推荐系统中的特征工程

王帅强 & 丁卓冶

京东零售-搜索与推荐平台部-数据科学实验室-推荐科学组

京东

目录

京东

特征构造：单品和素材

京东

特征构造：Dense & Embedding

- JD推荐系统大体分为单品推荐和内容素材推荐两类
- 单品推荐模型特征是内容素材推荐模型特征的子集
- 主要分为Dense特征和Embedding特征
- Dense特征可分为用户单边、商品单边、交互特征三类

用户单边特征

京东

- 用户近期的活跃度（1m内的行为数）
- 价格quantile偏好： (ord, cart) * (all, c3) * (mean, median, min, max)
- 好评率偏好：all, c3
- 基于cart的自营/POP偏好：(all, c3) * (jd, pop) * (tf, tf-idf)
- 促销敏感度偏好：购买促销商品频率，平均折扣率，使用京券的频率，使用京券平均折扣率，用户使用店铺京券的频率，使用店铺京券平均折扣率，用户使用京豆频率

商品单边特征

- 召回特征
  - 商品是否通过某路召回源召回
  - 主要召回源的召回统计量：最大值、平均值、求和
- 商品单边画像
  - 是否自营、Pop 、京东配送、图书、新品、全球购商品、山姆会员商品、秒杀商品、进口商品、奢侈品、高端商品
- 商品单边行为特征
  - 评论数、平均评分、好评数、差评数、好评率、差评率、好评率置信区间下限、差评率置信区间下限、热度分、店铺热度分、价格等级、京东价、log京东价、价格分位数

商品单边特征

- 威尔逊置信区间
  - 解决小样本置信度不高的问题

$$\frac{\hat{p}+\frac{1}{2n}z_{1-\frac{\alpha}{2}}^2 \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{1-\frac{\alpha}{2}}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\frac{\alpha}{2}}^2}$$

- 其中$\hat{p}$是赞成比例，$z_{1-\frac{\alpha}{2}}^2$是满足一定统计量的常数
- 一般取其下限w_-: $\begin{cases} w_- \to \hat{p}, & n \to \infty \\ w_- \to 0, & n \to 0 \end{cases}$

商品单边特征

- 商品单边短期（7 days）行为特征
  - Item level: SKU, C3, PW, Brand
  - General Metrics: CTR与 CVR, 销量、点击量
  - Other Metrics：行为总数、平均每天行为数、用户数、平均每天用户数、用户数占比（ Click & Order ）

交互特征：基于行为的相似

- 用户和商品的交互特征（占比）
  - (sku, c3, c2, pw, brand, gender, adj, ext) * (c, d, w, m) * (clk, follow, cart, ord)
- 用户偏好与目标sku的匹配度
  - 用户长期画像与sku的匹配度: Gender, C3, PW, EXT
  - 用户与用户在过去有行为的同c3/c2的sku的绝对价格偏好的相对差值：(clk, follow, cart, ord) * (c, d, w, m)
  - 用户的价格quantile偏好匹配： (ord, cart) * (all, c3) * (mean, median, min, max)
  - 好评率偏好匹配：all, cid3
  - 基于cart的用户各c3下的自营/pop偏好匹配：tf, tf-idf

交互特征：基于行为的相关

- 基于用户行为的相关交互特征：$rel(u,v) = \mathbf{u}_{c3}^{\mathrm{T}}\mathbf{M}\mathbf{v}_{c3}$
  - $\mathbf{u}_{c3}$：用户的c3偏好，时间窗口分别为1c、10c、20c、1d、7d、1m
  - $\mathbf{v}_{c3}$：商品的c3向量（one-hot）
  - $\mathbf{M}$：基于行为（共点击、共购买）的相关矩阵

交互特征：用户的匹配序列特征

- 用户近10次Click行为和当前商品的匹配序列特征 $\mathbf{v} = [v_0, \dots, v_9]$
  - Item level：SKU, C3, PW, related C3
  - 取值范围：$v_i = \begin{cases} 0, & \text{不匹配} \\ 1, & \text{匹配} \\ -1, & \text{无第}i\text{次行为} \end{cases}$

DNN模型中的Embedding特征

- 输入：用户近2天的c3, c2, pw, brand, shop, title term等
- 同时考虑dense特征和embedding特征的DNN模型
- Title term的长尾过滤
  - 目的：去除噪声，减小模型大小（减小至~30%）
  - 方法：（1）过滤低频词汇，保留>30的词（2）将词分为纯汉字和其他，保留top 50%纯汉字，top 30%其他
- Embedding的使用
  - 直接做dense特征
  - 构造复合特征：$\mathbf{u}^{\mathrm{T}}\mathbf{v}, \cos(\mathbf{u},\mathbf{v}), |\mathbf{u}-\mathbf{v}|, (\mathbf{u}-\mathbf{v})^{2}$
- Term长尾过滤 + Embedding复合特征的DNN模型，在发现好货
  - 人均点击+6.56% (p=0.000), 人均引流uv+7.06% (p=0.000), 转化率+7.68% (p=0.000), uv价值+8.21% (p=0.001)

