

# 第13章：无监督学习概论

导师：谈晨

---

# 无监督学习

## Unsupervised Learning

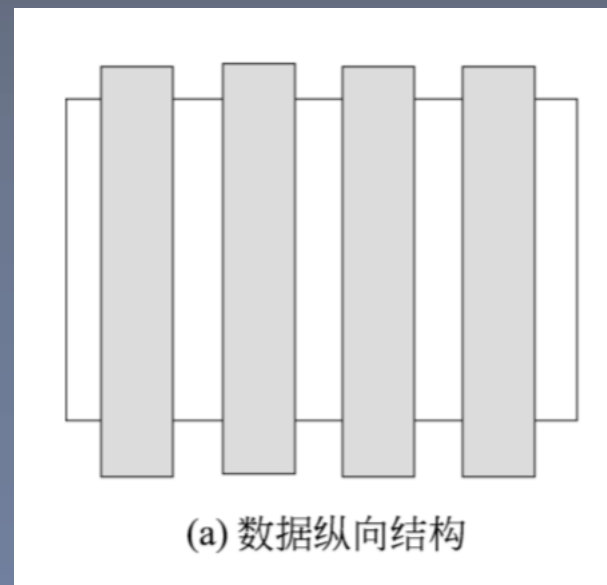
- 使用无标注数据  $U = \{x_1, x_2, \dots, x_N\}$       学习或训练，由特征向量组成
- 无监督学习的模型是函数  $z = g_\theta(x)$  ， 条件概率分布  $P_\theta(z|x)$  ， 或条件概率分布  $P_\theta(x|z)$
- 假设训练数据集由N个样本组成，每个样本是一个M维向量。训练数据可以由一个矩阵表示，每一行对应一个特征，每一列对应一个样本

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix}$$

# 无监督学习

## Unsupervised Learning

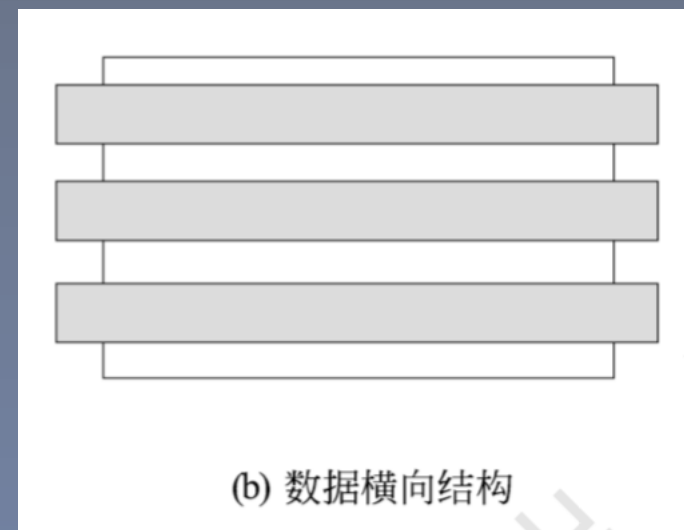
- 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构。假定损失最小的压缩得到的结果就是最本质的结构。
- 考虑发掘数据的纵向结构，  
把相似的样本聚到同类，  
即对数据进行聚类



# 无监督学习

## Unsupervised Learning

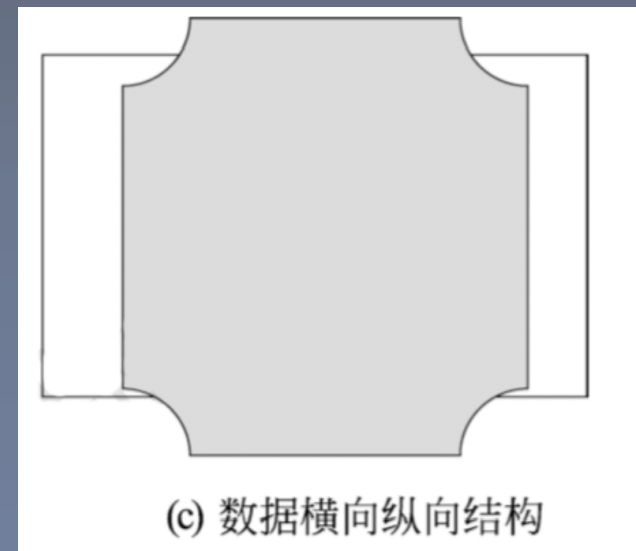
- 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构。假定损失最小的压缩得到的结果就是最本质的结构。
- 考虑发掘数据的横向结构，  
把高维空间的向量转换为  
低维空间的向量，  
即对数据进行降维。



