NLP 南京大学
自然语言处理研究组
Natural Language Processing Group

南京大学
NANJING UNIVERSITY

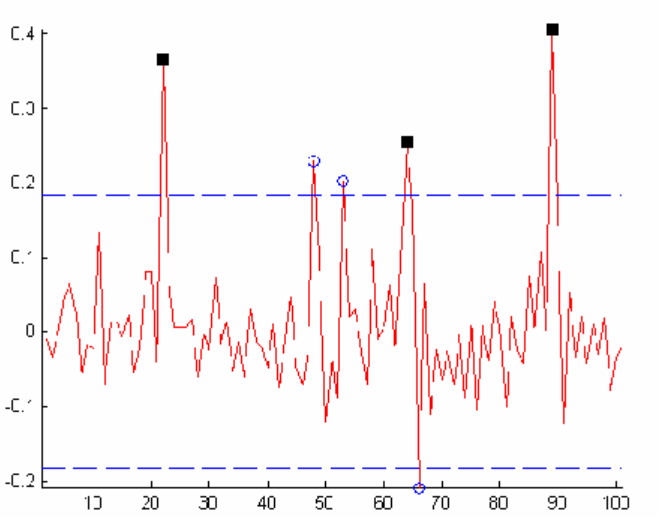
无监督学习

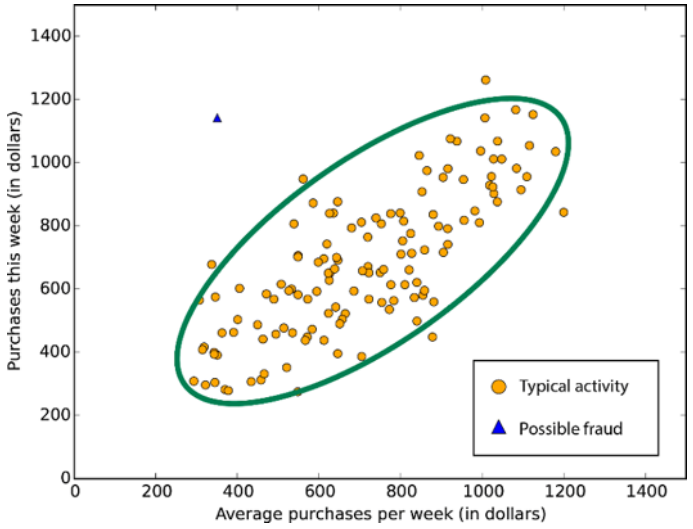
黄书剑

- 无监督学习
- 聚类分析
  - k均值聚类
- 关联规则
  - Apriori
- **异常检测**

异常检测

- **发掘数据中包含的不一致性、不规律（离群点）**
  - 检测异常行为（系统故障、欺诈、入侵等）
  - 检测异常的状态（生态系统失调、流感疫情爆发等）





异常的成因

- 数据来源于不同的类别
