机器学习技术在推荐系统中的应用

打造千人千面的个性化推荐引擎

推荐搜索部

刘思喆

2014 年 9 月 27 日

目录

推荐系统

目录

推荐系统

京东推荐产品

- 80+ 推荐产品，包括移动端和 Web 端
- 20+ 推荐服务，支撑 EDM、广告、微信端等
- 遍布用户网购的各个环节

推荐系统的价值

- 挖掘用户潜在购买需求
- 缩短用户到商品的距离
- 用户需求不明确时提供参考
- 满足用户的好奇心

推荐产品截图示例


根据浏览猜你喜欢
金士顿（Kingston）8G Class4 TF（micro SD）存储
(已有466365人
¥26.90
购买该商品的用户还购买了
康纳(CONNAL) ZTD100K-SD1 嵌
直降 ¥1299.00
加入购物车
快易典H5 平板学生电脑 小学初中高中同步学习机 英
您可能还需要以下商品
赛尔贝尔（Syllable） G03-002 柯莉
¥59.00
(已有9111人评价)
加入购物车
关注此商品的人还关注
快易典EH1 电子词典 翻译机 小学初
¥298.00
加关注
快易典（Koridy）A990全能词典王
¥339.00
加关注
快易典EH2 电子词典 学生英语辞典
¥418.00
加关注
好易通(besta) 无敌V4牛津高阶+剑
¥669.00
加关注
购买了该商品的用户还购买了
海尔(haier) ES50H-Q1(ZE) 50
热水器
999.0
LG WD-N12435D 滚筒洗衣机
色）
2499.0
浏览了该商品的用户最终购买
康纳(CONNAL) CXW-200-TD08A
欧式 吸油烟机
¥1799.00
华帝(VATTI) JZT-i10008C 台嵌两用
式 燃气灶(天然气)
1299.0

不同位置的推荐产品定位不同

- 单品页：购买意图
- 过渡页：提高客单价
- 购物车页：购物决策
- 无结果页：减少跳出率

- 订单完成页：交叉销售
- 关注推荐：提高转化
- 我的京东推荐：提高忠诚度

京东推荐算法优化方向

- 以数据分析为工具，提升数据的质量和覆盖度，增强对业务的理解（25%）
- 测试不同算法在不同数据源的效果，提高召回模型的质量，增加结果辨识度（50%）
- 以用户反馈为依据，融合不同类型、不同维度据源，对推荐结果重排序（15%）
- 增加数据的更新频率（5%）
- 其他（5%）

京东推荐算法优化方向

- 以数据分析为工具，提升数据的质量和覆盖度，增强对业务的理解（25%）
- 测试不同算法在不同数据源的效果，提高召回模型的质量，增加结果辨识度（50%）
- 以用户反馈为依据，融合不同类型、不同维度据源，对推荐结果重排序（15%）
- 增加数据的更新频率（5%）
- 其他（5%）