# 无监督学习

## Unsupervised Learning

- 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构。假定损失最小的压缩得到的结果就是最本质的结构。
- 同时考虑发掘数据的  
纵向与横向结构，假设数据由含有  
隐式结构的概率模型生成得到，  
从数据中学习该概率模型。



# 聚类

## Clustering

---

- 聚类（clustering）是将样本集合中相似的样本（实例）分配到相同的类，不相似的样本分配到不同的类。
- 聚类时，样本通常是**欧氏空间**中的向量，类别不是事先给定，而是从数据中**自动发现**，但类别的个数通常是事先给定的。样本之间的相似度或距离 由应用决定。
- 如果一个样本只能属于一个类，则称为**硬聚类**（hard clustering）
- 如果一个样本可以属于多个类，则称为**软聚类**（soft clustering）

# 聚类

## Clustering

---

- 硬聚类时，每一个样本属于某一类

$$z_i = g_{\theta}(x_i), i = 1, 2, \dots, N$$

- 软聚类时，每一个样本依概率属于每一个类

$$P_{\theta}(z_i | x_i), i = 1, 2, \dots, N$$



# 降维

## Dimensionality Reduction

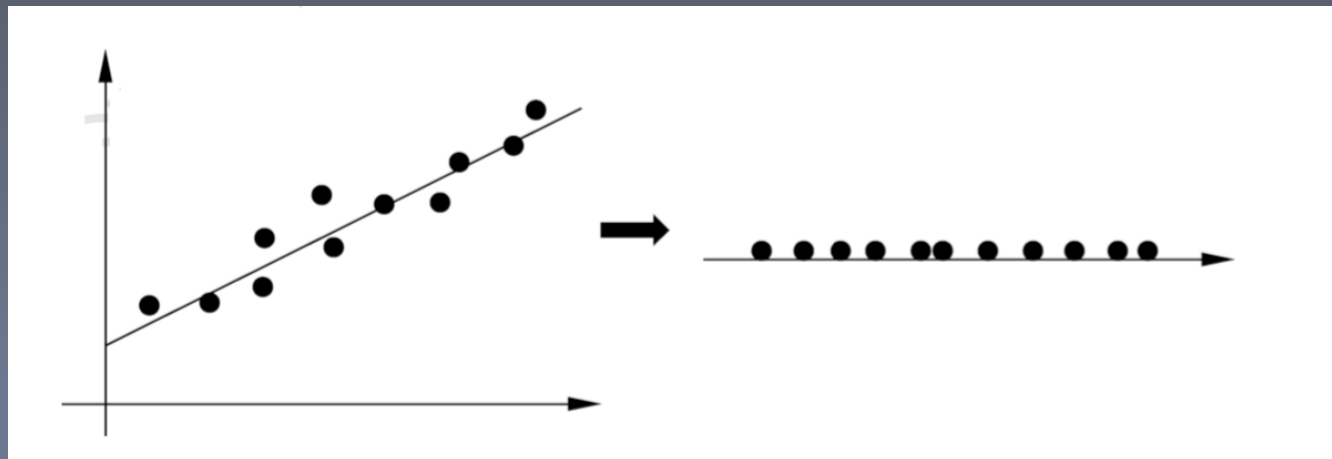
- 降维 (dimensionality reduction) 是将训练数据中的样本 (实例) 从高维空间转换到低维空间。
- 假设样本原本存在于低维空间, 或者近似地存在于低维空间, 通过降维则可以更好地表示样本数据的**结构**, 即更好地表示样本之间的**关系**。
- 高维空间通常是高维的欧氏空间, 而低维空间是低维的欧氏空间或者流形 (manifold)。
- 从高维到低维的降维中, 要保证样本中的**信息损失最小**。



# 降维

## Dimensionality Reduction

- 降维有**线性**的降维和**非线性**的降维。



- 二维空间的样本存在于一条直线的附近，可以将样本从二维空间转换到一维空间。通过降维可以更好地表示样本之间的关系。

# 降维

## Dimensionality Reduction

- 假设输入空间是欧氏空间  $X \subseteq \mathbf{R}^d$ ，输出空间也是欧氏空间  $Z \subseteq \mathbf{R}^{d'}, d' \ll d$ ，后者的维数低于前者的维数。降维的模型是函数  $z = g_{\theta}(x)$

