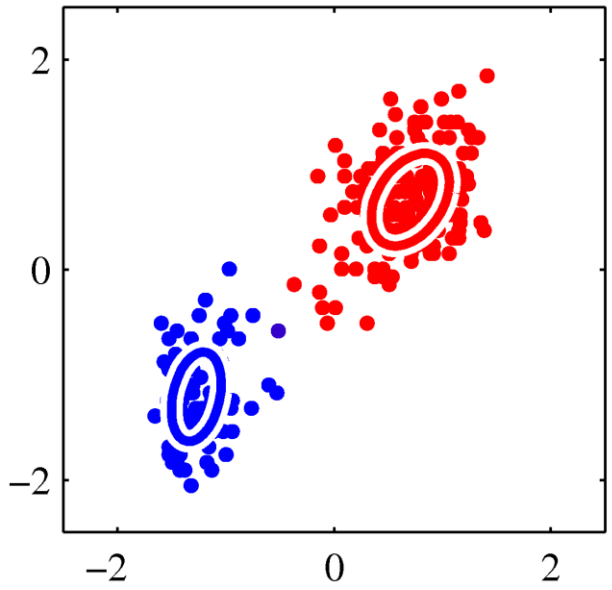
《机器学习及其在化学中的应用》2025年课程

Sec. 8

混合模型：聚类

（K-均值法；高斯混合模型）



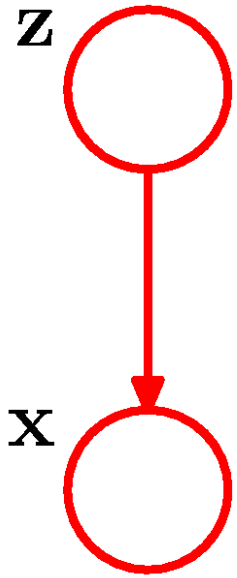
刘志荣 (LiuZhiRong@pku.edu.cn)

北京大学化学学院

2025.11.10

内容提要

- K-均值法
- 混合高斯模型
- 应用例子

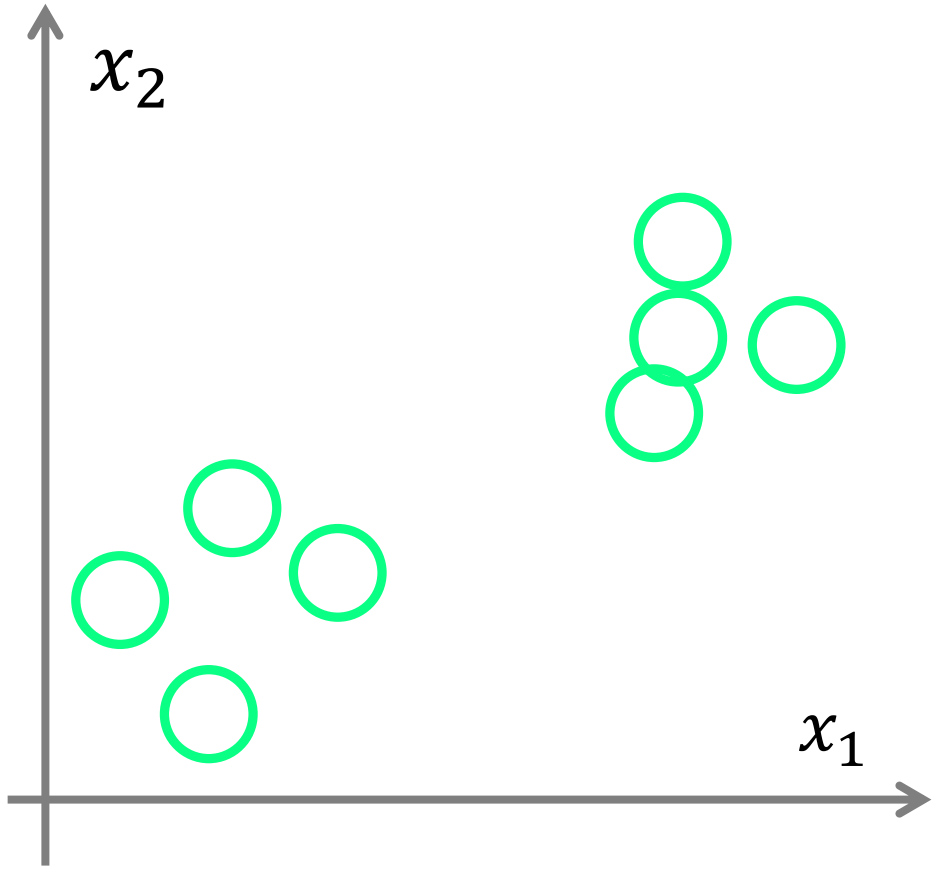


1. K-均値法…
（K-means clustering）

団簇ing…

无监督学习（Unsupervised Learning）

- 给学习算法一个数据集$\{\mathbf{x}_n\}\,(n = 1,2,\ldots,N)$，里面没有答案或标签，让算法据此寻找其中的规律与结构。
- 例如：聚类（clustering）
  - 市场分割（客户群）；
  - 产品分割（电商目录、手机套餐）；
  - 社交网络（社区）；
  - 新闻聚合；
- 例如：降维



聚类：K-均值法（K-means）

- 假设我们想把数据点分成K组，K已知。
- 直观想法：与组内点的距离小于与组外点的距离。
- 一种方法：找一些原型prototype $\boldsymbol{\mu}_k$（聚类中心cluster centroids）
- 在K-均值法中，定义畸变函数（Distortion function）：

$$J(\{r_{nk}\},\{\boldsymbol{\mu}_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

  - r_{nk}是第n个数据点分类结果的独热编码（的第k分量）。如果被分到第k个组，则$r_{nk} = 1$；否则为0。
- 任务变成：

$$\min_{r_{nk},\boldsymbol{\mu}_k} J(\{r_{nk}\},\{\boldsymbol{\mu}_k\})$$

求解方法…

arg min：使后面函数达到最小值时的变量的取值

（1） $\min\limits_{r_{nk}} J$ ⇒ $r_{nk}=\begin{cases}1, & \text{if } k=\arg\min\limits_j \|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2\\ 0, & \text{otherwise}\end{cases}$

$\mathbf{x}_n$离哪个$\boldsymbol{\mu}_j$近就分到哪个组。

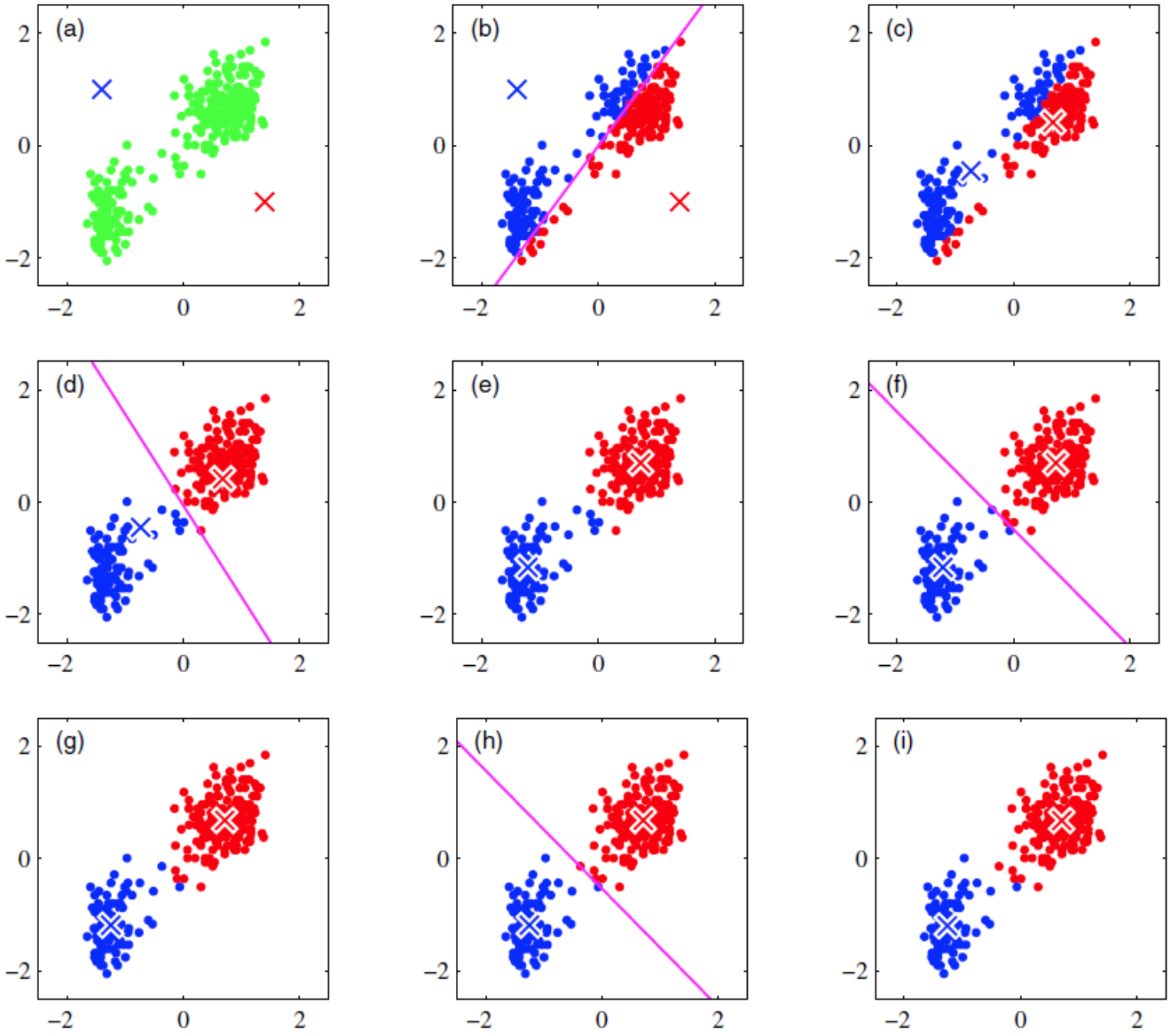
（2） $\min\limits_{\boldsymbol{\mu}_k} J$ ⇒ $\boldsymbol{\mu}_k=\frac{\sum_n^N r_{nk}\mathbf{x}_n}{\sum_n^N r_{nk}}$

