京东大数据 R 语言应用与实践

数据创造价值

JD.com

刘思喆

2015 年 6 月 28 日

目录

大数据应用

目录

应用背景

- 京东业务涉及用户、商品、商家、促销、反作弊、风险控制、精准营销、供应链优化、金融等。京东是中国唯一的数据覆盖全链路的电商平台。
- 2012 年 JD.com 正式启动了大数据平台的搭建，平台底层数据存储和离线运算由 hadoop 完成，在京东所有的程序均基于自由语言和平台开发
- 除搜索、推荐、广告等线上数据应用以外，同样也有适应于各条业务线的数据产品，如报表平台、数据调度、数据知识、数据工厂等
- 同传统行业相比，京东的业务场景变换迅速，数据有着多变、海量、复杂、高增长等特点

建模环境简述

数据情况：

- 客户维度：亿级 -> 千万级
- 商品维度：千万级 -> 百万级
- 数据量：GB-TB 级

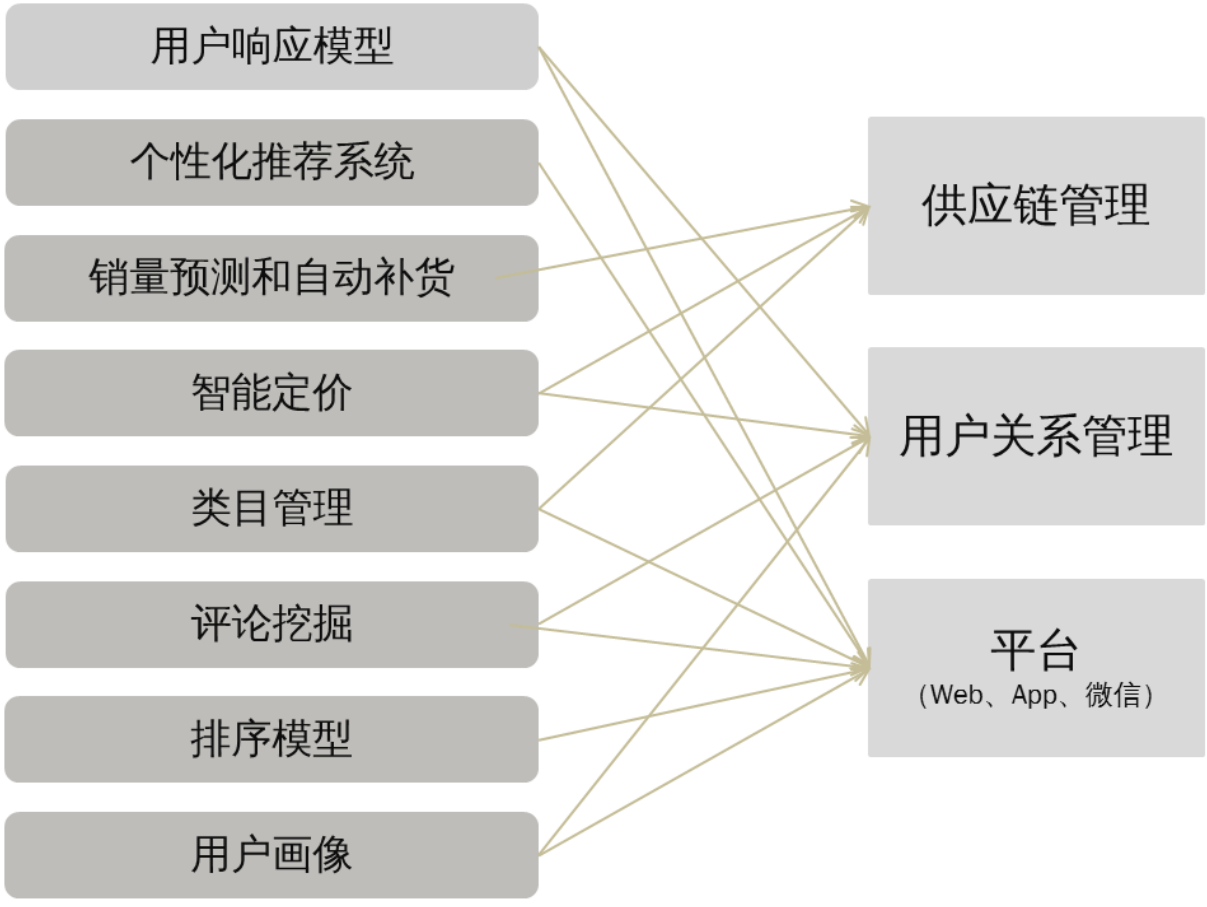
分析技术：

1. 探索分析：均值、方差、分位数、列联表
2. 基础分析：如假设检验、相关分析、主成分（因子）分析
3. 挖掘模型：回归、kmeans 聚类、概率图模型、决策树、随机森林、GBDT、关联规则、矩阵分解、时间序列等
4. 可视化图形：条图、直方图、概率密度图、定制化图形
5. 重复性分析：报表、周报、日报（邮件）

