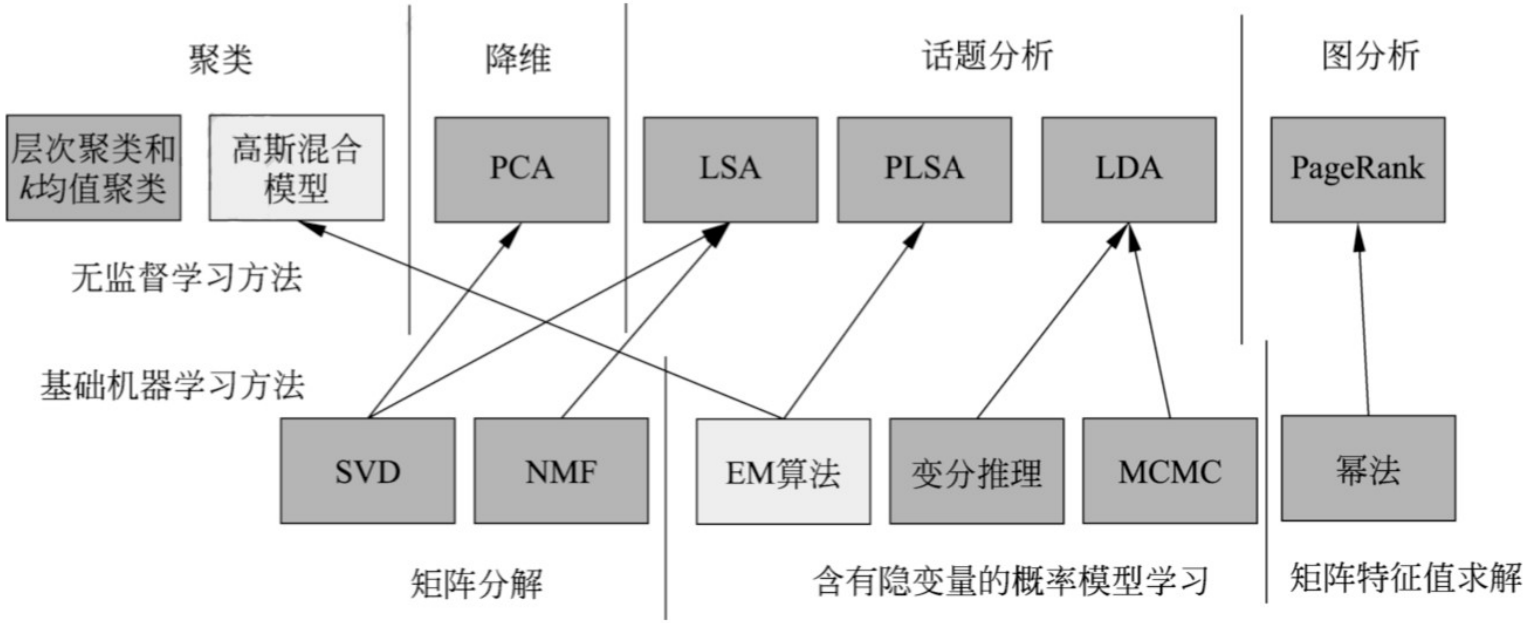
清華大學
Tsinghua University

第二十二章
无监督学习方法总结

各种方法之间的关系


聚类
降维
话题分析
图分析
层次聚类和k均值聚类
高斯混合模型
PCA
LSA
PLSA
LDA
PageRank
无监督学习方法
基础机器学习方法
SVD
NMF
EM算法
变分推理
MCMC
幂法
矩阵分解
含有隐变量的概率模型学习
矩阵特征值求解

各种方法之间的关系

- 无监督学习
  - 聚类
  - 降维
  - 话题分析
  - 图分析

- 聚类的方法
  - 层次聚类
  - K均值聚类
  - 高斯混合模型

各种方法之间的关系

- 无监督学习
  - 聚类
  - 降维
  - 话题分析
  - 图分析

- 降维的方法
  - PCA

各种方法之间的关系

- 无监督学习
  - 聚类
  - 降维
  - 话题分析
  - 图分析

- 话题分析的方法
  - LSA
  - PLSA
  - LDA

各种方法之间的关系

- 无监督学习
  - 聚类
  - 降维
  - 话题分析
  - 图分析 
    - 图分析的方法
      - PageRank

各种方法之间的关系

- 基础方法
  - 矩阵分解
  - 矩阵特征值求解
  - 含有隐变量的概率模型估计

线性代数问题

各种方法之间的关系

- 基础方法
  - 矩阵分解
  - 矩阵特征值求解
  - 含有隐变量的概率模型估计

} 概率统计问题

各种方法之间的关系

- 基础方法
  - 矩阵分解
  - 矩阵特征值求解
  - 含有隐变量的概率模型估计

- 矩阵分解的方法
  - SVD
  - NMF

各种方法之间的关系

- 基础方法
  - 矩阵分解
  - 矩阵特征值求解
  - 含有隐变量的概率模型估计

- 矩阵特征值求解的方法
  - 幂法

各种方法之间的关系

- 基础方法
  - 矩阵分解
  - 矩阵特征值求解
  - 含有隐变量的概率模型估计

- 含有隐变量的概率模型学习的方法
  - EM算法
  - 变分推理
  - MCMC

无监督学习方法

硬聚类

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

无监督学习方法

软聚类

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

无监督学习方法

线性降维

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

无监督学习方法

非概率模型

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

无监督学习方法

概率模型

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

基础机器学习方法

表 22.2　含有隐变量概率模型的学习方法的特点

算法	基本原理	收敛性	收敛速度	实现难易度	适合问题
EM 算法	迭代计算、后验概率估计	收敛于局部最优	较快	容易	简单模型
变分推理	迭代计算、后验概率近似估计	收敛于局部最优	较慢	较复杂	复杂模型
吉布斯抽样	随机抽样、后验概率估计	依概率收敛于全局最优	较慢	容易	复杂模型

话题模型

表 22.3　矩阵分解的角度看话题模型

方法	一般损失函数 $B(D\|UV)$	矩阵 U 的约束条件	矩阵 V 的约束条件
LSA	$\|D-UV\|_F^2$	$U^{\mathrm{T}}U=I$	$VV^{\mathrm{T}}=\Lambda^2$
NMF	$\|D-UV\|_F^2$	$u_{mk}\geqslant 0$	$v_{kn}\geqslant 0$
PLSA	$\sum_{mn} d_{mn}\log\frac{d_{mn}}{(UV)_{mn}}$	$U^{\mathrm{T}}1=1$ $u_{mk}\geqslant 0$	$V^{\mathrm{T}}1=1$ $v_{kn}\geqslant 0$

话题模型

表 22.4　话题模型 LSA 和 NMF 的约束条件

方法	变量 u_k 的约束条件	变量 v_n 的约束条件
LSA	正交	正交
NMF	$u_{mk} \geqslant 0$	$v_{kn} \geqslant 0$