- 数据自然变异
- 数据测量和收集的误差
- ……

- 异常v.s.噪音
  - 噪音不一定导致异常的结果
  - 噪音的观察价值较小（倾向于随机发生）

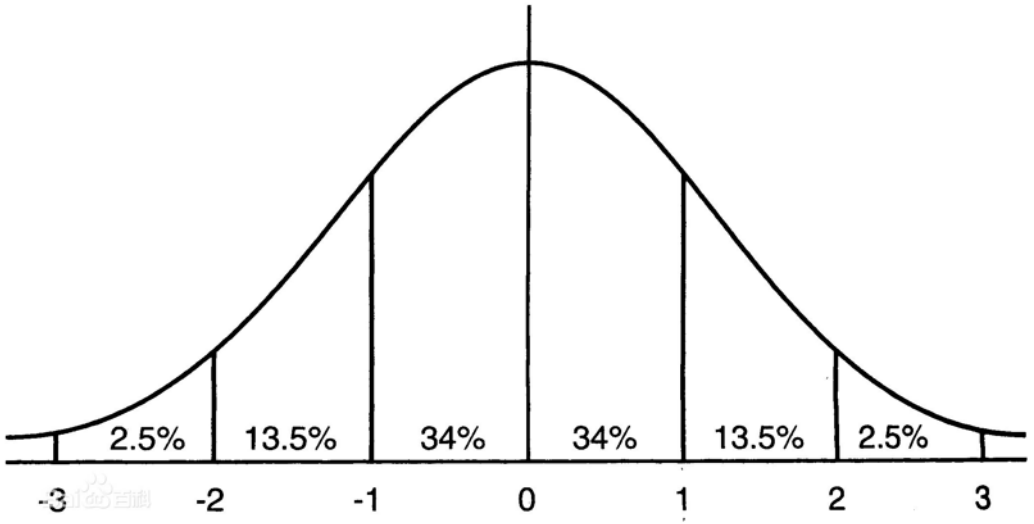
异常的检测

- 有监督的检测
- 无监督的检测

- 基于模型的技术
  - 正态分布
- 基于邻近度的技术
  - k-近邻
- 基于密度的技术

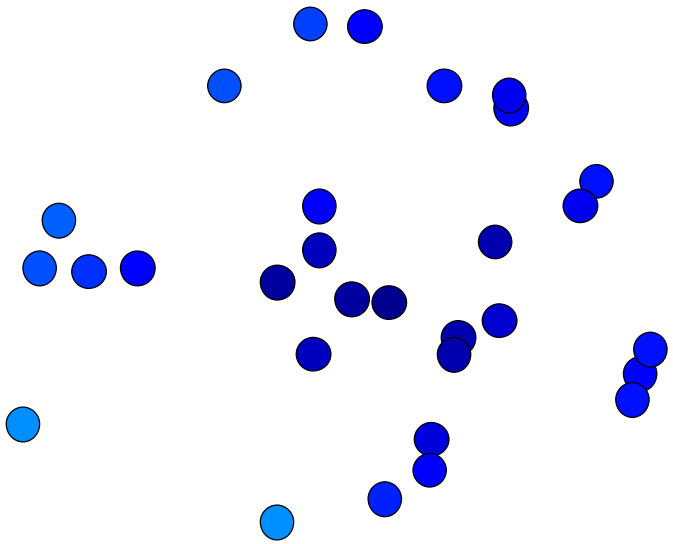
基于正态分布的离群点预测

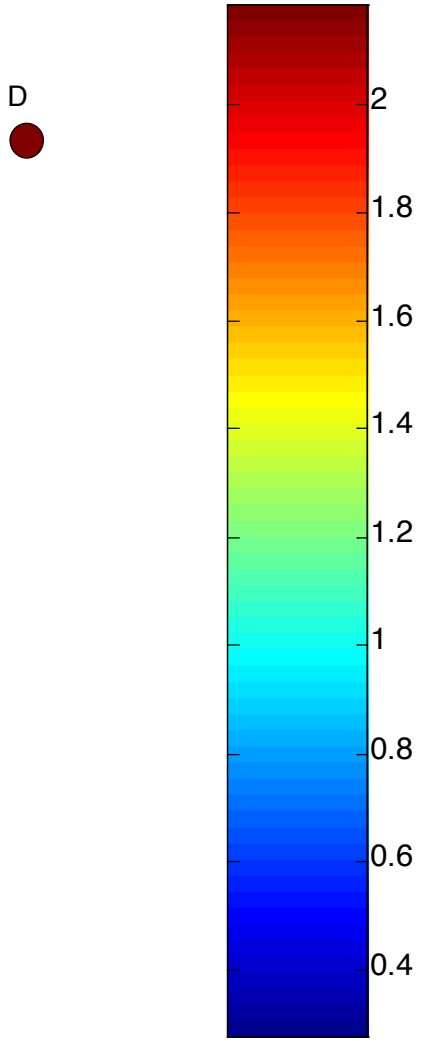
- 标准正态分布，是以0为均值、以1为标准差的正态分布，记为N(0，1)
  - 离群程度为其出现的概率