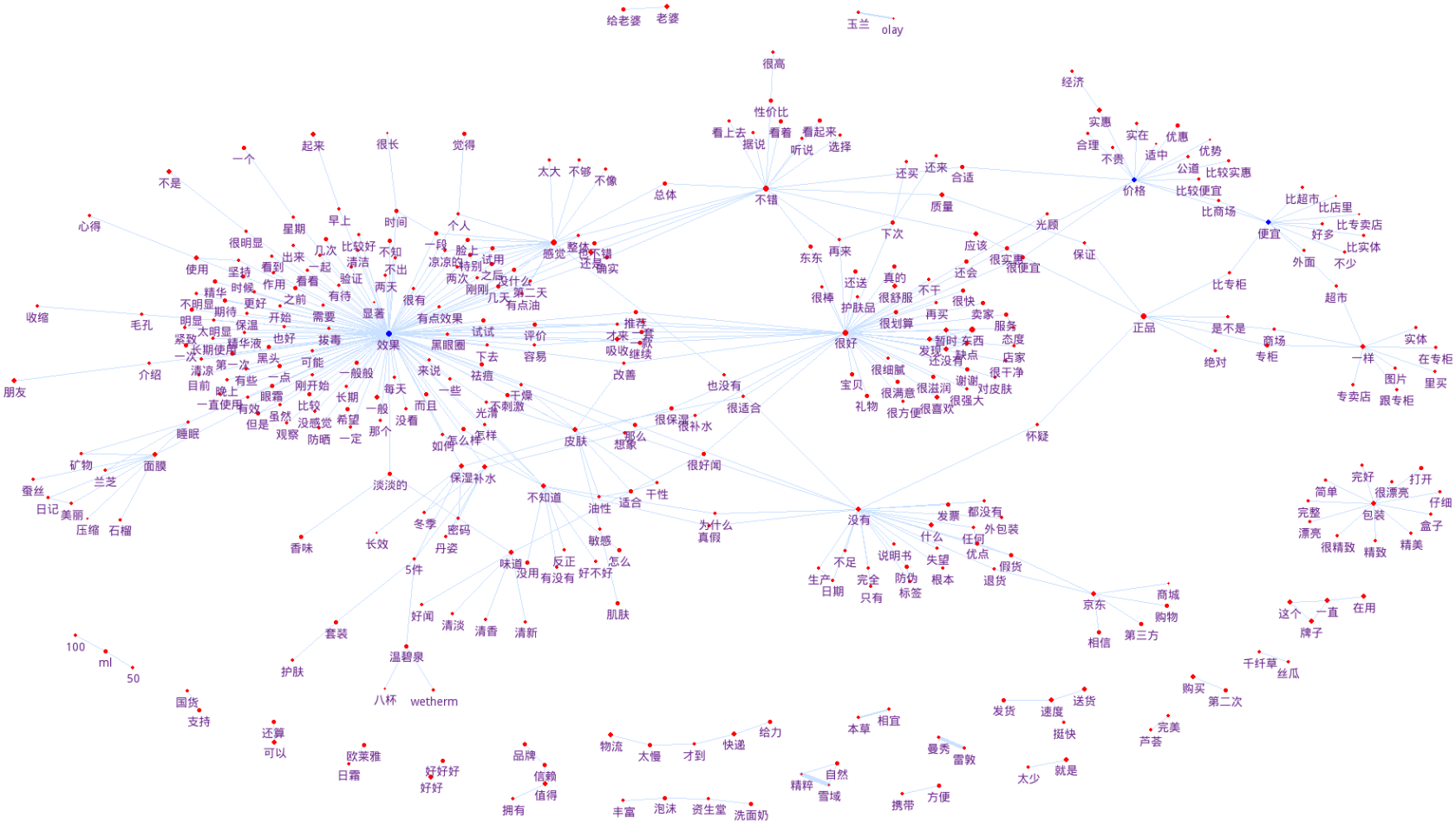
从 R 的角度来看

- 数据挖掘领域应用最广泛的软件和语言（KDnuggets 2012-2015）
- 完整且丰富的统计、机器学习、可视化平台
- 数据编程的完美实现
- 便捷的、可扩展的并行方案（如同 Hadoop、Spark、Redis）