$\boldsymbol{\mu}_k$等于组内数据点$\mathbf{x}_n$的平均值。

（3）不断循环（1,2），迭代求解。

- J不断下降，但有可能收敛到局部极小值。可用不同初值多次求解。

例子…



在线学习

- $\boldsymbol{\mu}_k^{(\text{new})} = \boldsymbol{\mu}_k^{(\text{old})} + \eta_n \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{(\text{old})} \right)$

基于任何距离（相似性）定义$\mathcal{V}$

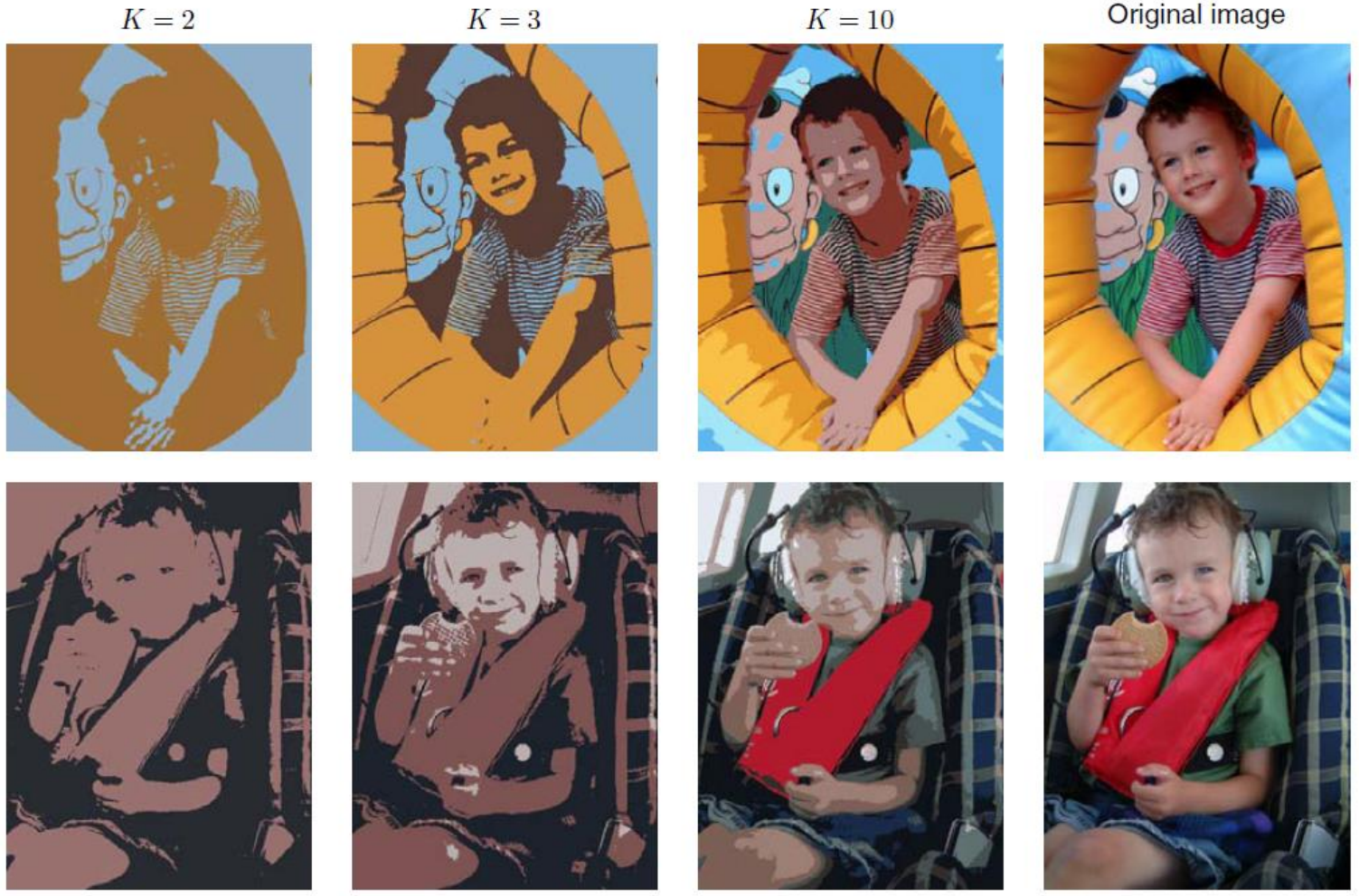
不一定使用欧氏距离

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

应用例子：调色板选择与图像压缩

- 有损压缩。全彩色：3 bytes

  调色板：4.2%, 8.3%, 16.7%。

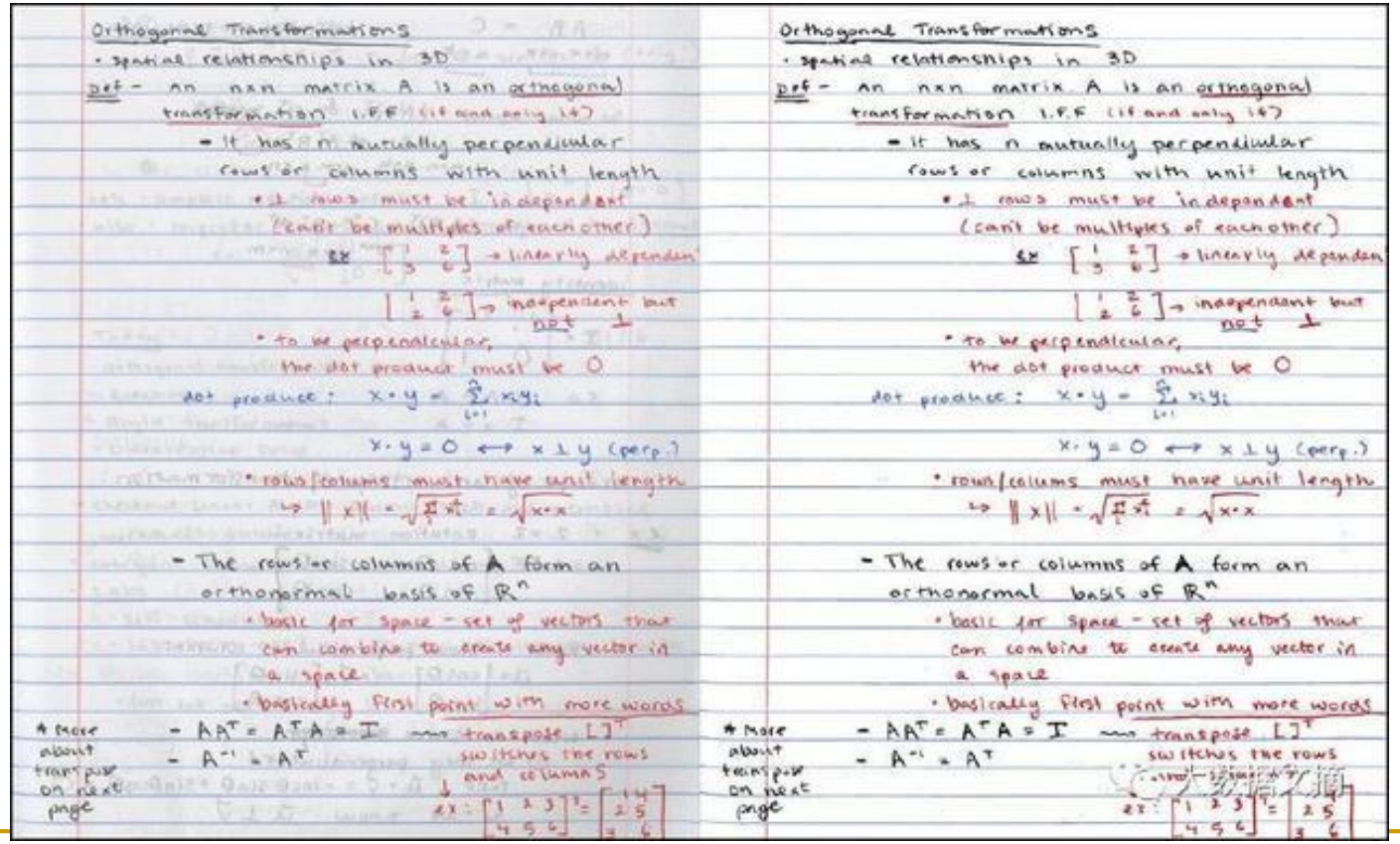


Bishop Fig. 9.3

应用例子：扫描笔记去噪

- 左：输入扫描件（300 DPI，7.2MB PNG / 790KB JPG.）
- 右：输出图片（8种颜色，300 DPI，121KB PNG）。

扫描笔记去噪-代码：
noteshrink-master.zip



https://baijia.baidu.com/s?id=1595977285675528555
手把手：扫描图片又大又不清晰？这个**Python**小程序帮你搞定！