目录

推荐系统

典型推荐系统技术

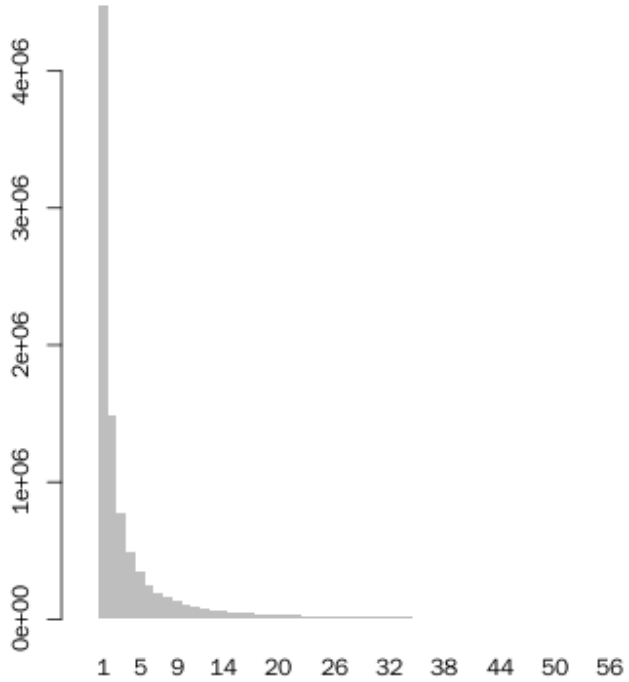
按照数据的分类： 协同过滤、内容过滤、社会化过滤

按照模型的分类： 基于近邻的模型、矩阵分解模型、图模型

京东对推荐数据的理解

用户行为

1. 浏览
2. 点击
   - 普通点击
   - 搜索点击
3. 加入购物车（或关注）
4. 购买
   - 订单
   - 用户
5. 评分



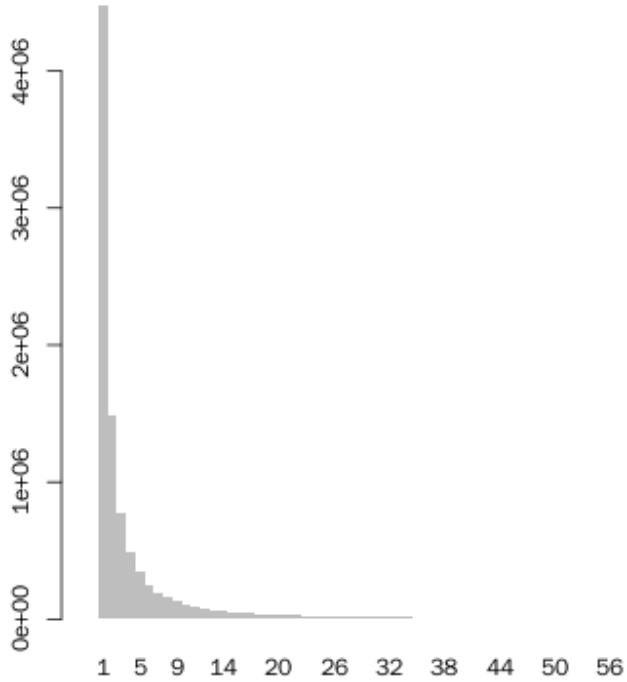
基于内容

- 标题
- 扩展属性
- 评论
- 描述
- …

京东对推荐数据的理解

用户行为

1. 浏览
2. 点击
   - 普通点击
   - 搜索点击
3. 加入购物车（或关注）
4. 购买
   - 订单
   - 用户
5. 评分



基于内容

- 标题
- 扩展属性
- 评论
- 描述
- ...

目录

推荐系统

推荐的 CTR 预测

什么是推荐商品的 CTR（Click Through Rate）？

- 关联推荐的情境下，根据给定主商品推出的推荐商品，在用户浏览后被点击的概率。
- 可以理解为条件概率 $P(Y = 1|X)$

为什么要预测推荐商品的 CTR？

1. 调整推荐商品的排序
2. 用于多模型的融合
3. 发现影响推荐商品点击率的重要因素

特征表征方法

用目标问题所在的特定领域知识或者自动化方法来生成、提取、删减或组合变化来得到特征。

领域经验法

- 条件关系（=,!=）
- 几何运算
- 分段及比例
- 其他

自动化技术

- PCA, ICA, NMF
- Linear Discriminant Analysis
- Collaborative Filtering
- AutoEncoder

最优子集（Feature selection）的优点

- 提高模型的可解释性
- 减少训练和预测的时间
- 有效降低过拟合，提升模型的适应能力

Feature selection methods I

- Filters select subsets of variables as a pre-processing step, independently of the chosen predictor.
- Wrappers utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power.
- Embedded methods perform variable selection in the process of training and are usually specific to given learning machines.

Feature selection methods II

- Subset Selection (Stepwise and Stagewise Selection)
- Shrinkage Methods (Ridge Regression and Lasso)
- Methods Using Derived Input Directions(Principal Components Regression, Partial Least Squares)

如何对商品属性进行描述

对商品的形容：

品牌词、中心词、修饰词；类目属性、扩展属性；

基于用户行为的在商品上的反映：

- 销量、PageRank、评论数、好评度
- 商品的标签（如时间标签、地域标签、性别标签等）

对于商品标签（以时间差异构建的时间 feature 为例）：
假设 9：00 - 19：00 为白天（D），19：00 - 9：00 为夜间（N），则在这两个时间段内的用户购买则构成了该商品的时间标签，该商品标签的一般性定义为：

$$\frac{\sum_{u\in D} M_{u,i}}{\sum_{u\in D} M_{u,i} + \sum_{u\in N} M_{u,i}} - \frac{\sum_{u\in D} M_{u}}{\sum_{u\in D} M_{u} + \sum_{u\in N} M_{u}}$$