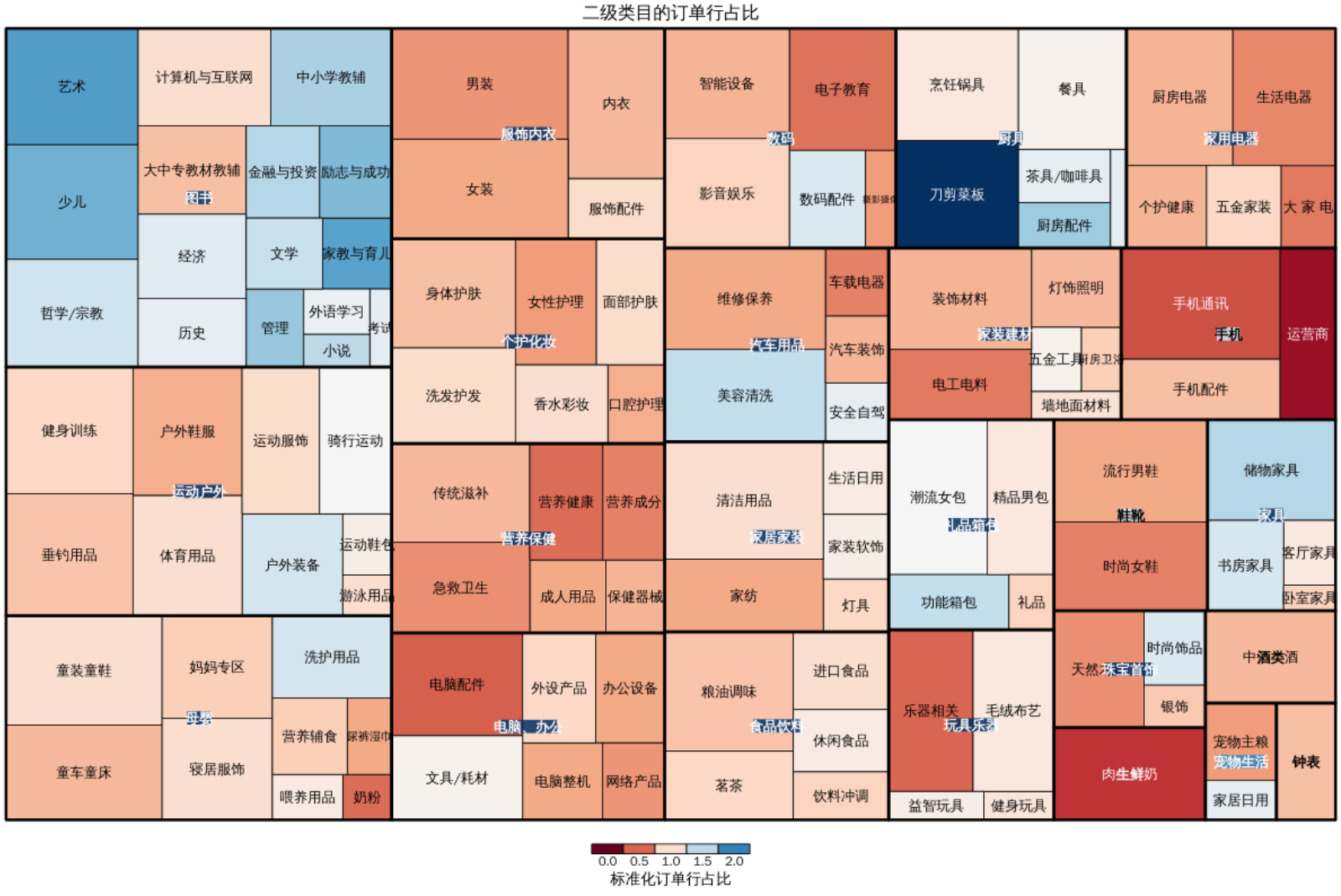
分析和挖掘模型服务对象



数据分析 -洞察业务的利器 I



数据分析 -洞察业务的利器 II



注：根据 JD.com 数据发布要求，引用数据已经做过随机化处理

数据分析 -洞察业务的利器 III

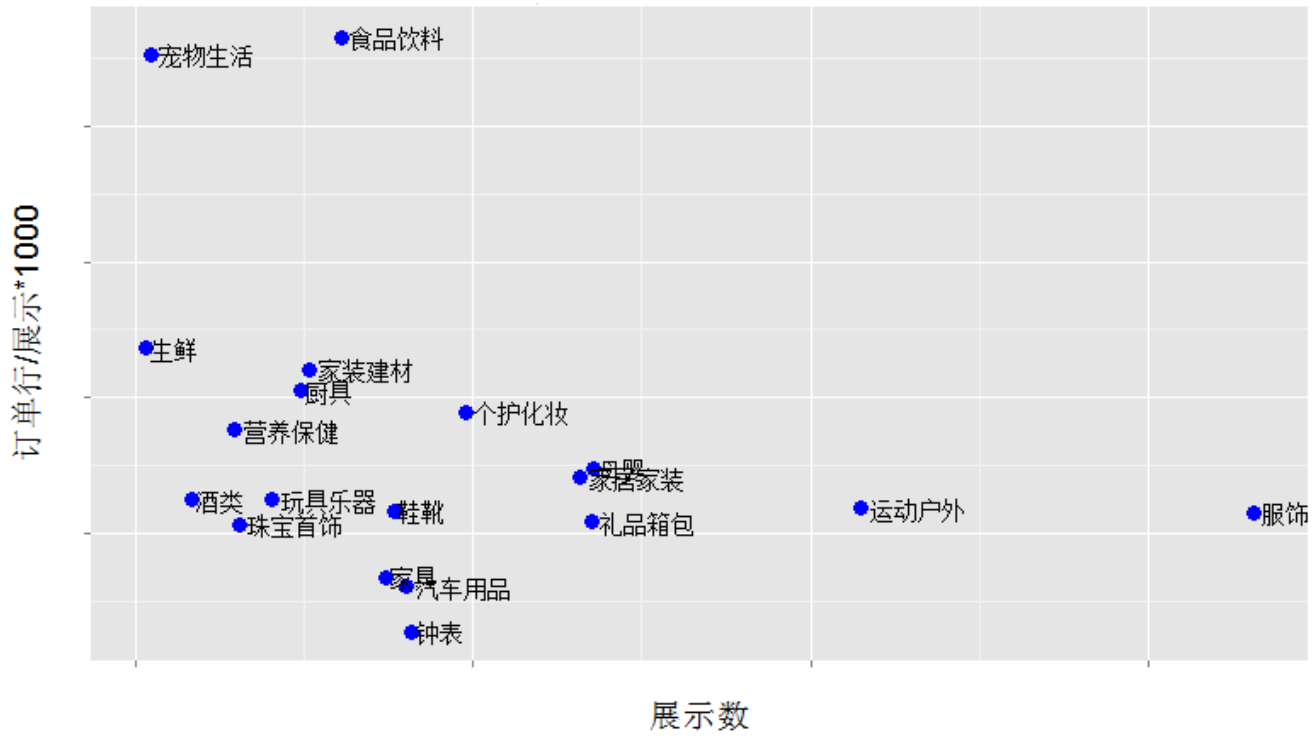


Figure: 某推荐位在在各个一级品类下的展示和转化的对比情况（横纵坐标已隐去）

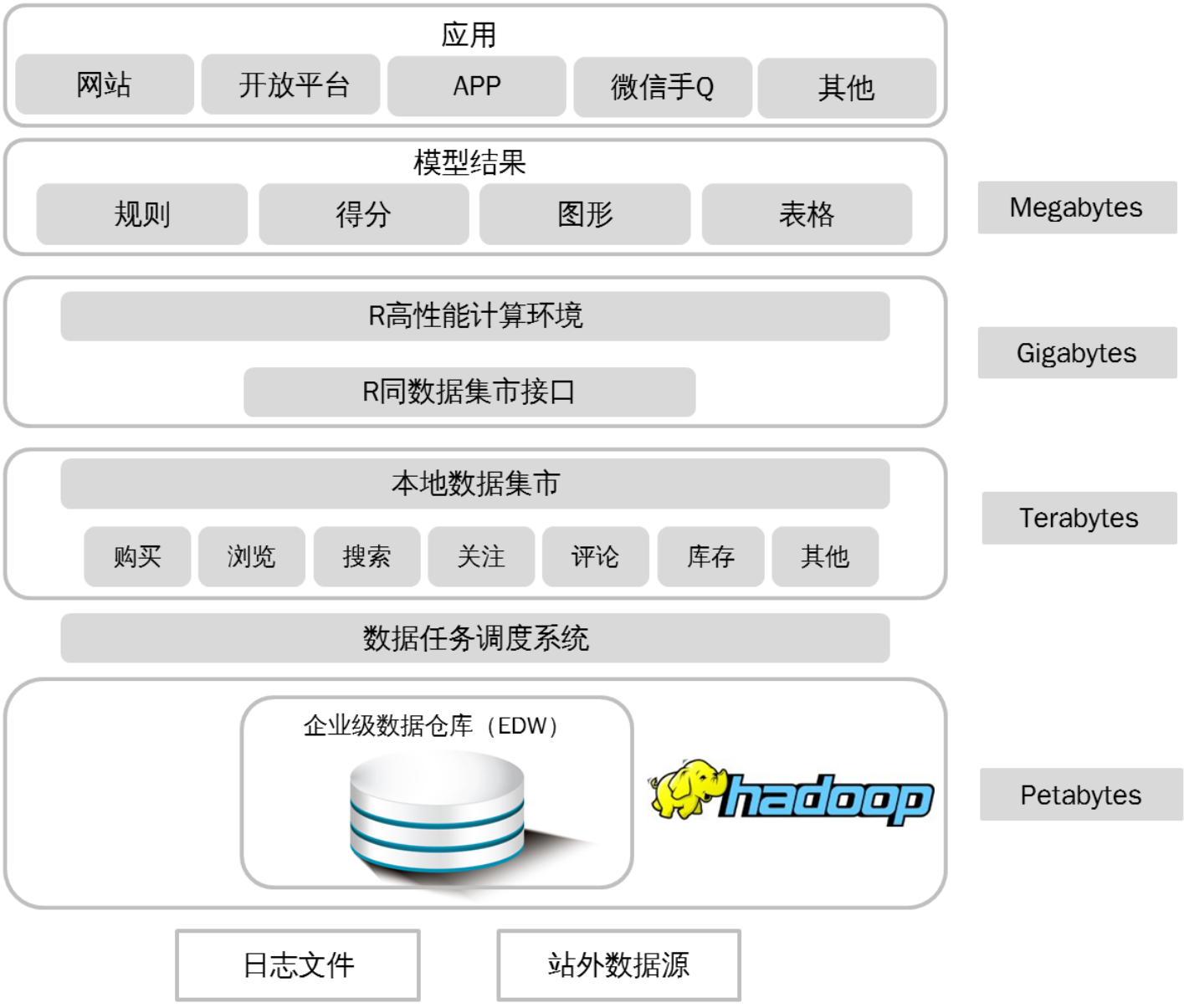
目录

大数据应用

典型的工作流程

1. 通过 Hive 或 Pig 集群获取目标数据
2. 在 R 环境下进行数据探索、清洗、转换工作
3. R 环境下分析建模 (Feature Selection, Benchmark)
4. 评估 (离线评估和分流量测试)