- 其中  $x \in X$  是样本的高维向量， $z \in Z$  是样本的低维向量， $\theta$  是参数。

函数可以是线性函数也可以是非线性函数。

- 降维的过程就是学习降维模型的过程。降维时，每一个样本从高维向量转换为低维向量

$$z_i = g_{\theta}(x_i), \quad i = 1, 2, \dots, N。$$

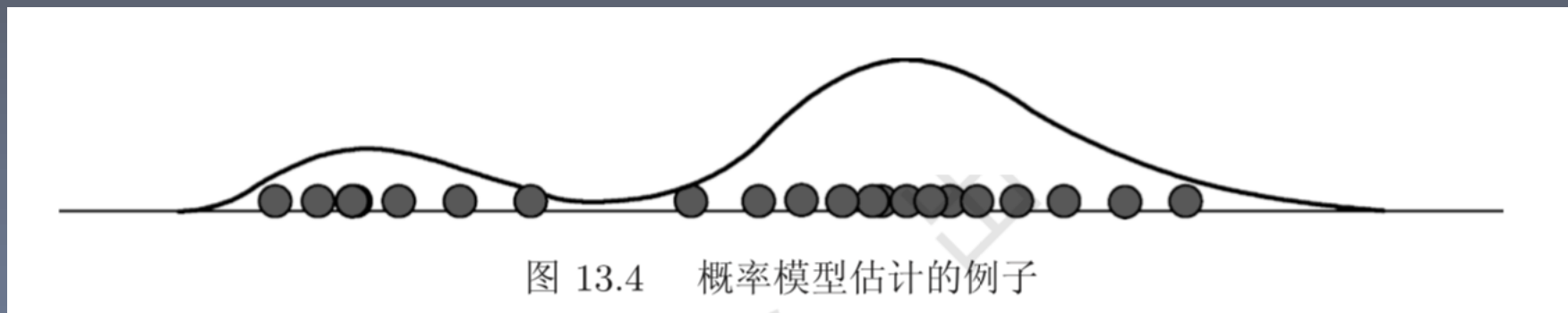
# 概率模型估计

## Estimation of Probability Model

- 假设训练数据由一个概率模型生成，由训练数据学习概率模型的**结构和参数**。
- 概率模型的结构类型，或者说概率模型的集合事先给定，而模型的具体结构与参数从数据中**自动学习**。学习的目标是找到最有可能生成数据的结构和参数。
- 概率模型包括**混合模型、概率图模型**等。
- 概率图模型又包括**有向图模型和无向图模型**。

# 概率模型估计

Estimation of Probability Model



- 假设数据由高斯混合模型生成，学习的目标是估计这个模型的参数。

# 概率模型估计

## Estimation of Probability Model

- 概率模型表示为条件概率分布  $P_{\theta}(x|z)$
- 随机变量 $x$ 表示观测数据，可以是连续变量也可以是离散变量
- 随机变量 $z$ 表示隐式结构，是离散变量
- 随机变量  $\theta$  表示参数
- 模型是混合模型时， $z$ 表示成分的个数
- 模型是概率图模型时， $z$ 表示图的结构

# 概率模型估计

## Estimation of Probability Model

- 概率模型估计是从给定的训练数据  $U = \{x_1, x_2, \dots, x_N\}$  中学习模型  $P_\theta(x|z)$  的结构和参数，计算出模型相关的任意边缘分布和条件分布。
- 注意随机变量 $x$  是多元变量，甚至是高维多元变量
- 软聚类也可以看作是概率模型估计问题。根据贝叶斯公式

$$P(z|x) = \frac{P(z)P(x|z)}{P(x)} \propto P(z)P(x|z)$$

- 假设先验概率服从均匀分布，只需要估计条件概率分布  $P_\theta(x|z)$  。这样，可以通过对条件概率分布  $P_\theta(x|z)$  的估计进行软聚类

# 无监督学习三要素

## 3 Major Points of Unsupervised Learning

- 模型

- 函数  $z = g_{\theta}(x)$  , 条件概率分布  $P_{\theta}(z|x)$  , 或条件概率分布  $P_{\theta}(x|z)$