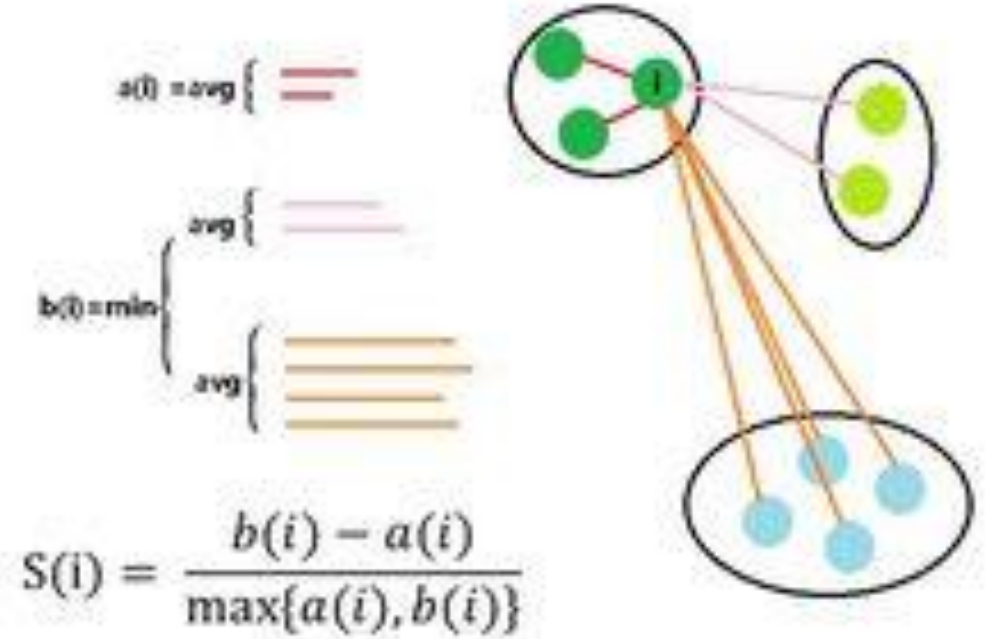
选择聚类数K

- 非监督学习，因此没有所谓的“最佳”K值。
- 通常是需要根据不同的问题，人工进行选择的。选择的时候思考采用聚类的动机是什么，然后选择能最好服务于该目标的聚类数。
- 从数据本身的特征来讲，最佳K值对应的类别下应该是类内距离最小化并且类间距离最大化。有些辅助指标可以用来评估这种特征，比如平均轮廓系数、类内距离/类间距离等都可以做此类评估。

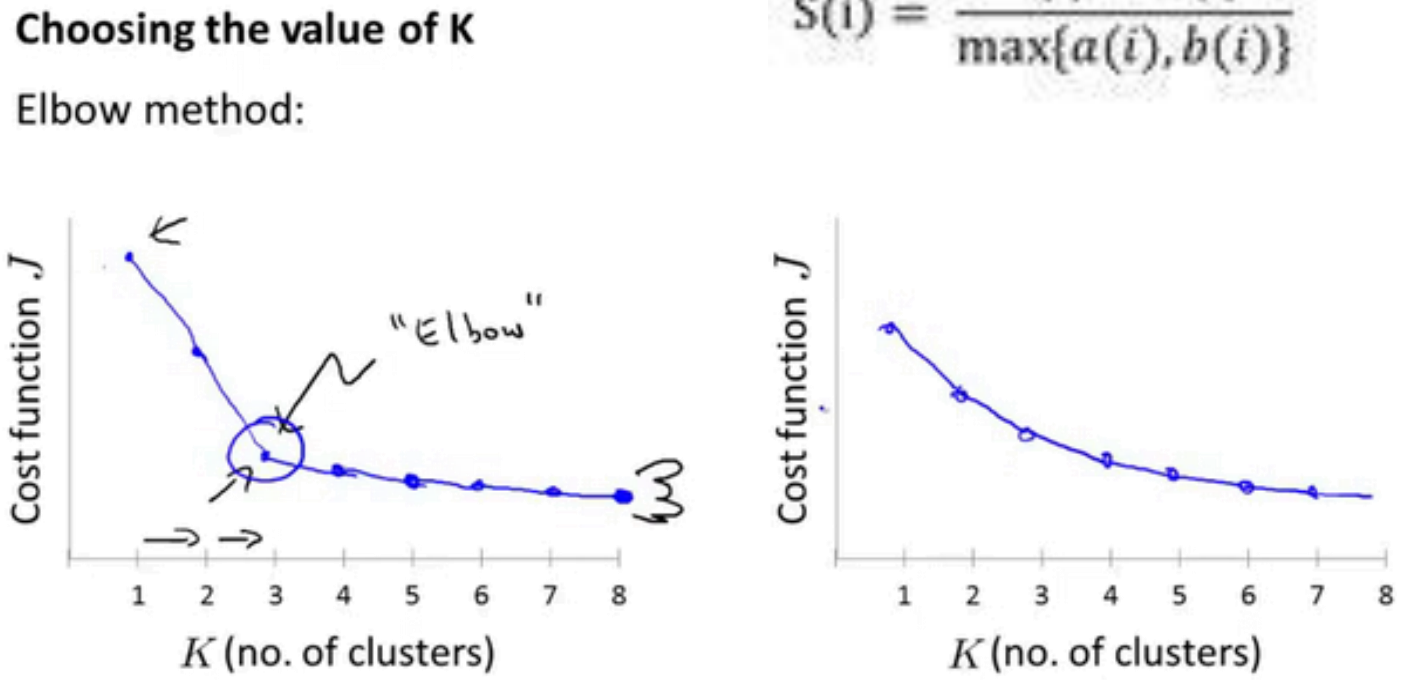
- 平均轮廓系数：　　$S(i) \in [-1,1]$

$$S(i) = \frac{b(i)-a(i)}{\max[a(i),b(i)]}$$

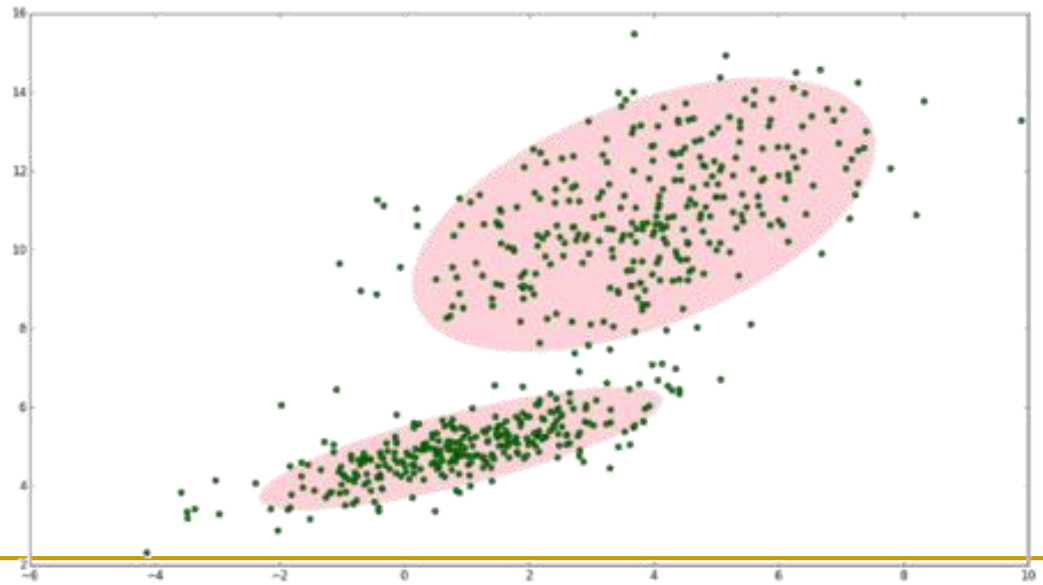
  - 其中$a(i)$与$b(i)$分别是第i类的类内平均距离及其与最近类的类间平均距离。



- “肘部法则”