5. 线上集成 (R, Hive QL, Java, C++, Python...)

数据的流动


应用
网站
开放平台
APP
微信手Q
其他
模型结果
规则
得分
图形
表格
Megabytes
R高性能计算环境
R同数据集市接口
Gigabytes
本地数据集市
购买
浏览
搜索
关注
评论
库存
其他
Terabytes
数据任务调度系统
企业级数据仓库（EDW）
hadoop
Petabytes
日志文件
站外数据源

涉及数据挖掘、分析技术的相关 R 包

- **数据传递及服务**（RHive、RServe、rJava、RJDBC）
- **清洗及预处理**（sqldf、stringr、XML、data.table）
- **抽样、预测、分类、关联规则、特征选择、稀疏矩阵运算、矩阵分解、社交网络、分词、模型评估等**
- **高性能计算**（rhdfs、rmr2、Rcpp、snow）
- **自动化报告**（knitr、slidify）
- **其他**

目录

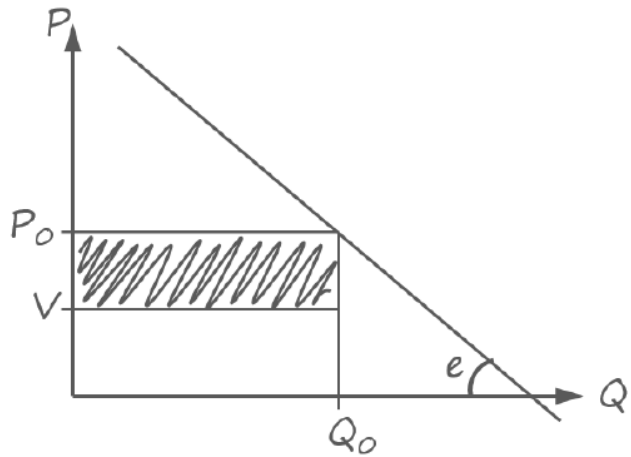
大数据应用

问题陈述

对于任意的零售商来说，通常的利润依照以下公式计算：

$$G = Q \times (P - V) - C$$

这里 G 是利润，Q 是销售量，P 是单位商品价格，V 是商品成本价，C 是固定费用（比如分摊管理费用）。



一般需求变化百分比和价格变化的百分比之间的关系定义为价格弹性：

$$e = \frac{\Delta Q/Q}{\Delta P/P}$$

原型展示

```
P <- 50:150
Q <- 1e6/P
PQ.lm <- lm(log(Q) ~ log(P))
> PQ.lm

Call:
lm(formula = log(Q) ~ log(P))

Coefficients:
(Intercept)       log(P)
      4.605       -1.000
```