内容素材特征：以发现好货为例

京东

- 新的信息：文本，达人作者等
- 素材文本单边特征
  - 正文长度、发布距今时间
  - Title/text是否包含品牌词、产品词
  - 是否包含jd, pop, 新品，奢品, 全球购, plus, JD精选, 旗舰店sku
  - Term的(7d, 30d) * (exp, click, order) * (uv, pv, uniq_uv, uniq_pv, ctr, cvr, ctr_lcb, cvr_lcb)的最大值
  - Term覆盖的(上柜sku, jd, pop, 新品，奢品, 全球购, plus, JD精选, 旗舰店sku, 收藏sku)数量和占比的最大值

内容素材特征：以发现好货为例

京东

- 素材作者单边特征（ c3 ）
  - 文章总数、占比
  - 30天：(feedback, exp, clk, up, share, cart, ord, clk_uv) * (max, min, mean, sum)
  - 30天：(ctr, ctr_lcb, uctr, uctr_lcb) * (max, min, mean, sum)
- Term交互
  - (click, cart) * (2w, 3m) * (title, sku title, text) * (count, cos)

Highlights

- Dense特征
  - 细粒度的信息提取：用户的自营、价格/购买力、促销等偏好画像及交互
  - 基于置信区间的特征correction
  - 最近10次点击的交互序列特征提取
  - 基于行为的相关特征提取
- Embedding特征
  - Term长尾过滤与复合特征的构建

特征上线：检查和监控

京东

特征评估：数据特征重要性估计

京东

- 数据集中的特征重要性
  - 特征在数据集中自身的重要性，模型无关
  - 计算：特征与label的相关性
- 模型中的重要性
  - 在学习到的模型中，特征的重要性
  - 接下来介绍
- 二者的一致性分析

特征评估：GBDT模型的特征重要性估计

京东

- 特征重要性的2种计算方法
  - 信息增益：特征在所有决策树中的信息增益的平均值
  - 权重：特征在所有决策树中出现的次数
- 使用方法
  - 新加入的特征是否进入特征重要性的top
  - 新加入的特征能否导致模型离线指标显著提升

特征评估：DNN模型的特征重要性估计

- 方法1：增加特征后，计算离线指标的增益
- 方法2

$$Imp(x_i) = \sqrt{Var\{E[f(\mathbf{x}|x_i)]\}} = \sqrt{Var\left(\left\{\frac{1}{n}\sum_{k=1}^{n} f\left(\mathbf{x}^{(k)}\middle|x_i = x_i^{(j)}\right)\right\}_{j=1,2,\dots,n}\right)}$$

- 思路：改变某特征的值，导致模型结果的变化越大，特征越重要
- 算法
  - 对每一维特征x_i，
    - 对其每个取值$x_i^{(j)}$，计算排序函数输出结果的期望$E_i^{(j)}, j = 1,2,\dots,n$
    - 计算$\left\{E_i^{(j)}\right\}_{j=1,2,\dots,n}$的标准差，得到其重要性

特征评估：DNN模型的特征重要性估计

- 方法3

$$Imp(x_i) = \mathrm{cov}(x_i, E[f(\mathbf{x}|x_i)])$$

$$= \mathrm{cov}\left(\left\{\left(x_i^{(j)}, \frac{1}{n}\sum_{k=1}^{n} f\left(\mathbf{x}^{(k)}\middle|x_i = x_i^{(j)}\right)\right)\right\}_{j=1,2,\ldots,n}\right)$$

- 思路：某特征的值和模型结果的变化越相关，特征越重要
- 算法
  - 对每一维特征x_i，
    - 对其每个取值$x_i^{(j)}$，计算排序函数输出结果的期望$E_i^{(j)}$，$j = 1,2,\ldots,n$
    - 计算$\left\{\left(x_i^{(j)}, E_i^{(j)}\right)\right\}_{j=1,2,\ldots,n}$的皮尔逊相关系数的绝对值，得到其重要性

特征上线：线上线下的数据一致性

京东

- 全量feature log训练模型
- 确保线下使用的特征和线上的计算是一致的

特征监控：异常特征的识别和监控

京东

- Observer特征监控
  - 计算0/-1的占比, mean, variance, min, max

谢 谢 ！

王帅强 & 丁卓冶
京东零售-搜索与推荐平台部-数据科学实验室-推荐科学组

京东