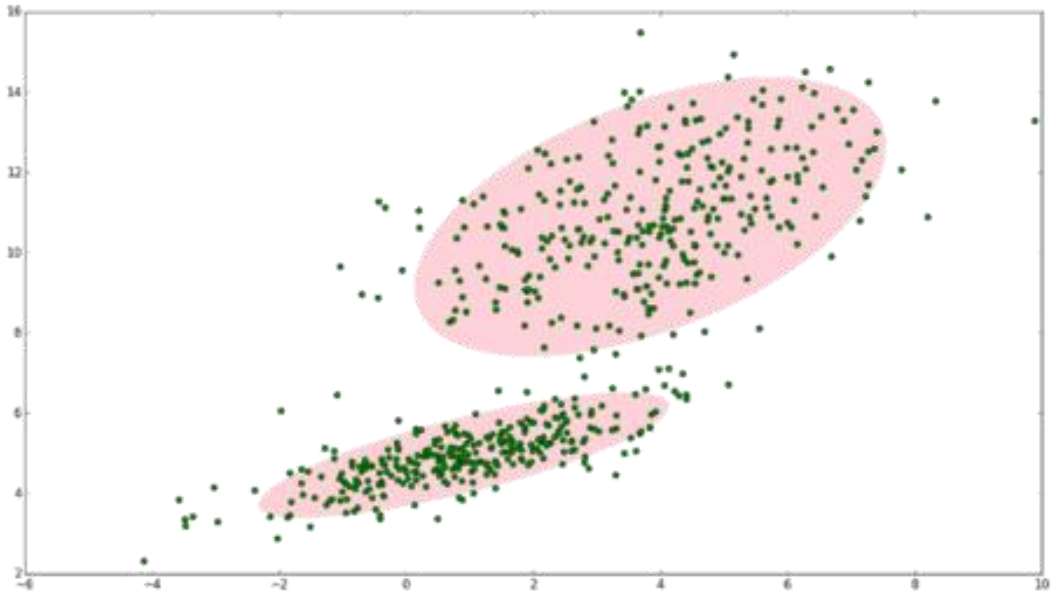
2. 高斯混合模型…
（Gaussian mixture models）



高斯混合模型

- 假设$\mathbf{x}$的分布是由多个高斯分布混合而成：

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



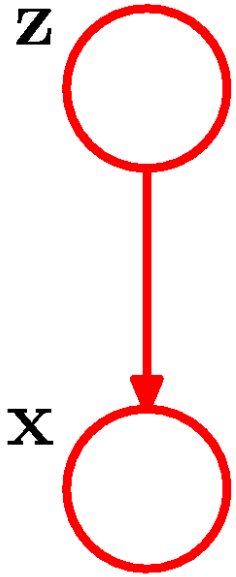
- 通过引入具有K个离散取值的隐藏变量$\mathbf{z}$（采用独热编码），可以变成简单的贝叶斯网络：

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

其中

$$p(z_k = 1) = \pi_k$$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Expectation-Maximization（EM）算法
期望-最大化

$$p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_n p(\mathbf{x}_n) = \prod_n \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_n \ln\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]$$

- 最大化这个似然函数。
- 对$\boldsymbol{\mu}_k$求导，得（Bishop 9.2.2）

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n$$

  - 其中 $\gamma(z_{nk}) \equiv p(z_k = 1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$

    代表$\mathbf{x}_n$来自第k个高斯函数的概率，而

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

- 对$\boldsymbol{\Sigma}_k$求导（复杂，略），得

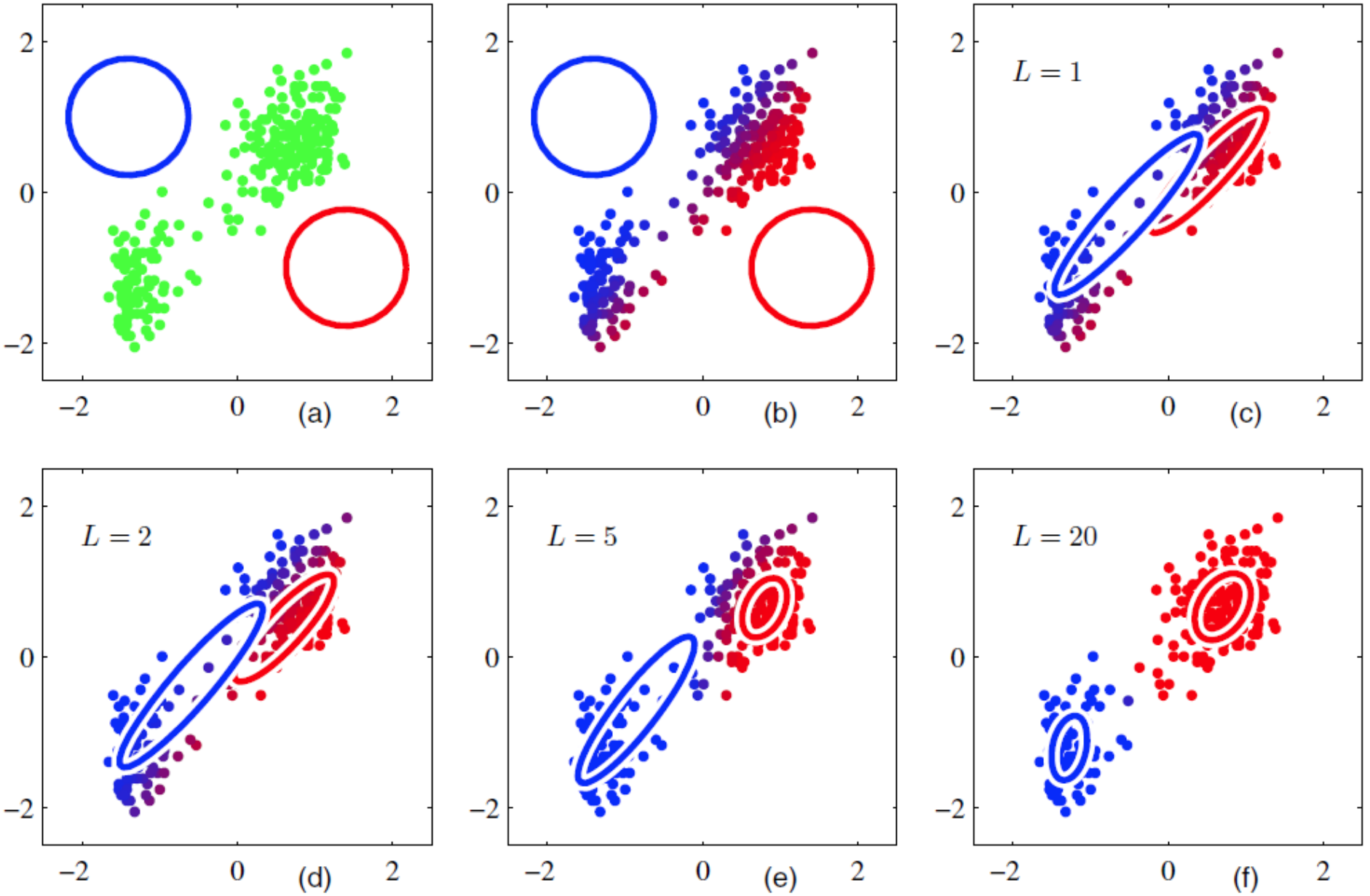
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

- 对π_k求导，得

$$\pi_k = \frac{N_k}{N}$$

- EM算法：
  - 初始化$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$。
  - E步骤：计算$\gamma(z_{nk})$
  - M步骤：用上述公式更新$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$
  - 重复E与M步骤直至收敛。

例子

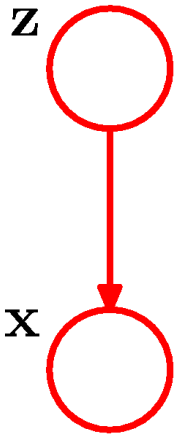


一般隐藏变量下的EM算法

- 假设x是可观测量，z是不可观测的隐藏量。我们想通过对x的观测来求解$p(x,z|\theta)$中的参数θ。
- $p(x,z|\theta)$比较简单（如高斯函数），求$\ln p(\mathbf{X},\mathbf{Z}|\theta)$ 对θ的极值比较容易。但$\mathbf{Z}$未知，我们需考虑

$$p(\mathbf{X}|\theta)=\prod_n p(x_n|\theta)=\prod_n \int p(x_n,z_n|\theta)dz_n$$

$$\ln p(\mathbf{X}|\theta)=\sum_n \ln \int p(x_n,z_n|\theta)dz_n$$



- 对θ求导数：

$$\frac{\partial \ln p(\mathbf{X}|\theta)}{\partial\theta}=\sum_n \frac{\int \frac{\partial p(x_n,z_n|\theta)}{\partial\theta}dz_n}{\int p(x_n,z_n|\theta)dz_n}=\sum_n \int \left[\frac{p(x_n,z_n|\theta)}{\int p(x_n,z_n|\theta)dz_n}\cdot\frac{\partial \ln p(x_n,z_n|\theta)}{\partial\theta}\right]dz_n$$

$p(z_n|x_n,\theta)$

$$\frac{\partial \ln p(\mathbf{X}|\theta)}{\partial \theta} = \sum_n \int \left[p(z_n|x_n,\theta) \cdot \frac{\partial \ln p(x_n,z_n|\theta)}{\partial \theta} \right] dz_n = 0$$

- 利用上式可得到某些类型的迭代求解方法。在某些情况下（如混合高斯模型），如果把$p(z_n|x_n,\theta)$看做固定的值，即与θ无关，

$$\sum_n \int \left[p(z_n|x_n,\theta_0) \cdot \frac{\partial \ln p(x_n,z_n|\theta)}{\partial \theta} \right] dz_n = 0$$

有简单的封闭形式的解，不需迭代。

- 总体的迭代EM算法：
  - E步骤： $\theta_0 \leftarrow \theta$，计算$p(z_n|x_n,\theta_0)$；
  - M步骤：利用上面的方程求解θ；
  - 循环（E, M）步骤，直至收敛。

EM算法可以保证收敛到一个稳定点，不过不能保证收敛到全局的极大值点

与K-均值法的联系

- 在高斯混合模型中（假设$\boldsymbol{\Sigma}_k = \epsilon\mathbf{I}$），

$$p(\mathbf{x}|z_k = 1, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{2\pi\epsilon}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right\}$$