通过构建不变弹性模型，我们发现价格 P 和销量需求 Q 的弹性为 -1，即价格和销量需求是负向关系。

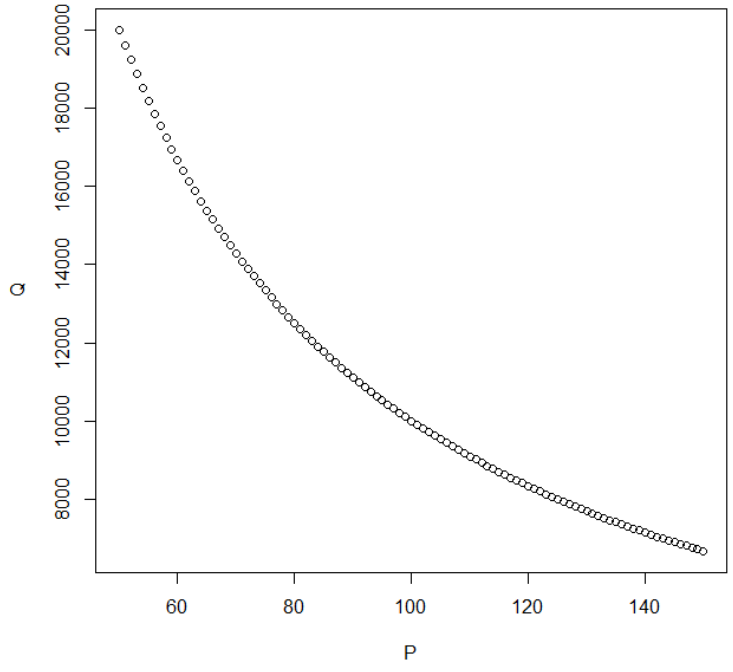


Figure: 销量 Q 随着价格的升高而降低，其实际的度量为每升高 1% 的价格，则会导致 1% 的销量的下降

不变弹性模型的优势

- 从定义上看，弹性模型并不关注变量的绝对变化，而是相对变化，这样对于不同的 sku 则可以进行对比；
- 不变弹性模型可以反映商品在生命周期上特征（导入期、发展期、成熟期、衰退期的弹性迥异）；
- 可以对商品需求量的时间序列预测进行修正。

关于价格的定义

这里使用的并非京东价，而是重新定义的“印象价格”，即用户的付出成本。

应用一：商品细分

假设所观测的 sku 生命周期比较稳定（适用不变价格弹性模型），那么

1. 价格弹性大的 sku 通过降价，可以获得更高的需求（销量），从营销角度讲，有比较好的（降价）促销空间
2. 而价格弹性比较小的 sku 对价格的变动并不敏感，并不会随着价格的升高而导致需求的变化（刚性的），则可以考虑适当提价，增加单位商品的利润率，或采用其他非价格促销方案（如捆绑）
3. 同理对于 PV 弹性来说，高 PV 弹性的 sku 适合引入流量，而低 PV 弹性则需考虑其他方式提升销量。

应用二：销量预测的修正

在求解不变弹性以后，根据未来价格的变化预计（促销降价力度），修正销量（需求）的变化

1. 计算时间范围 T_1 的价格对销量不变弹性 E_d；
2. 根据 T_1 时间范围的销量数据，求解未来 T_2 时间范围的销量 S；
3. 估算未来 T_2 时间范围的价格变化程度 P_d(百分比)；
4. 同时考虑在多次降价后，其效应会有所下降[1]，引入调整因子 adjust（设置为 o.8）
5. 根据 E_d、P_d、adjust 来修正 S

[1]我们在引入数据一般优先考虑销量较高的 sku，这类 sku 恰恰处于商品生命周期的成熟期，而接下来的则会进入衰退期，sku 弹性有下降趋势。因此在预测时会考虑将预测期的弹性调低

