问题，密度聚类对应区域重叠问题不太好办
- 我们有一堆个样本，可以构建成全连接图，并且两两样本之间总是可以去求相似度的
- 两两样本之间构建邻接矩阵来表示图
- 邻接矩阵上面的值，是用高斯相似度计算得来
- 然后对角线都是0，可以根据高斯相似度看出来，这样有了矩阵W
- 除了对角线都有相似度的值，然后把它们按行或列加和得对角阵D
- L=D-W，这样的L矩阵叫做Laplace矩阵

$$W_{ij} = S_{ij} = exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$$

$$\mathbf{D} = \begin{pmatrix} d_1 & \cdots & \cdots \\ \cdots & d_2 & \cdots \\ \vdots & \vdots & \ddots \\ \cdots & \cdots & d_n \end{pmatrix}$$

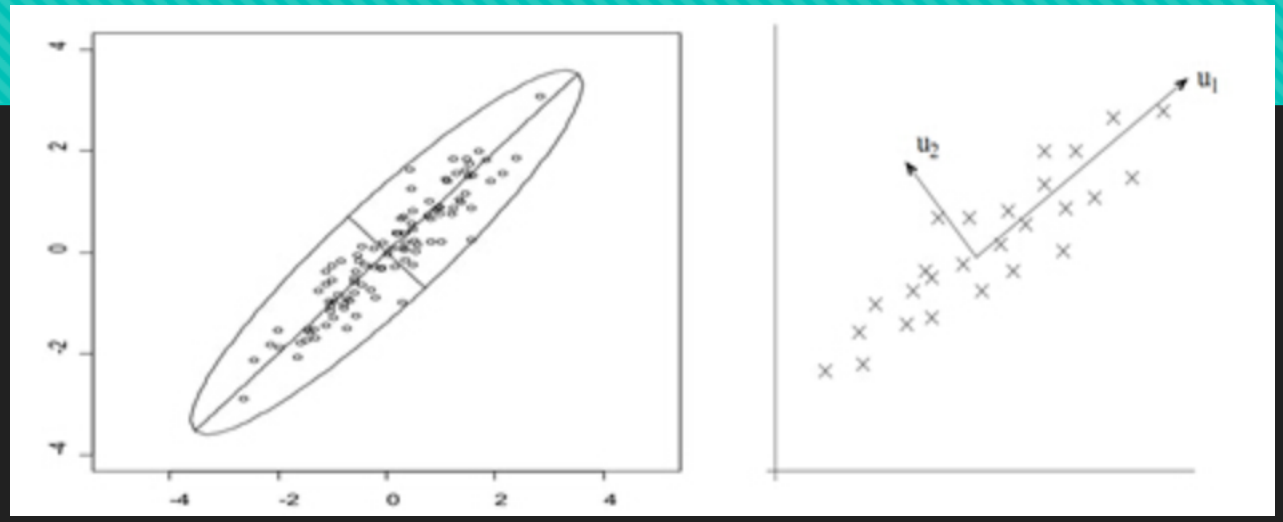
$$d_i = \sum_{j=1}^{n} w_{ij}$$

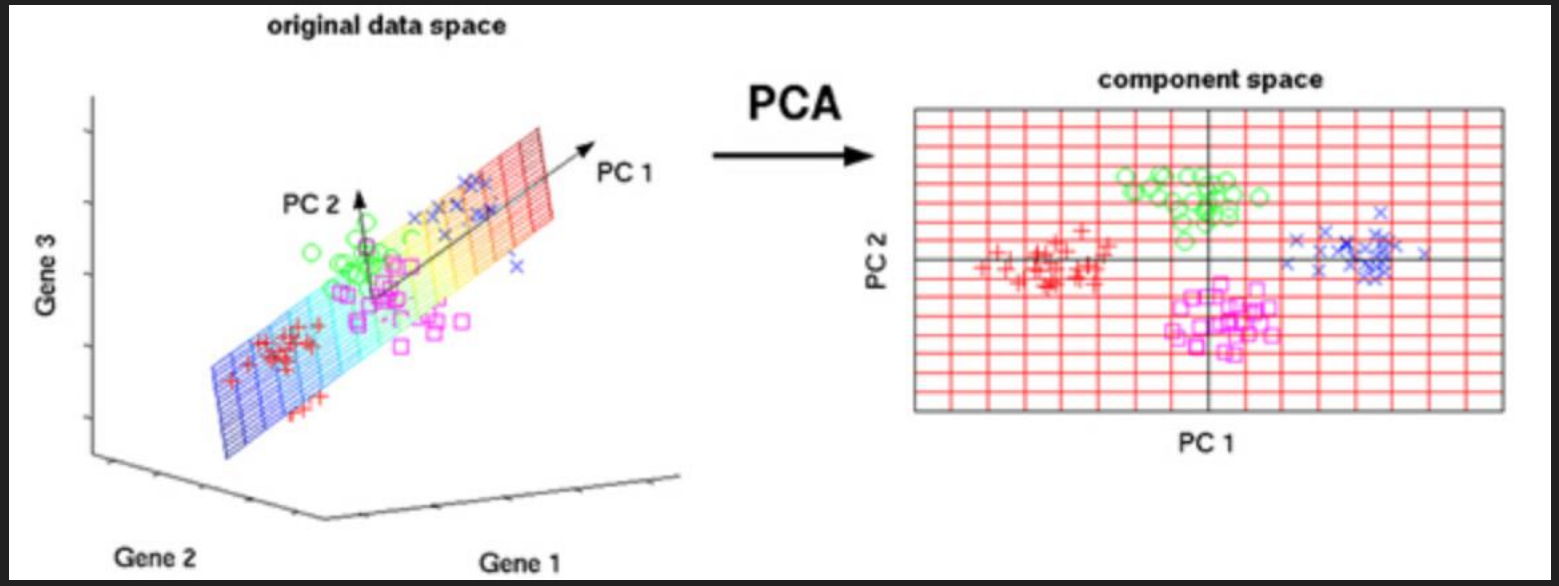
谱聚类

- L矩阵是N*N的，N是样本个数，实数形成的对数矩阵，求特征值和特征向量
- L*ui = lambda i*ui，lambda是特征值，ui是特征向量，一组lambda有从大到小可以排序
- 每个对应的lambda都对应一个ui，每个ui是一个个的列向量，比如u11,u21,u31,un1
- 根据排序默认从小到大，逆序之后我们就取前面的几个ui列向量就可以了，其实这是一种降维啊！
- 然后我们的前面的这几个列向量ui就成了新的对应每个样本的几个重要的特征！
- 最后我们用K-Means聚类算法对样本进行聚类就好了
- 谱聚类和PCA的关系？？？
- 就是Laplace矩阵做一个主成分分析PCA，然后做K均值聚类