- 则在E步骤中，

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right\}}{\sum_k \pi_k \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right\}}$$

  - 当ϵ趋向于0时，离$\mathbf{x}_n$最近的$\boldsymbol{\mu}_k$所对应的$\gamma(z_{nk}) = 1$，其余为0，这其实就是K-均值法中对r_{nk}的求解（E步骤）。

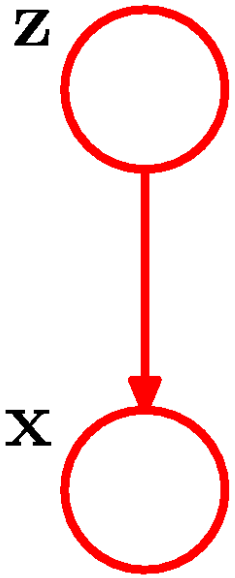
- 对$\boldsymbol{\mu}_k$的求解（M步骤）与K-均值法中的相同。

- 因此，K-均值法可看做是一种硬边界的高斯混合模型。

与K-均值法的比较

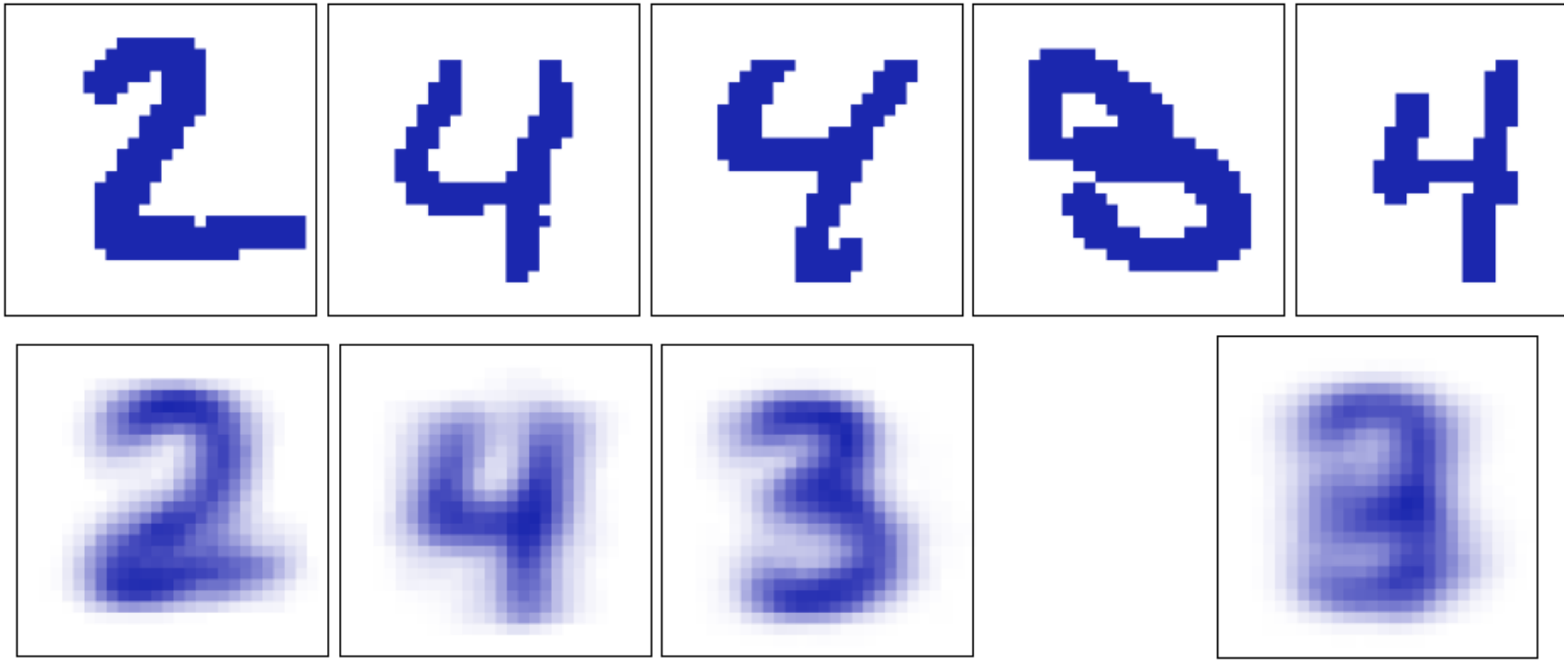
- 高斯混合模型假设数据点是高斯分布的（可以是椭圆），这是一个限制较少的假设；而K-均值法假设它们是圆形的。
- 高斯混合模型有概率信息，可描述混合类。
- 高斯混合参数多，求解难度比K-均值法大。

离散x分布



- 也可以考虑离散的x分布，例如其分量$x_i = \{0,1\}$
- 在z固定的条件下，进一步假设x分量的分布是独立的。

  这类似于朴素贝叶斯，但此时没有已知标签。
- 通过EM算法可以从数据中学习z，并计算（预测）任一新的x属于任一分类（隐藏 z 值）的几率。
- 例子：抛铜钱（各种类型，如狄青钱）。
- 应用：数字识别的无监督学习EM算法。



- 伯努利混合模型。

Bishop Fig. 9.10

其它聚类方法（略）

- 层次聚类（Hierarchical clustering）
- InfoMap
- DBSCAN
- 图论中的团体检测（community detection）
- …

应用例子1：

http://www.dataivy.cn/blog/ad_clustering_with_keans/

基于K-Means的广告效果聚类分析.mht

问题描述

- 某企业由于投放的广告渠道比较多，需要对其做广告效果分析以实现有针对性的广告效果测量和优化工作。
- 数据记录数：889
- 数据预处理：
  - 缺失值替换：例如，替换为均值；
  - 字符串分类转换为整数分类；
  - 数据标准化：Min-Max标准化

- 数据维度：

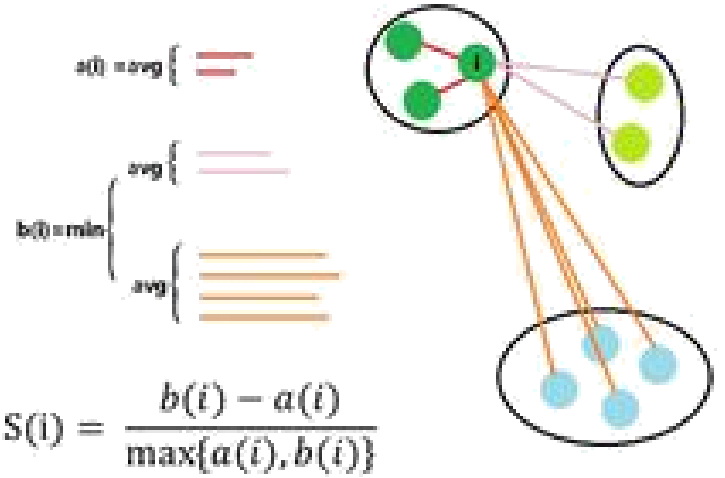
1. 日均UV：每天的平均独立访客，从一个渠道中带来的一个访客即使一天中到达多次都统计为1次
2. 平均注册率：日均注册的用户数量/平均每天的访问量
3. 平均搜索量：平均每个访问的搜索次数
4. 访问深度：总页面浏览量/平均每天的访问量
5. 平均停留时间：总停留时间/平均每天的访问量
6. 订单转化率：总订单数量/平均每天的访问量
7. 投放总时间：每个广告媒介在站外投放的天数
8. 素材类型：广告素材类型，包括jpg、gif、swf、sp
9. 广告类型：广告投放类型，包括banner、tips、横幅、通栏、暂停以及不确定（不知道到底是何种形式）
10. 合作方式：广告合作方式，包括roi、cpc、cpm和cpd
11. 广告尺寸：每个广告投放的尺寸大小，包括140*40、308*388、450*300、600*90、480*360、960*126、900*120、390*270
12. 广告卖点：包括打折、满减、满赠、秒杀、直降、满返。

K值的确定

- K值的确定一直是K-Means算法的关键，而由于K-Means是一个非监督式学习，因此没有所谓的“最佳”K值。但是，从数据本身的特征来讲，最佳K值对应的类别下应该是类内距离最小化并且类间距离最大化。有多个指标可以用来评估这种特征，比如平均轮廓系数、类内距离/类间距离等都可以做此类评估。
- 平均轮廓系数：

$$S(i) = \frac{b(i)-a(i)}{\max[a(i),b(i)]}$$

  - 其中$a(i)$与$b(i)$分别是第i类的类内平均距离及其与最近类的类间平均距离。



- *******K value and silhouette summary:**********
- [2. 0.46692821]
- [3. 0.54904646]
- [4. 0.56968547]
- [5. 0.48186604]
- [6. 0.45477667]
- [7. 0.48204261]
- [8. 0.50447223]
- [9. 0.52697493]]
- Best K is: 4 with average silhouette of 0.5697