实际商品案例

1. 九阳（joyoung）豆浆机 DJ13B-D08：
   - 观察 4 月 1 日至 5 月 6 日共计 36 个数据，其价格销量弹性为 -7.208，PV 销量弹性为 1.363
   - 根据 arima(1,1,1) 模型对其后的四天的销量进行预测，得到的预测数据为 53, 54, 54, 54。通过修正，预测数据调整为 145,149,148,148，而实际发生的销量为 106,141,131,101。
   - 未修正 MAPE 值为 54.88%，修正 MAPE 值为 23.70%
2. ThinkPad E420（1141-AA5）14 英寸笔记本电脑：
   - 同样采用上面的时间窗口，arima(1,1,1) 预测
   - 价格弹性 -11.38，PV 弹性 0.55（客户只对降价感兴趣）
   - 未修正 MAPE 值为 28.54%，修正 MAPE 值为 2.94%

对使用价格弹性对需求预测修正的关键点

并非所有的 sku 都适用于此种方法，其原因有如下几种情况

1. sku 销售的时间窗口不能太短，至少满足大于等于 30 天。如果时间窗口太短，会导致回归拟合出现问题（新品）；
2. sku 销量不能太低，想象一下：如果某个 sku 每日销量均为个位数，拟合的回归方程会不能通过假设检验，甚至拟合没有实际意义（长尾商品）；
3. 在需求预测实际修正过程中，要求尽量保持预测期的价格变化稳定。比如根据促销方案，7 天预测期的价格大致会降低 5%，而实际情况是前 3 天降 10%，后 4 天降 1%，这种状况就会导致修正错误。
4. 有些 sku 的生命周期很短，迅速经历导入期、生长期、成熟期、衰退期，这时候弹性变化是非常剧烈的，这时候不宜使用不变弹性模型。

目录

大数据应用

如何评价一款商品的好坏？

1. 普通消费者在购买商品后，会对商品有一个全面的感受，比如销售者满意度、客户体验等。但如何收集消费者对商品的感受数据则是难点。
2. 消费者对一件（一组）商品满意度非常差，则消费者会选择不再回来购物。这种流失的源头是出现了 ``不良'' 商品，需要商家进行优化（比如配送、商品质量、商品价格等）

可选的数据解决方案

- 客户投诉数据
- 商品评论的文本数据
- 基于用户购买行为

前两种方案优点是非常明显的，是了解商品满意度的直接途径。但也存在部分缺点：

- 投诉数据量有限，有些用户并不倾向通过投诉来表示不满，而是直接``用脚投票''
- 使用评论数据的前提是：所有的不满意用户都会在网站上留言，但事实并非如此
- 用户对于商品挑剔程度不同：有的用户所有的评论都呈现攻击性态度，但这些差评并不影响未来的购物

基本原理

在京东第一次购物的用户体验非常重要，体验好则成为存量用户，反之则失去用户。已知这些新用户第一次购买的商品清单和用户未来的状态（未来是否再发生购物行为），则可以生成如下矩阵：

status		p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
1	u1	0	0	0	1	1	0	0	0	0	0
1	u2	0	0	1	1	0	0	0	1	0	0
1	u3	1	1	0	0	0	0	0	0	1	0
1	u4	1	0	0	1	0	0	0	0	0	0
1	u5	0	1	0	0	0	1	0	0	0	0
0	u6	0	0	1	0	0	0	0	0	1	0
0	u7	0	1	0	0	0	0	1	0	0	1
0	u8	1	0	0	0	1	0	0	0	0	0
0	u9	0	0	0	1	0	0	0	0	1	0
0	u10	0	0	1	0	0	0	0	1	0	1

- 如果用户未来 X 个月又购买了某种商品，则在这个矩阵 status 对应的位置标记为 1，反之为 0。
- 矩阵的行代表了用户，列代表了商品：对于第一行来说，u1 用户购买了 p4、p5 两件商品。

基本原理

在京东第一次购物的用户体验非常重要，体验好则成为存量用户，反之则失去用户。已知这些新用户第一次购买的商品清单和用户未来的状态（未来是否再发生购物行为），则可以生成如下矩阵：

status		p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
1	u1	0	0	0	1	1	0	0	0	0	0
1	u2	0	0	1	1	0	0	0	1	0	0
1	u3	1	1	0	0	0	0	0	0	1	0
1	u4	1	0	0	1	0	0	0	0	0	0
1	u5	0	1	0	0	0	1	0	0	0	0
0	u6	0	0	1	0	0	0	0	0	1	0
0	u7	0	1	0	0	0	0	1	0	0	1
0	u8	1	0	0	0	1	0	0	0	0	0
0	u9	0	0	0	1	0	0	0	0	1	0
0	u10	0	0	1	0	0	0	0	1	0	1

- 如果用户未来 X 个月又购买了某种商品，则在这个矩阵 status 对应的位置标记为 1，反之为 0。
- 矩阵的行代表了用户，列代表了商品：对于第一行来说，u1 用户购买了 p4、p5 两件商品。