- 策略

- 目标函数的优化

- 算法

- 迭代算法, 通过迭代达到对目标函数的最优化



# 聚类

## Clustering

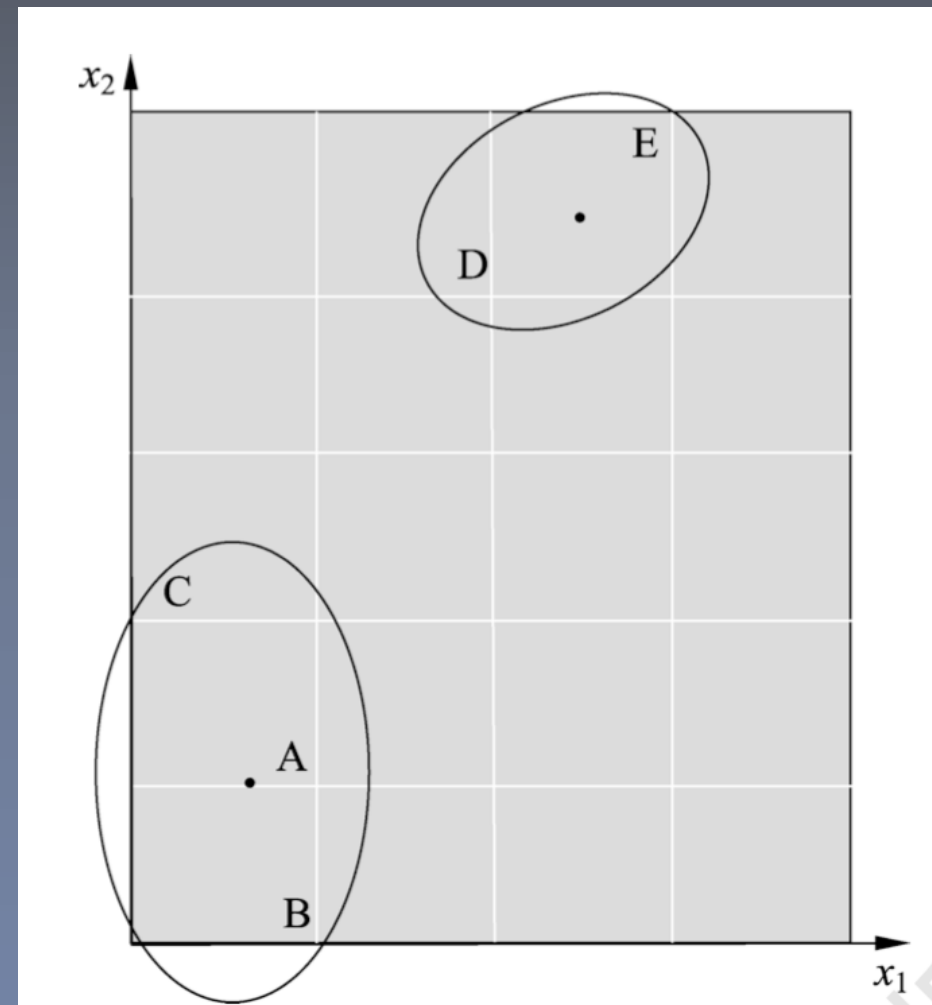
- 有5个样本A、B、C、D、E，每个样本有二维特征 $x_1, x_2$ 。
- 通过聚类算法，可以将样本分配到两个类别中。

|       | A | B | C | D | E |
|-------|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 2 | 3 |
| $x_2$ | 1 | 0 | 2 | 4 | 5 |

# 聚类

## Clustering

- 假设用k均值聚类， $k=2$ 。开始可以取任意两点作为两个类的中心
- 依据样本与类中心的**欧氏距离**的大小将样本分配到两个类中
- 然后计算两个类中样本的**均值**，作为两个类的新的**类中心**
- 重复以上操作，直到两类不再改变
- 最后得到聚类结果，A、B、C为一个类，D、E为另一个类。