如果平均轮廓得分值小于0，意味着聚类效果不佳；

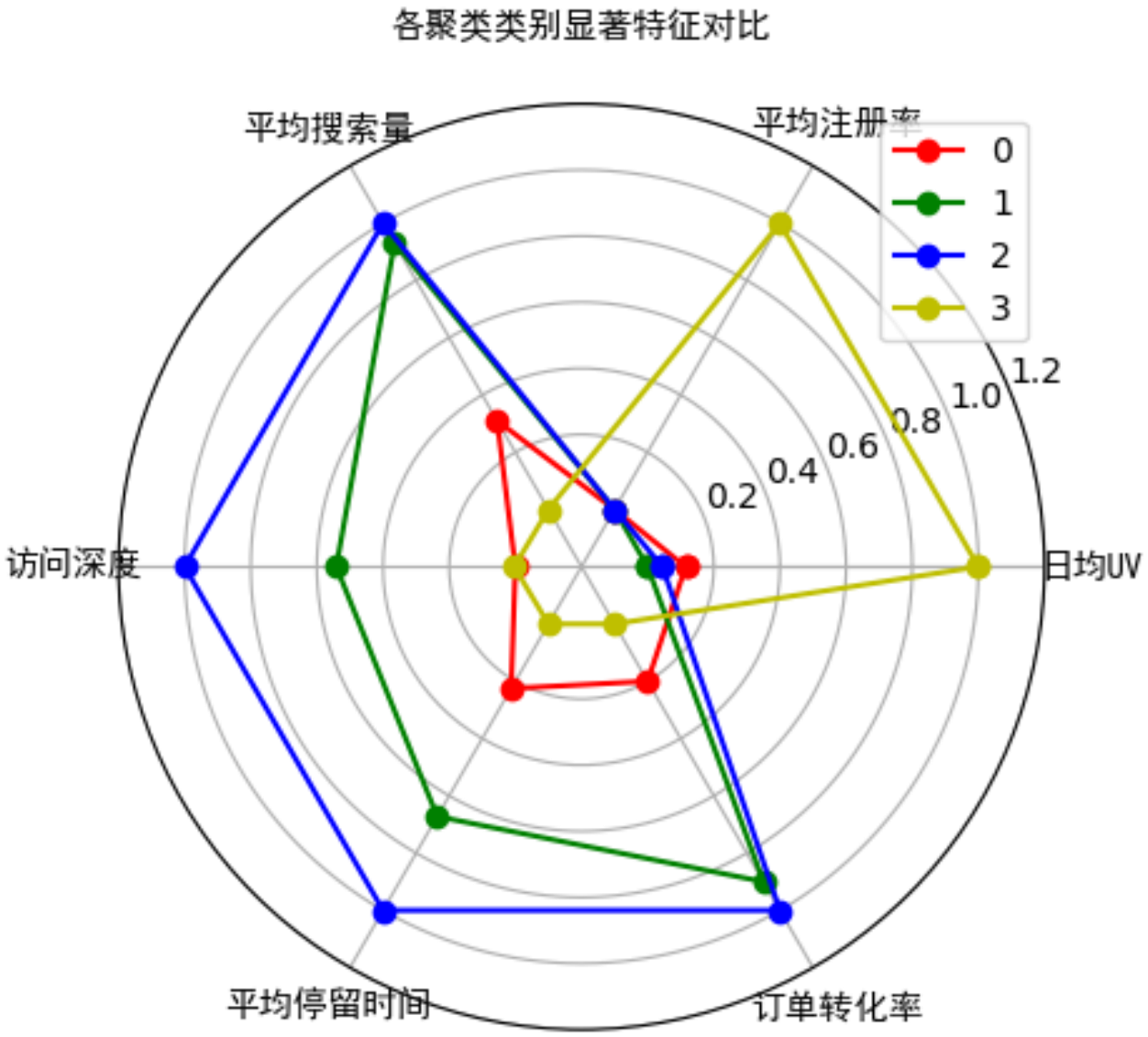
如果值大于0且小于0.5，那么说明聚类效果一般；

如果值大于0.5，则说明聚类效果比较好。

聚类结果的特征分析

- clusters 0 1 2 3
- counts 411 297 27 154
- percentage 0.46 0.33 0.03 0.17
- 日均UV 1369.81 1194.69 1263.03 2718.7
- 平均注册率 0.003 0.003 0.003 0.005
- 平均搜索量 0.082 0.144 0.151 0.051
- 访问深度 0.918 5.728 9.8 0.948
- 平均停留时间 165.094 285.992 374.689 104.14
- 订单转化率 0.009 0.016 0.017 0.007
- 投放总时间 8.462 8.57 7.996 8.569
- 素材类型 swf jpg swf jpg
- 广告类型 不确定 不确定 通栏 banner
- 合作方式 cpc cpc cpc cpc

- 聚类0（46%）：效果比较平庸；
- 聚类1（33%）：除了注册转化率较低外在各指标表现较好，是规模较大且综合效果较好的媒体。
- 聚类2（3%）：与1类似，并且表现更好，属于少量的“精英”类渠道。
- 聚类3（17%）：日均UV和平均注册率突出，其它差，是一类“引流”+“拉新”的渠道；



应用例子2：

David McKay, Robert F. Moran, Daniel M. Dawson, John M. Griffin, Simone Sturniolo, Chris J. Pickard, Andrew J. Berry, and Sharon E. Ashbrook.

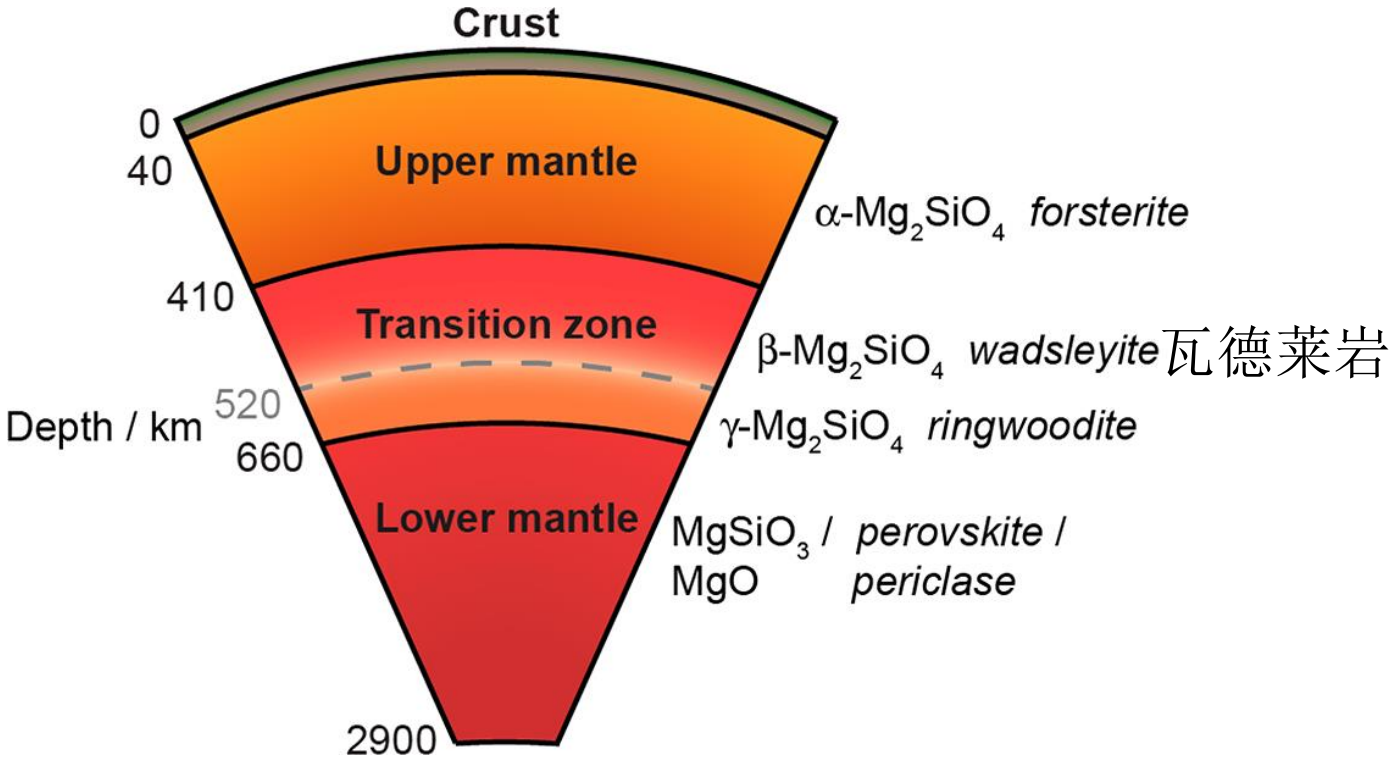
A Picture of Disorder in Hydrous Wadsleyite under the Combined Microscope of Solid-State NMR Spectroscopy and Ab Initio Random Structure Searching。