实际的清单结果及解释

Table: 部分``不良''商品的清单

sku_id	score	product_name
10xxxxxxxx	-0.52	carslan 卡姿兰 xxxxxxx
10xxxxxxxx	-0.41	xxxx xxxx 秋冬新款男士亮面休闲 xxx xxx xxx
10xxxxxxxx	-0.55	xxxx 秋冬新款男装修身中长款 xxxxx xxxxx xxx 黑色 XL
xx41xx	-0.37	xxxx xxxx 18 升电烤箱 xxx xxxx xx
xxxx76	-0.40	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx81	-0.38	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx23	-0.33	联想 xxxx 3G 手机 xxxxxxxx 双卡双待单通
xxxx25	-0.52	xxxxx xxxx 4G 录音笔
xxxx32	-0.41	HTC xxxxx 3G 手机（xxxx）TD-SCDMA/GSM

Table: 对于三类商品的标记以及表现

表现	表现	id	名称	购买用户数	再次购买	不再购买
不良：	-	10xxxxxxxx	xxxxxxxxx 纯棉男袜	32	3	29
正常：	0	38xxxxxx	xxx 超薄干爽纸尿裤箱装 xxxx 片	13	6	7
优秀：	+	30xxxxxx	xx 螺旋藻 xxxxxxxxxx*1 桶	15	14	1

实际的清单结果及解释

Table: 部分``不良''商品的清单

sku_id	score	product_name
10xxxxxxxx	-0.52	carslan 卡姿兰 xxxxxxx
10xxxxxxxx	-0.41	xxxx xxxx 秋冬新款男士亮面休闲 xxx xxx xxx
10xxxxxxxx	-0.55	xxxx 秋冬新款男装修身中长款 xxxxx xxxxx xxx 黑色 XL
xx41xx	-0.37	xxxx xxxx 18 升电烤箱 xxx xxxx xx
xxxx76	-0.40	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx81	-0.38	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx23	-0.33	联想 xxxx 3G 手机 xxxxxxxx 双卡双待单通
xxxx25	-0.52	xxxxx xxxx 4G 录音笔
xxxx32	-0.41	HTC xxxxx 3G 手机（xxxx）TD-SCDMA/GSM

Table: 对于三类商品的标记以及表现

表现	表现	id	名称	购买用户数	再次购买	不再购买
不良：	-	10xxxxxxxx	xxxxxxxxx 纯棉男袜	32	3	29
正常：	0	38xxxxxx	xxx 超薄干爽纸尿裤箱装 xxxx 片	13	6	7
优秀：	+	30xxxxxx	xx 螺旋藻 xxxxxxxxxx*1 桶	15	14	1

有益的效果

按品类，造成用户流失的原因分析略 . . .

- 识别过程更加规整化、流程化。为日常运营中干预 ``不良'' 商品提供了一个有效、快速、便捷的方式。
- 有效减少不良商品对于客户的负面影响。阻止这些客户流失或流向竞争对手，对其他（潜在）顾客的负面影响降低至最低。
- 对于使用以天为记录单位的不良商品识别方法的应用，每天大约能记录 5-10 种不良商品，平均覆盖 100-150 个客户。保守地，按照每位客户一年再购买一次商品，客单价 250 计算，未来一年累计额外带来 900 万 -1350 万的销售额。

有益的效果

按品类，造成用户流失的原因分析略 . . .

- 识别过程更加规整化、流程化。为日常运营中干预 ``不良'' 商品提供了一个有效、快速、便捷的方式。
- 有效减少不良商品对于客户的负面影响。阻止这些客户流失或流向竞争对手，对其他（潜在）顾客的负面影响降低至最低。
- 对于使用以天为记录单位的不良商品识别方法的应用，每天大约能记录 5-10 种不良商品，平均覆盖 100-150 个客户。保守地，按照每位客户一年再购买一次商品，客单价 250 计算，未来一年累计额外带来 900 万 -1350 万的销售额。

总结和思考

- 工具并不是我们数据科学家所关注的，顺手即可。最重要的还是针对于问题的思想和思路
- 不是所有的场景都适合使用 R，但数据挖掘工作这种非线性的工作更适合像 R 这样的开源工具，尤其适合中小企业
- 一般情况下，简洁的统计模型优于复杂的推断模型
- 模型无价值，数据无价值，价值在于如何映射业务
- 数据具有时效性，并需要准确的表现数据背后的信息

- 邮件：liusizhe<at>jd.com
- 博客: `http://www.bjt.name`
- 微博：@ 刘思喆

Jump to first slide