第22章 无监督学习方法总结

无监督学习方法的关系和特点

第2篇详细介绍了八种常用的统计机器学习方法，即聚类方法（包括层次聚类与k均值聚类）、奇异值分解（SVD）、主成分分析（PCA）、无监督学习方法总结 22.1无监潜在语义分析（LSA）、概率潜在语义分析（PLSA）、马尔可夫链蒙特卡罗法（CMC，包括 Metropolis-Hastings-算法和吉布斯抽样）、潜在狄利克雷分配（LDA）、 PageRank算法。此外，还简单介绍了另外三种常用的统计机器学习方法，即非负矩阵分解（NMF）变分推理、幂法。这些方法通常用于无监督学习的聚类、降维、话题分析以及图分析。

表 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

表 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

表 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	$\|\|D-UV\|\|_F^2$	$U^TU=I$	$VV^T=\Lambda^2$
NMF	$\|\|D-UV\|\|_F^2$	$u_{mk}\geq 0$	$v_{kn}\geq 0$
PLSA	$\sum_{mn} d_{mn}\log\frac{d_{mn}}{(UV)_{mn}}$	$U^T1=1$ $u_{mk}\geq 0$	$V^T1=1$ $v_{kn}\geq 0$

本文代码更新地址：https://github.com/fengdu78/lihang-code

中文注释制作：机器学习初学者公众号：ID:ai-start-com

配置环境：python 3.5+

代码全部测试通过。