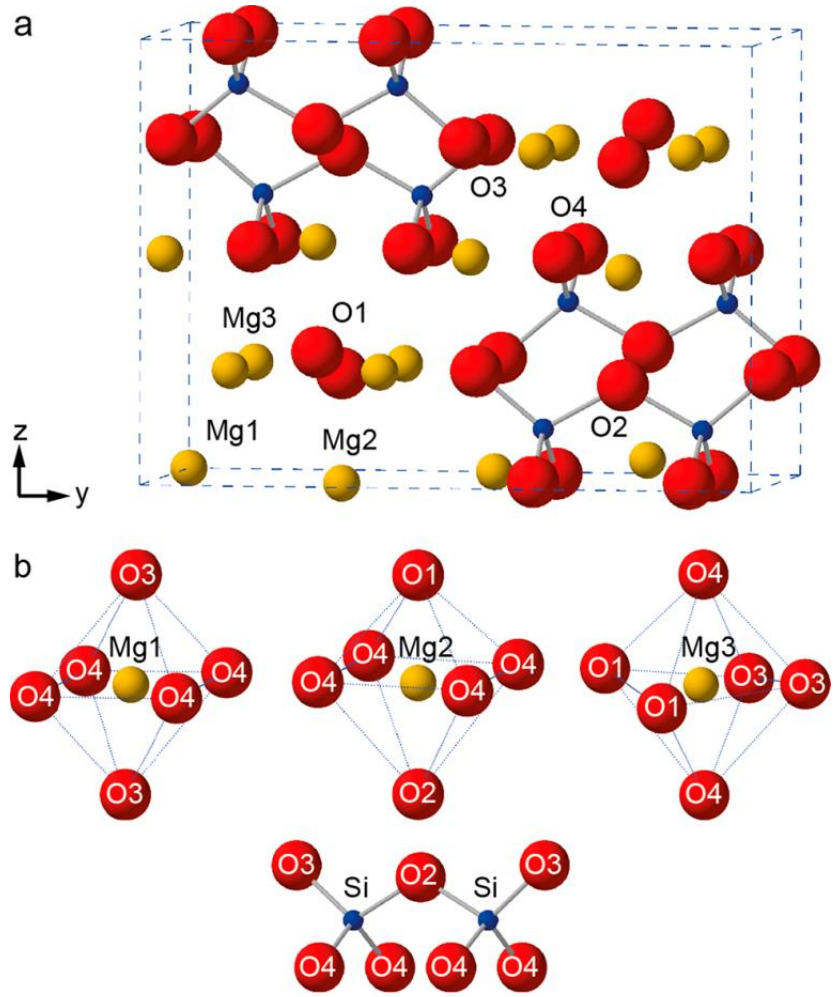
J. Am. Chem. Soc. 141, 3024-3036 (2019).

学习资料：jacs3024.pdf, jacs3024s.pdf

问题描述

- 地壳中β-Mg_2SiO_4很重要。
- 存在多种质子化可能。





研究方案

(1) ab initio random structure searching (AIRSS)

(2) DFT geometry optimization during the AIRSS process

(3) K-means clustering。

(4) DFT optimization with increased accuracy on the selected structures

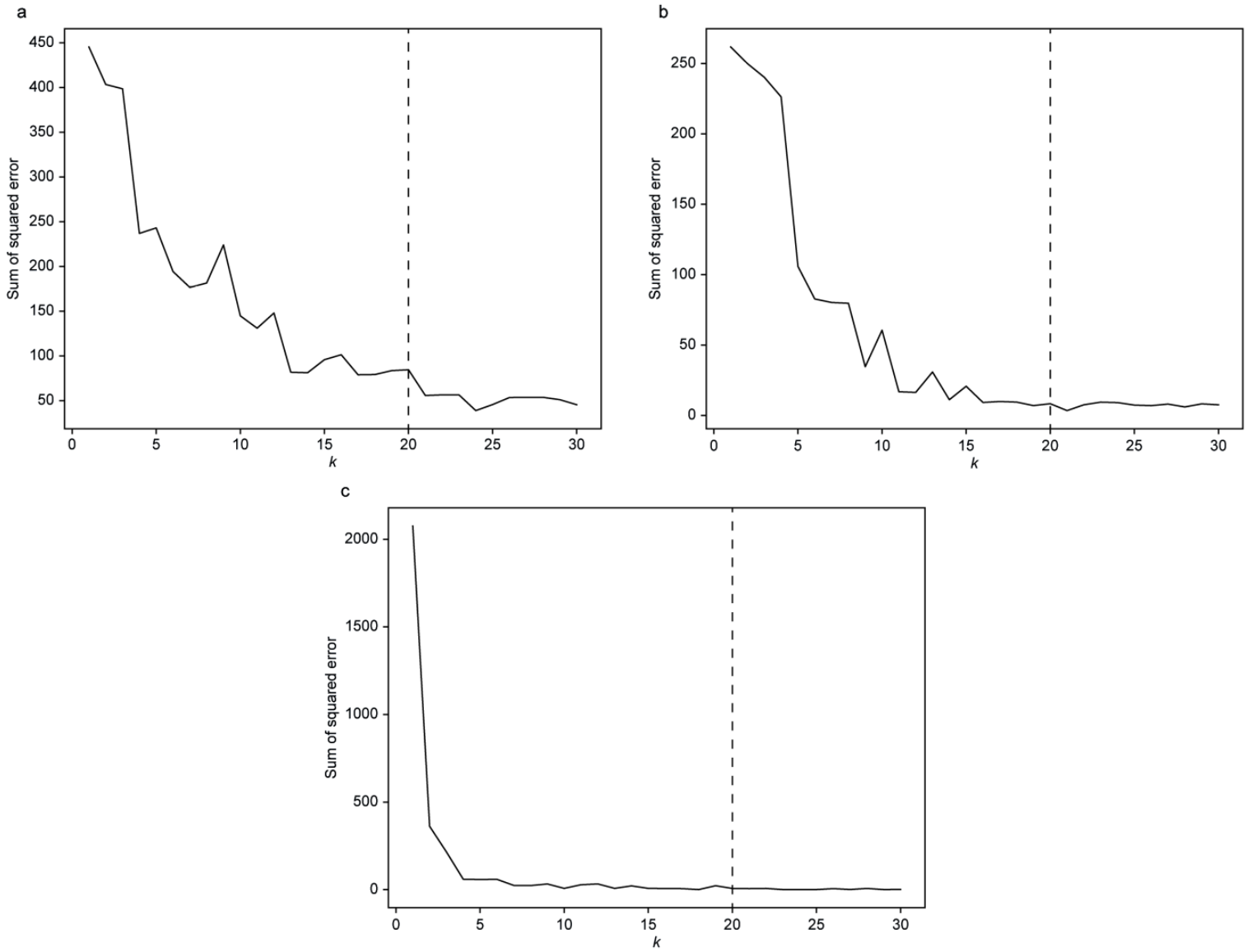
(5) GIPAW NMR calculations

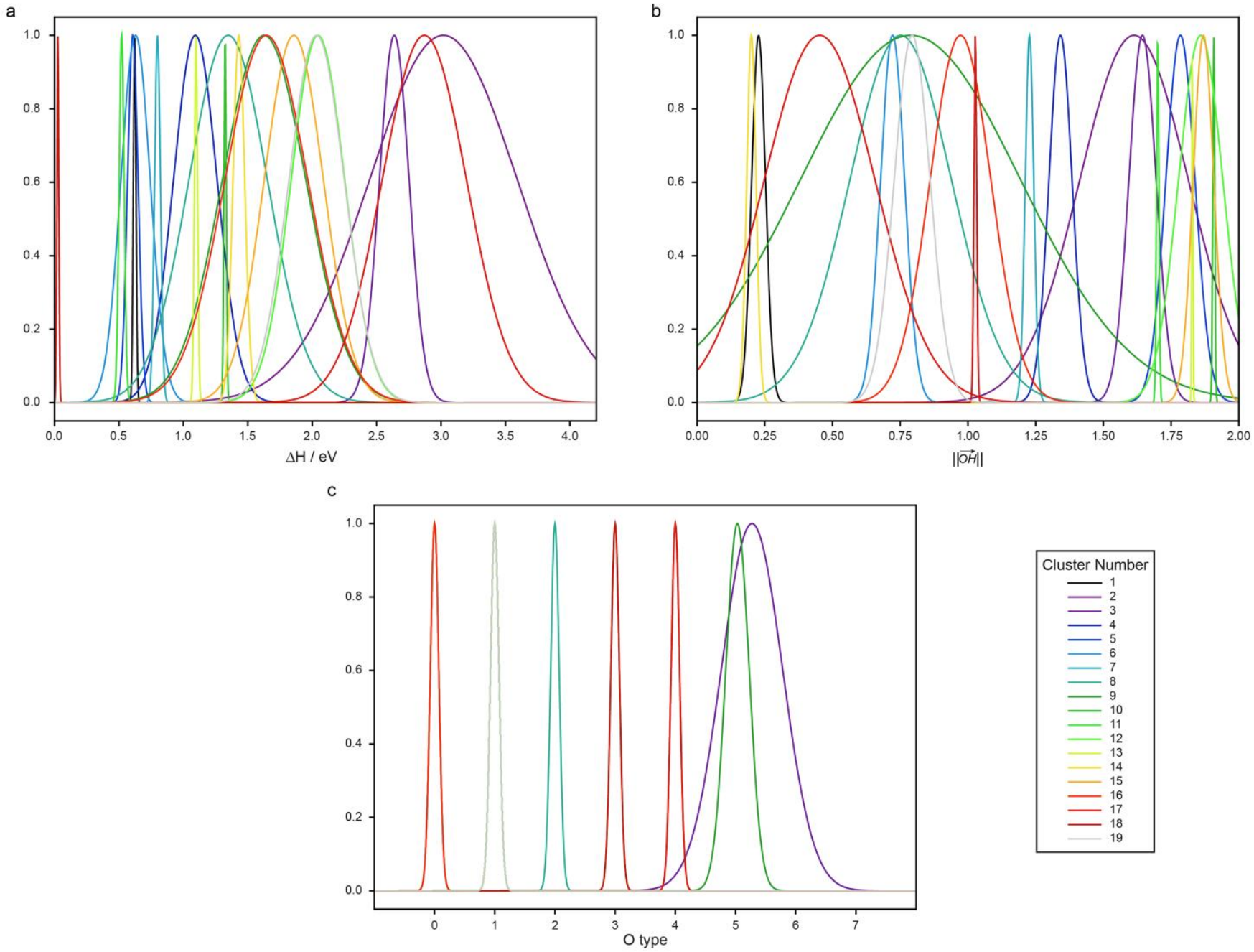
Of the original 1287 AIRSS-generated candidates, K-means clustering identified a total of 88 candidate structures for further study.