商品的组合属性

基于单一属性组合产生的属性，有以下三种：

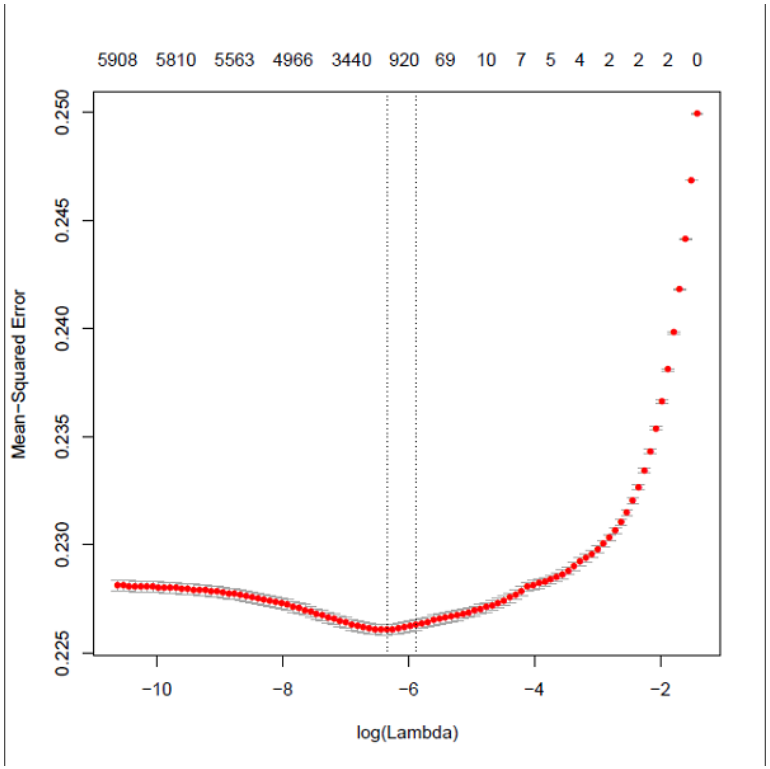
- 相同类属性的组合：如时序上的销量（趋势系数），销量的方差
- 不同类属性的组合：如商品的展示和点击组合（如 CTR）、点击和购买的组合（如 CVR）
- 推荐主商品和推荐品属性的组合。比如品牌词是否一致，价格的比值是否在一定范围内。

推荐主商品和推荐品三级类目关系需要使用两两配对的 feature 表征形式。

数据预处理及建模过程

- 去掉样本量较小的类，共 25 个一级类需要预测
- 对不均衡样本采取了 undersampling 策略，同时配置 5 次重复抽样预测
- 训练数据量为 500w，在并行 CV 选取 λ 的时间为 15-20 分钟
- 预测重排序数据为 6 亿条
- 预测所有数据，16 线程情况约为 1 小时

不同 λ 交叉验证的 MSE 曲线



部分三级类组合系数展示

1	9863	1478	-0.08	便携桌椅床	床
2	9756	12100	-0.08	板鞋	跑步鞋
3	1695	1698	-0.07	篮球	羽毛球
4	9790	12153	-0.07	其它	太阳镜
5	1474	1478	-0.07	便携桌椅床	睡袋/吊床
6	12123	12131	-0.07	户外风衣	冲锋衣裤
7	2629	1471	-0.07	户外鞋袜	户外配饰
8	1355	9767	-0.07	套装	T 恤
.	...	...	...	...	...
9	1698	9757	0.23	篮球鞋	篮球
10	1671	1694	0.24	乒乓球	纸品湿巾
11	12122	9756	0.24	跑步鞋	轮滑滑板
12	5152	1476	0.26	户外照明	军迷用品
13	1392	12121	0.27	骑行装备	面膜
14	1478	12153	0.29	其它	便携桌椅床
15	9765	12103	0.29	运动配饰	T 恤
16	2690	12131	0.29	户外风衣	户外服装
17	1474	12127	0.39	休闲衣裤	睡袋/吊床

过渡页购买还购买 CTR 预测模型实验

实验效果

- 实验流量 10%
- 观测时长 30 天
- 请求点击率：提升 14%
- 千次请求订单行数：提升 1%

不同平台（架构）的特点

平台	算法支持	成熟度	特征表征	训练速度	预测速度	工程化	扩展性
SPARK	有限	持续开发	较复杂	很快	很快	较难	较高
MPI	重新开发	较高	复杂	很快	很快	较难	较高
Vowpal wabbit	较多	较高	很方便	很快	较慢	非常方便	一般
R	很多	较高	需写函数	较快	很快	需开发	较高

目录

推荐系统

总结和回顾

- 数据的理解高于算法的理解，简单模型配以优质有效数据有更加的效果
- CTR 预测模型可以迅速学习出合理的模式做推断外延，关键点在于特征工程的合理程度
- ...

谢谢！
Thank you!

北京市朝阳区北辰西路8号北辰世纪中心A座6层
6F Building A, North-Star Century Center, 8 Beichen West Street,
Chaoyang District, Beijing 100101
T. 010-5895 1234 F. 010-5895 1234
E. xingming@jd.com www.jd.com

JD.京东.COM