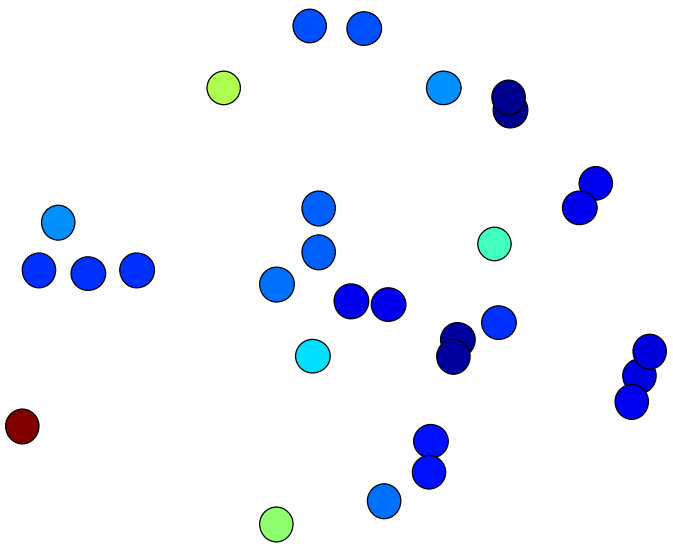
基于邻近度的离群点检测

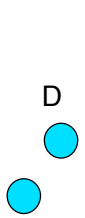
- 离群程度定义为其到k近邻的距离
  - k=1

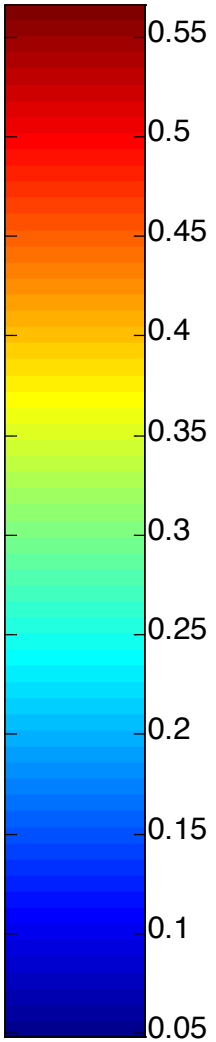




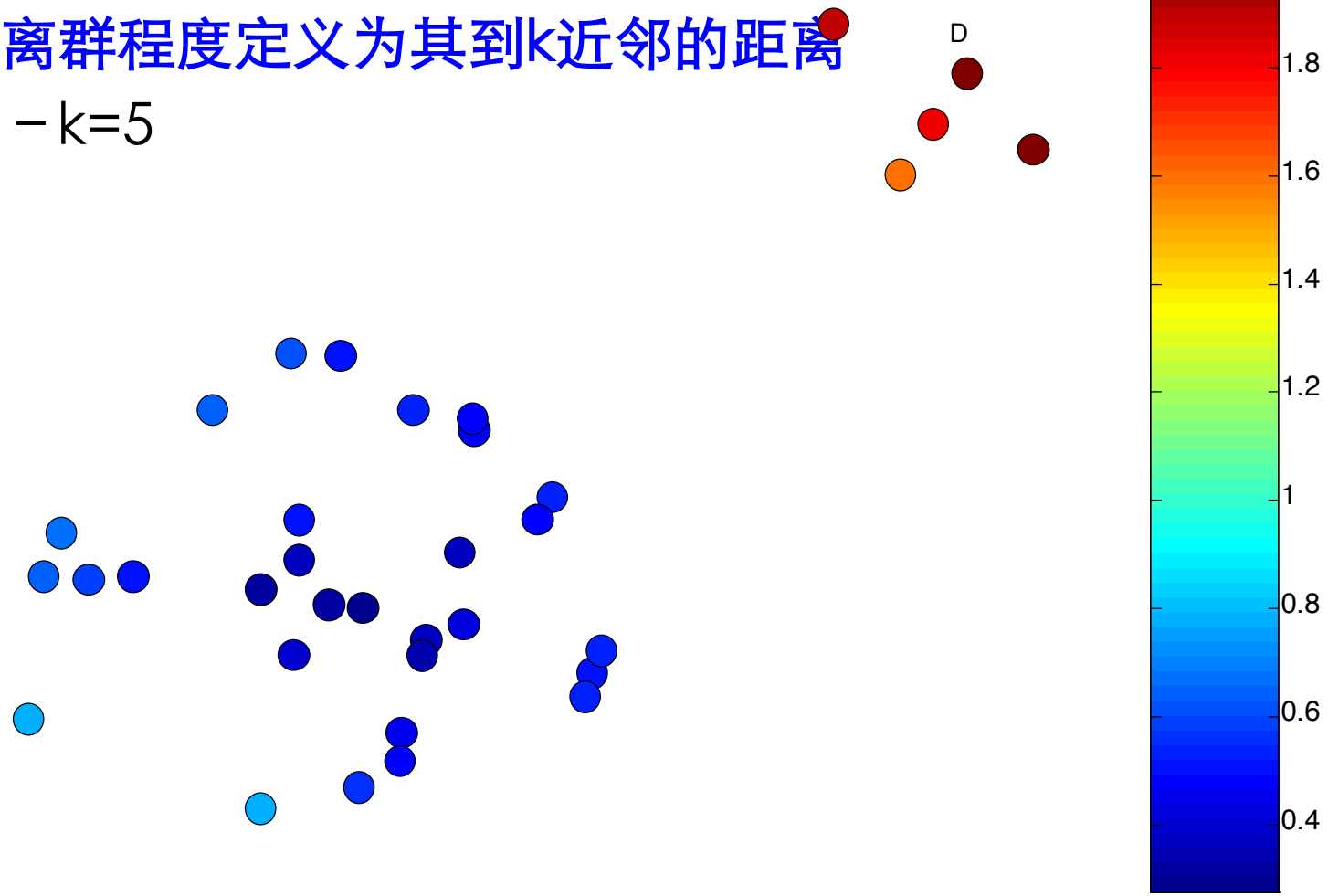
- 离群程度定义为其到k近邻的距离
  - k=1