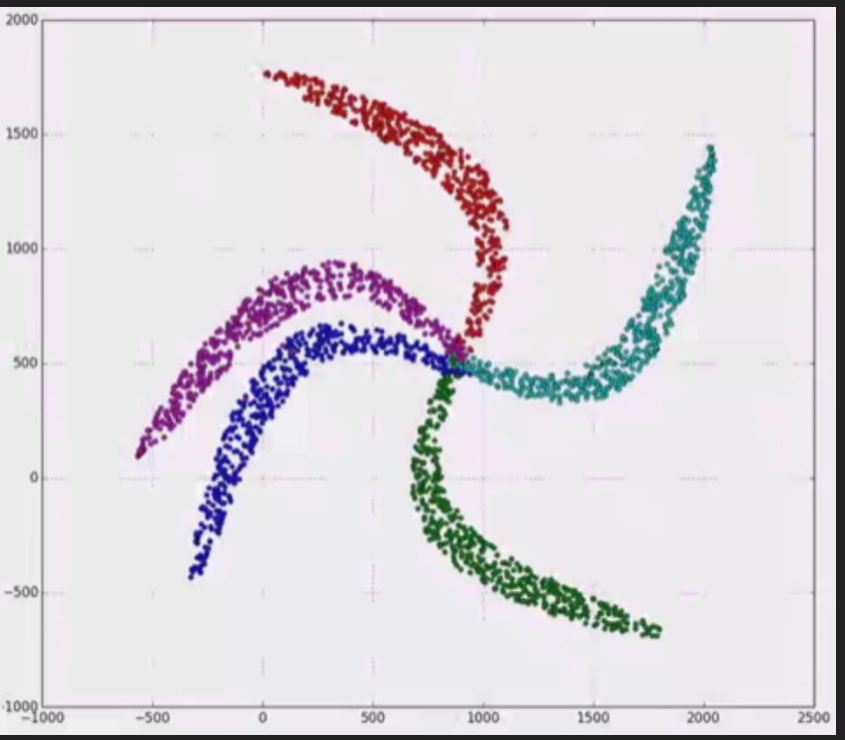
PCA降维

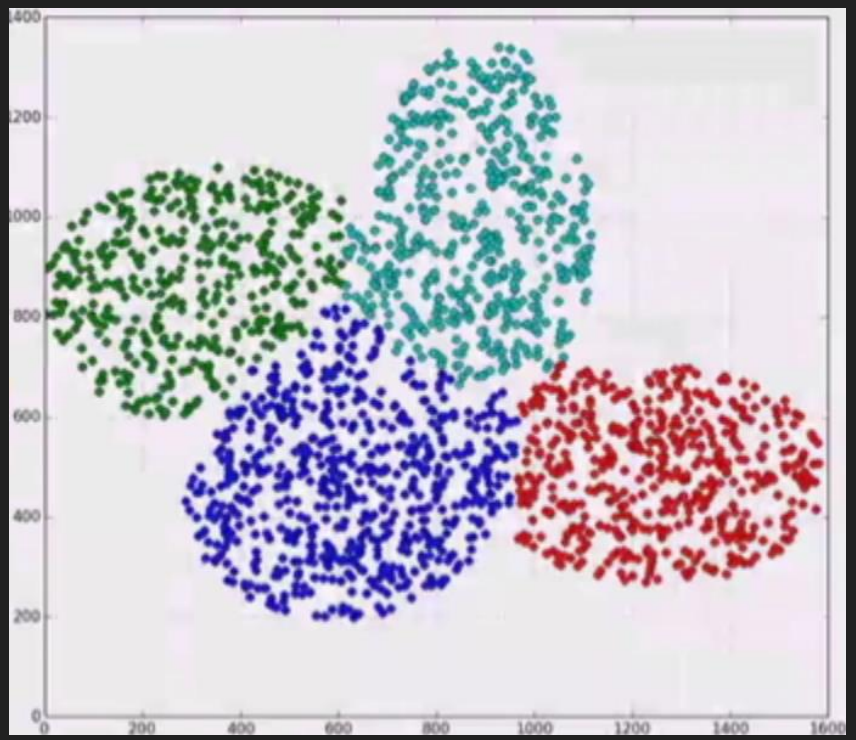
- 盲目减少指标会损失很多信息
- 容易产生错误的结论
- 各变量间存在一定的相关关系
- 先均值归一化，映射到原点
- 协方差矩阵
- 协方差（Covariance）
- 用于衡量两个变量的总体误差

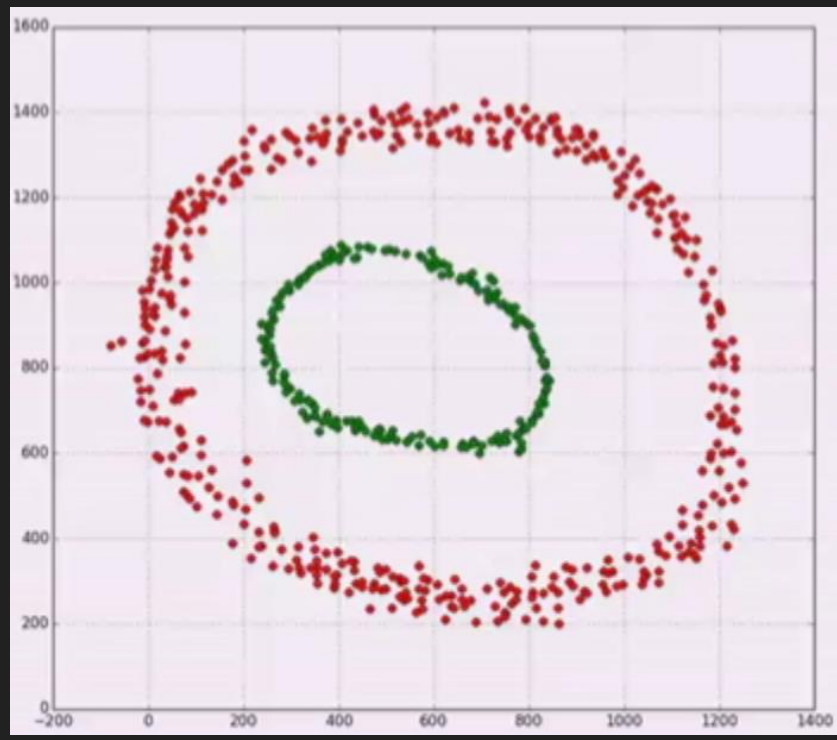




谱聚类


2000
1500
1000
500
0
−500
−1000
−1000
−500
0
500
1000
1500
2000
2500


1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600


1600
1400
1200
1000
800
600
400
200
0
−200
0
200
400
600
800
1000
1200
1400