# 降维

## Dimensionality Reduction

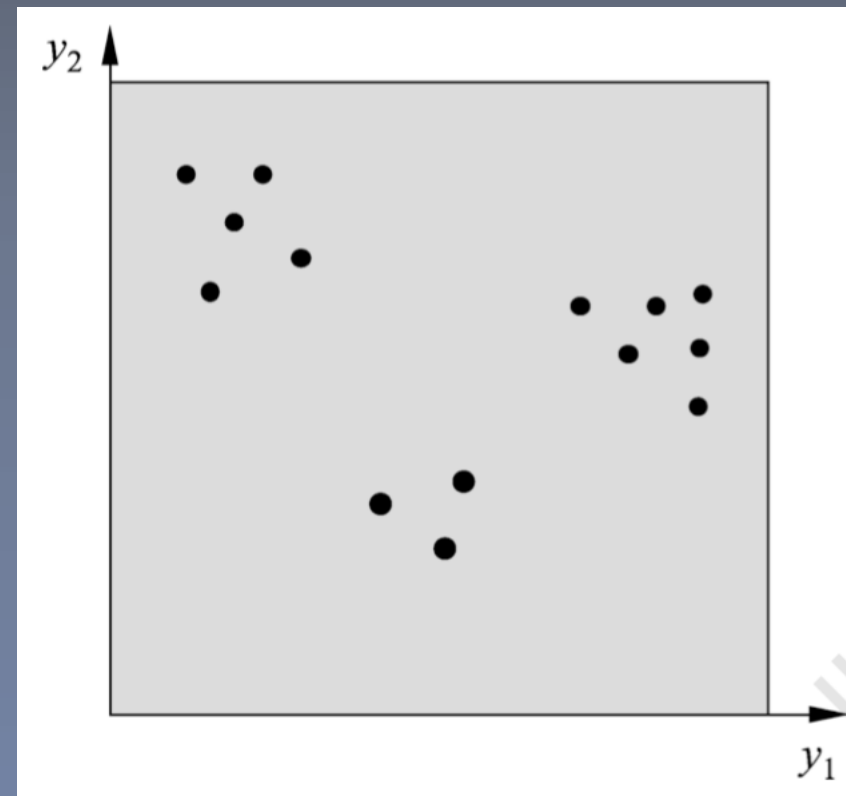
- 给出一个简单的数据集。有14个样本A、 B、 C、 D等，每个样本有9 维特征  $x_1, x_2, \dots, x_9$

|       | A   | B    | C   | D   | ... |
|-------|-----|------|-----|-----|-----|
| $x_1$ | 3   | 0.25 | 2.8 | 0.1 | ... |
| $x_2$ | 2.9 | 0.8  | 2.2 | 1.8 | ... |
| $x_3$ | 2.2 | 1    | 1.5 | 3.2 | ... |
| $x_4$ | 2   | 1.4  | 2   | 0.3 | ... |
| $x_5$ | 1.3 | 1.6  | 1.6 | 0   | ... |
| $x_6$ | 1.5 | 2    | 2.1 | 3   | ... |
| $x_7$ | 1.1 | 2.2  | 1.2 | 2.8 | ... |
| $x_8$ | 1   | 2.7  | 0.9 | 0.3 | ... |
| $x_9$ | 0.4 | 3    | 0.6 | 0.1 | ... |

# 降维

## Dimensionality Reduction

- 由于数据是高维（多变量）数据，很难观察变量的样本区分能力，也很难观察样本之间的关系。
- 对样本集合进行降维（**主成分分析**），  
结果在新的二维实数空间中，  
有二维新的特征 $y_1, y_2$ , 14个样本分布在不同位置。
- 通过降维，可以发现样本可以分为三个类，  
二维新特征由**原始特征**定义。



# 话题分析

## Topic Analysis

---

- 话题分析是文本分析的一种技术。
- 给定一个**文本集合**，话题分析旨在发现文本集合中每个文本的话题，而话题由**单词**的集合表示。
- 注意，这里假设有**足够数量**的文本，如果只有一个文本或几个文本，是不能做话题分析的。
- 话题分析可以形式化为**概率模型估计**问题，或**降维**问题。

# 话题分析

## Topic Analysis

- 给出一个文本数据集合。有6个文本，6个单词，表中数字表示单词在文本中的出现次数。

| 单词 \ 文本 |      |      |      |      |      |      |
|---------|------|------|------|------|------|------|
|         | doc1 | doc2 | doc3 | doc4 | doc5 | doc6 |
| word1   | 1    | 1    |      |      |      |      |
| word2   | 1    |      | 1    |      |      |      |
| word3   |      | 1    | 1    |      |      |      |
| word4   |      |      |      | 1    | 1    |      |
| word5   |      |      |      | 1    |      | 1    |
| word6   |      |      |      |      | 1    | 1    |