方法：K-均值法

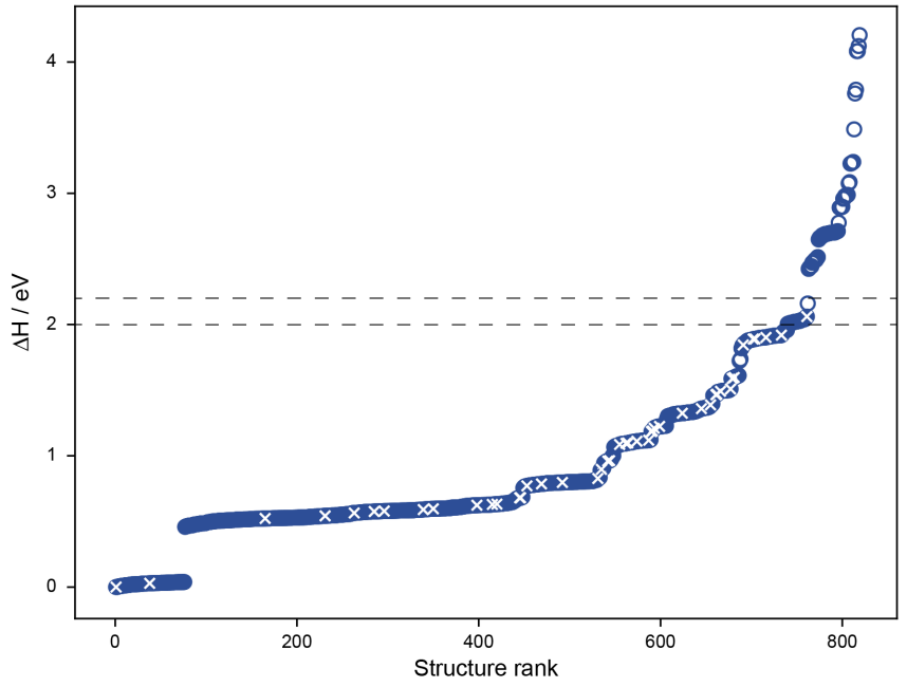
- 特征：
  - the relative enthalpy, ΔH;
  - hydroxyl O type;
  - vacancy type;
  - H···H distance;
  - H···vacancy distance;
  - O–H bond length;
  - OH···O hydrogen-bond length and the magnitude of the combined hydroxyl orientation vector

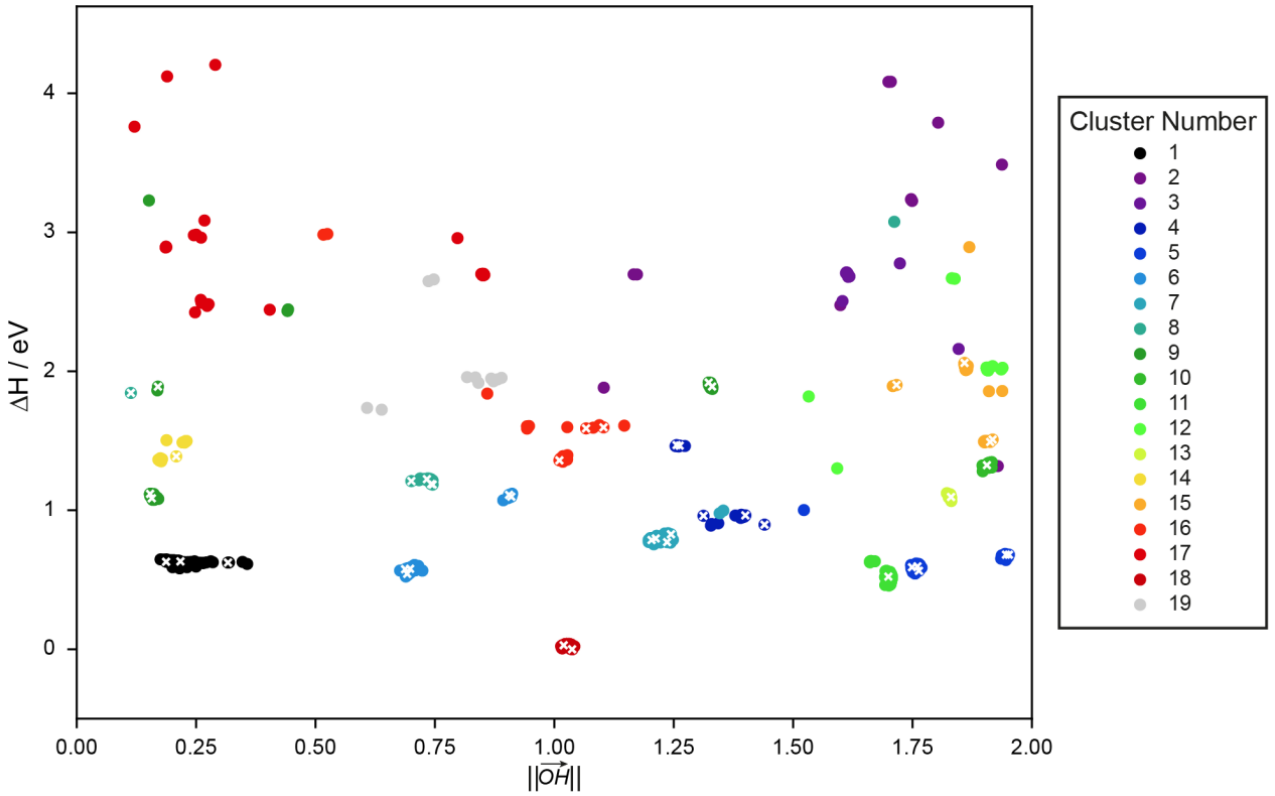
- K的选择：plotting the sum of squared errors within each cluster against K



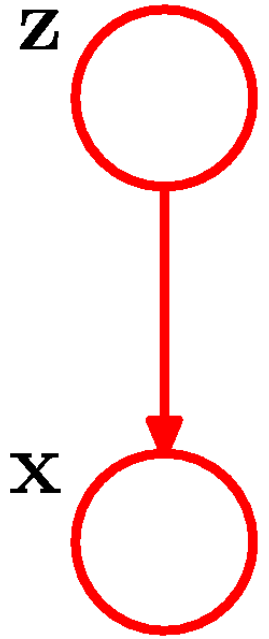

a
b
c
ΔH / eV
||$\overrightarrow{OH}$||
O type
Cluster Number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

利用聚类结果指导结构的选择





小结



- K-均值法（K-means）
  - 每个数据点被分到离自己最近的聚类中心所在的类。
  - 畸变函数（Distortion function）：

$$J(\{r_{nk}\},\{\boldsymbol{\mu}_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

- 高斯混合模型
  - 生成式模型
  - 假设分布是由多个高斯分布混合而成：$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - EM算法：交替计算数据点组别概率$\gamma(z_{nk})$与组别特征$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$

Scikit-Learn相关内容

https://scikit-learn.org/
https://sklearn.apachecn.org/

- **2.3. Clustering**
- **2.3.2. K-means**
  - **cluster.KMeans**
- **2.1. Gaussian mixture models**
  - **mixture.GaussianMixture**

- Reference:
  - Bishop 9.1-9.3.3；
  - Elements 13。
  - 实战 10。
  - 吴恩达 13。

扩展阅读：

- https://zhuanlan.zhihu.com/p/26144586

机器理解大数据的秘密：聚类算法深度详解.mht

- https://blog.csdn.net/u011511601/article/details/81951939

【聚类】五种主要聚类算法.mht

- http://www.sohu.com/a/215759232_642762

干货丨EM算法原理总结.mht

- http://www.dataivy.cn/blog/ad_clustering_with_keans/

基于K-Means的广告效果聚类分析.mht

- https://baijia.baidu.com/s?id=1595977285675528555

手把手：扫描图片又大又不清晰？这个Python小程序帮你搞定！.mht

扫描笔记去噪-代码：noteshrink-master.zip

- https://36kr.com/p/1488545128907139

人的情绪岂止6种？Google发布大规模数据集GoEmotions，情感类别提升到28种.mhtml

- 1992 年提出的六种基本情绪：愤怒（anger）、惊讶（surprise）、沮丧（disgust）、快乐（joy）、恐惧（fear）和悲伤（sadness）。
- 新的分类包括12种积极情绪、11种消极情绪、4种模棱两可的情绪类别和1种中立情绪

Positive		Negative		Ambiguous
admiration 👏	joy 😃	anger 😡	grief 😢	confusion 😕
amusement 😂	love ❤️	annoyance 😒	nervousness 😬	curiosity 🤔
approval 👍	optimism 🤞	disappointment	remorse 😔	realization 💡
caring 🤗	pride 😌	disapproval 👎	sadness 😞	surprise 😲
desire 😍	relief 😅	disgust 🤮		
excitement 🤩		embarrassment 😳		
gratitude 🙏		fear 😨		

谢谢大家！