第 9 周 学习笔记

本周的学习内容主要是推荐系统工具 surprise ,PySpark 基本介绍及使用和关联规则，以下是对本周学习的总结：

（一）surprise 工具

surprise 是一种很流行的推荐系统库 ,可以支持多种推荐算法 ,如基础算法、协同过滤、矩阵分解等。Surprise 包括多个模块，分别是：（1）surprise.similarities 相似度的度量，余弦相似度 cosine、均方差相似度 msd、皮尔逊相关系数 pearson、基线皮尔逊相关系数 pearson_baseline；（2）surprise.accuracy 评价指标，包括均方根误差 rmse、平均绝对误差 mae、一致序列对比率评分（FCP ,Fraction of Concordant Pairs）；（3）surprise.dataset 数据集模块，可以调用系统内置的数据集，验证推荐系统的优劣；（4）surprise.model_selection 模块, 顾名思义 , 就是选择模型的模块 , 包括用于交叉验证所需要的数据集切分、自动 CV、网格搜索 GridSearchCV 等 , 通过该模块可以对模型的参数进行调优。

（二）PySpark 基本介绍

PySpark 是 spark 的 python 版本，利用该库，可以方便使用 python 利用 Spark 框架 ,而不用对 Spark 本身进行繁琐的操作。PySpark 提供了 PySpark Shell ,PySpark Shell 将 Python API 链接到 Spark 核心并初始化 Spark 上下文。在安装之后，需要安装 JAVA 环境，并在 jupyter 中配置 JAVA 环境，之后 PySpark 才能正常的运行。

PySpark 可以方便地进行机器学习建模 , 其中的 mllib API 包含了常见的机器学习任务所需的模块，如分类（mllib.classification）、回归（mllib.regression）、聚类（mllib.clustering）、推荐（mllib.recommendation）等。

就数据分析而言 , PySpark 和 pandas 库有许多类似的相似的地方 , 通常二者可以相互

转换，如读取数据、数据分组计算、数据转换、数据透视表、数据基本统计分析等。

（三）推荐系统中的关联规则分析

事物之间是普遍联系的，如何找到事物的关联将有助于推荐系统的构建。关联规则就是分析事物联系的经典方法。在关联规则中，有几种基本的定义：（1）支持度，公式表示为 $Support(X,Y)=P(X,Y)=\frac{num(XY)}{num(Allsamples)}$，表示 X 和 Y 两个事物集合共同出现的次数在总的观测次数中所占的比重，显然支持度越高，X 和 Y 越容易一起发生；（2）置信度，公式表示为 $Confidence(X \Leftarrow Y)=P(X|Y)=P(XY)/P(Y)$，表示在 Y 发生后 X 跟着发生的概率，类似于条件概率，由 X 与 Y 共同出现的次数：Y 出现的次数得到，置信度越高，说明 X 伴随 Y 发生的概率越大；（3）提升度，表示含有 Y 的条件下，同时含有 X 的概率，与 X 总体发生的概率之比，公式表示为 $Lift(X \Leftarrow Y)=P(X|Y)/P(X)=Confidence(X \Leftarrow Y)/P(X)$，该值越大，说明 X 受到 Y 的影响越大。

频繁项集是指满足一定的条件，即认为可以频繁出现的集合。Apriori 算法是基于频繁项集的方法，其基于这样的事实来判断项集是否为平凡项集，即——如果一个项集是频繁项集，则它的所有子集都是频繁项集；如果一个集合不是频繁项集，则它的所有父集（超集）都不是频繁项集。

关联分析的目标：发现频繁项集,即发现满足最小支持度的所有项集；发现关联规则，从频繁项集中提取所有高置信度的规则。

Apriori 算法的原理是：（1）先搜索出候选 1 项集及对应的支持度，剪枝去掉低于支持度的 1 项集，得到频繁 1 项集。（2）对剩下的频繁 1 项集进行连接，得到候选的频繁 2 项集，筛选去掉低于支持度的候选频繁 2 项集，得到真正的频繁二项集。（3）以此类推，迭代下去，直到无法找到频繁 k+1 项集为止，对应的频繁 k 项集的集合即为算法的输出结果。

以下是 Apriori 算法的伪代码：

输入：数据集合，支持度阈值

输出：最大的频繁 k 项集

1）扫描整个数据集，得到所有出现过的数据，作为候选频繁 1 项集。k=1，频繁 0 项集为空集。

2）挖掘频繁 k 项集

a) 扫描数据计算候选频繁 k 项集的支持度

b) 去除候选频繁 k 项集中支持度低于阈值的数据集,得到频繁 k 项集。如果得到的频繁 k 项集为空，则直接返回频繁 k-1 项集的集合作为算法结果，算法结束。如果得到的频繁 k 项集只有一项，则直接返回频繁 k 项集的集合作为算法结果，算法结束。

c) 基于频繁 k 项集，连接生成候选频繁 k+1 项集。

3） 令 k=k+1，转入步骤 2。