# 话题分析

## Topic Analysis

- 对数据进行话题分析，如LDA分析，得到由**单词集合**表示的话题，以及由**话题集合**表示的文本。

表 13.4 话题分析（LDA 分析）的结果

| 单词 \ 话题 | 话题     |        | 文本 \ 话题 | 话题     |        |
|---------|--------|--------|---------|--------|--------|
|         | topic1 | topic2 |         | topic1 | topic2 |
| word1   | 0.33   | 0      | doc1    | 1      | 0      |
| word2   | 0.33   | 0      | doc2    | 1      | 0      |
| word3   | 0.33   | 0      | doc3    | 1      | 0      |
| word4   | 0      | 0.33   | doc4    | 0      | 1      |
| word5   | 0      | 0.33   | doc5    | 0      | 1      |
| word6   | 0      | 0.33   | doc6    | 0      | 1      |

- 具体地话题表示为**单词的概率分布**，文本表示为**话题的概率分布**。LDA是含有这些概率分布的模型。



# 图分析

## Graph Analytics

---

- 图分析 (graph analytics) 的目的是发掘隐藏在图中的**统计规律**或**潜在结构**。
- PageRank算法是无监督学习方法，主要是发现有向图中的**重要结点**。
- 给定一个有向图，定义在图上的随机游走即马尔可夫链。
- 随机游走者在有向图上随机跳转，到达一个结点后以等概率跳转到链接出去的结点，并不断持续这个过程。
- PageRank算法就是求解该**马尔可夫链的平稳分布**的算法。

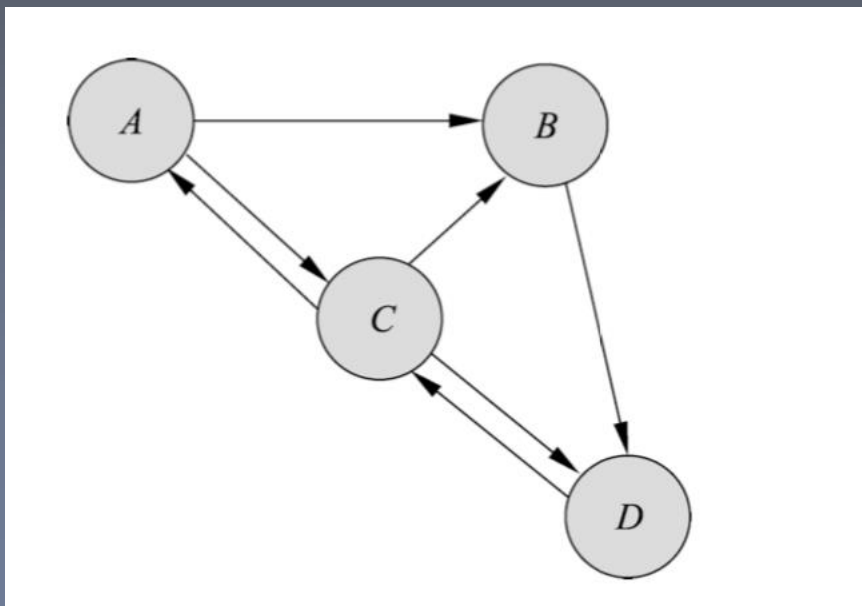
# Page Rank

## Page Rank

- 一个结点上的**平稳概率**表示该结点的重要性，称为该结点的PageRank值。
- 被指向的结点越多，该结点的PageRank值就越大。
- 被指向的结点的PageRank值越大，该结点的PageRank值就越大。
- PageRank值越大结点也就越**重要**。

# Page Rank原理

The principle of Page Rank



- 上图是一个简单的有向图，有4个结点 A,B,C,D。
- 给定这个图，PageRank算法通过**迭代**求出结点的PageRank值。

# Page Rank原理

## The principle of Page Rank

- 首先，对每个结点的概率值初始化，表示各个结点的到达概率，假设是等概率的。
- 下一步，各个结点的概率是上一步各个结点可能跳转到该结点的概率之和。
- 不断迭代，各个结点的到达概率分布趋于平稳分布，也就是PageRank值的分布。

| 步骤<br>结点 | 第 1 步 | 第 2 步  | 第 3 步  |
|----------|-------|--------|--------|
| <i>A</i> | $1/4$ | $2/24$ | $3/24$ |
| <i>B</i> | $1/4$ | $5/24$ | $4/24$ |
| <i>C</i> | $1/4$ | $9/24$ | $9/24$ |
| <i>D</i> | $1/4$ | $8/24$ | $8/24$ |



深度之眼  
deepshare.net

联系我们：

电话：18001992849

邮箱：[service@deepshare.net](mailto:service@deepshare.net)

QQ：2677693114



公众号



客服微信