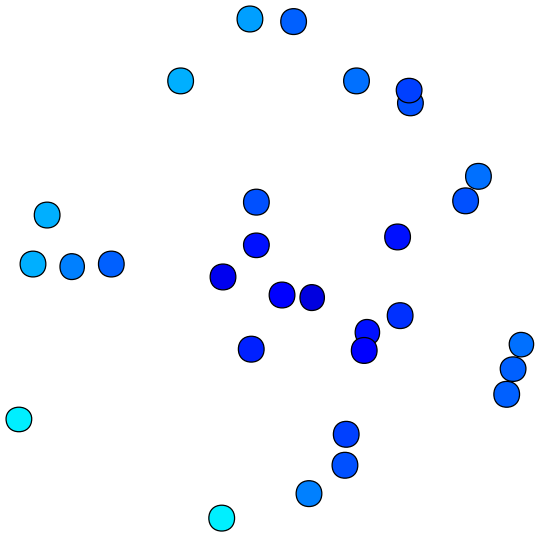


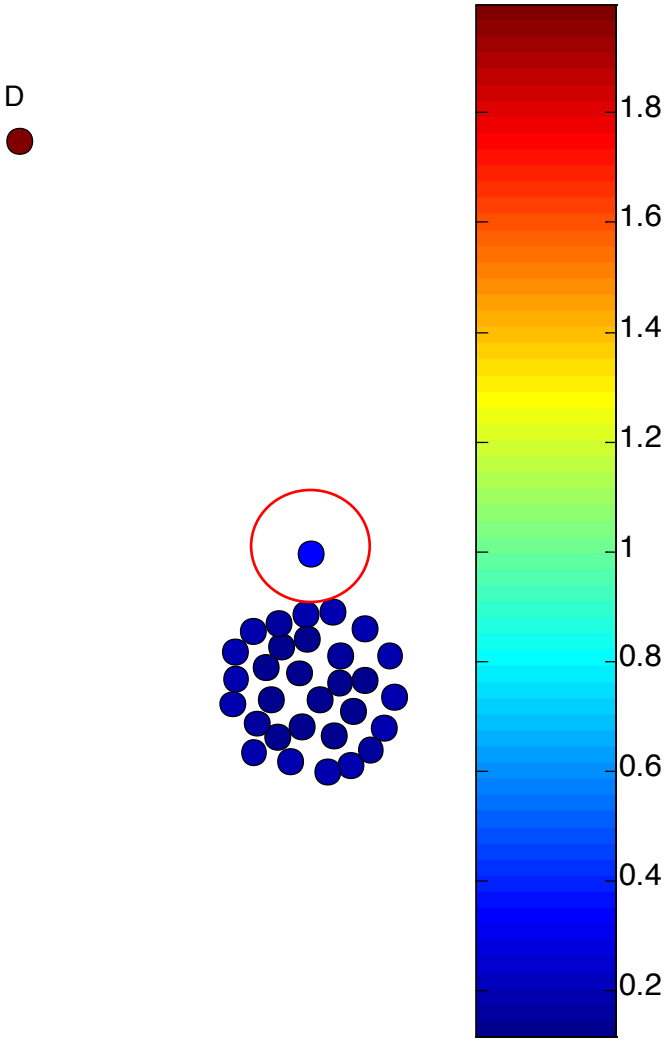


- 离群程度定义为其到k近邻的距离
  - k=5



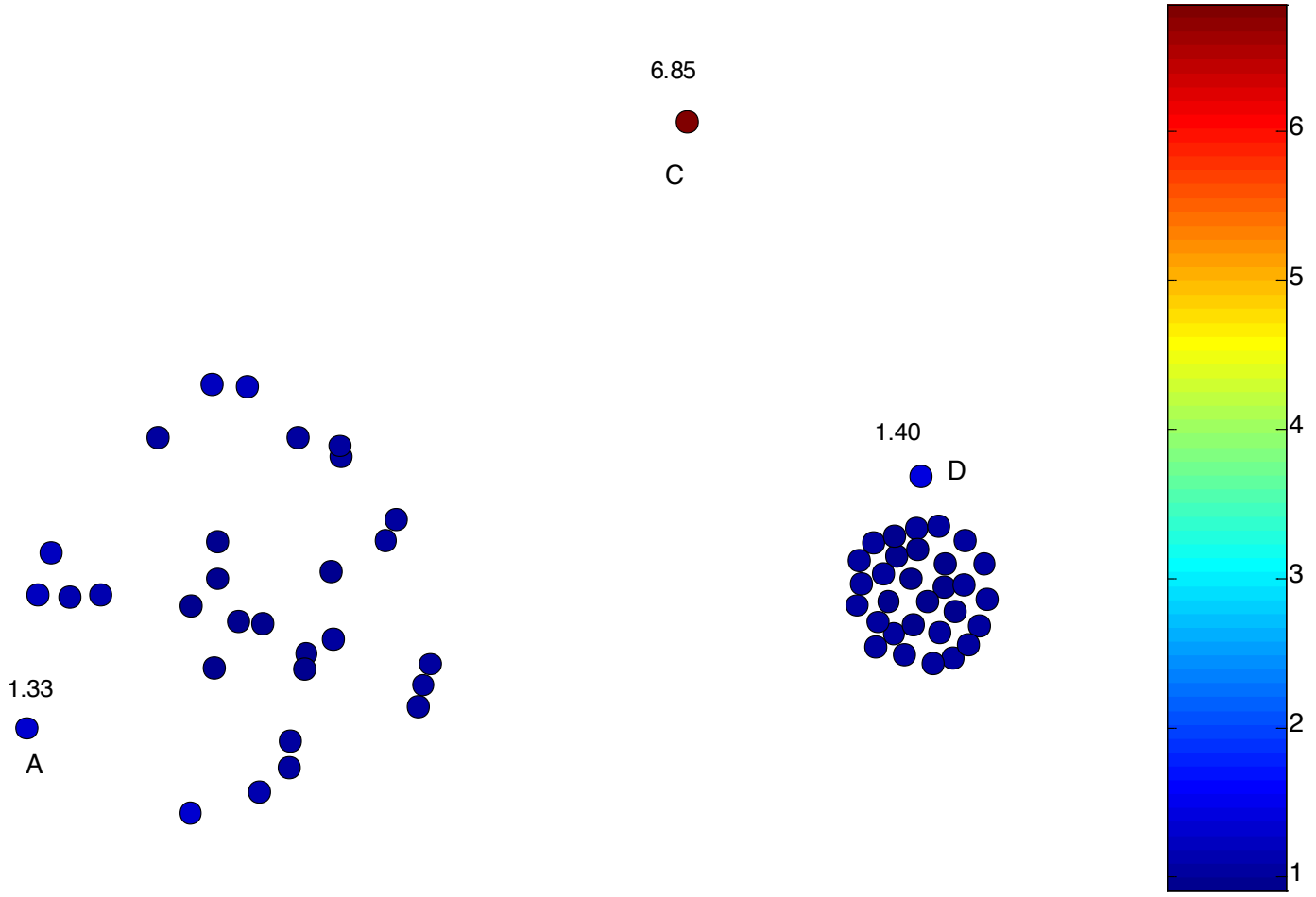
- 不同密度数据





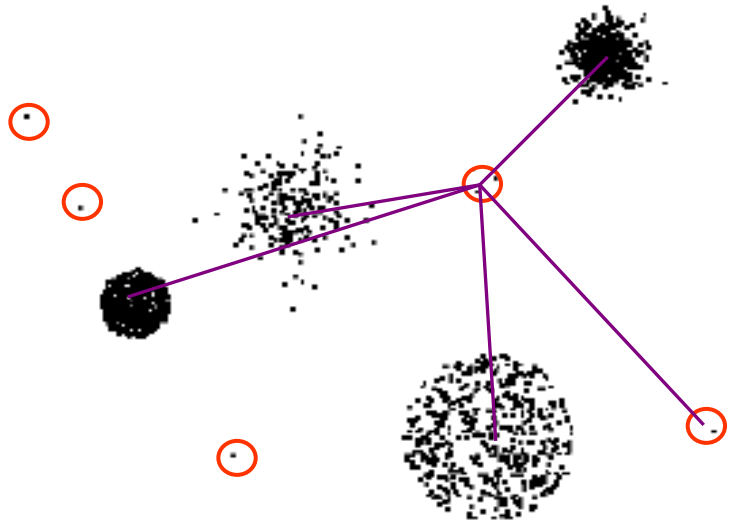
基于密度的离群点检测

$$average\ relative\ density(\mathbf{x},k) = \frac{density(\mathbf{x},k)}{\sum_{\mathbf{y}\in N(\mathbf{x},k)} density(\mathbf{y},k)/|N(\mathbf{x},k)|}$$



基于聚类的技术

- 综合聚类结果、密度等



练习五

- 尝试实现一个简单的Apriori算法，比较不同实现的性能差距
- 尝试观察原有数据中的异常分布

参考资料

- 本章大部分内容来源于以下两个课程的相关部分：
  - Introduction to Data Mining (Second Edition) https://www-users.cs.umn.edu/~kumar001/dmbook/index.php
  - Data Mining: Concepts and Techniques, 3rd ed. https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm