初赛问题一：区域销售预测

问题描述

为了提升配送时效，优化顾客购物体验，京东采用了分地理区域进行销售配送的方式。对商品在各个区域内销量的精准预测至关重要。我们将提供京东商品在多个销售区域过去**两年**的销售情况[1]，参赛队伍需要对目标商品在各个销售区域内未来**一个月**的销量进行预测。

评测方法

对商品销量的预测将直接作为库存计划的输入量。在库存计划中，对预测精度的需求往往不仅是对均值的预测，而是对销量分布的整体预测，即销量分布上不同分位点对应的值。为简化问题，我们根据库存计划的需求事先确定了每个商品所需要预测的目标分位点，我们将用分位点误差公式（quantile loss function）来衡量参赛队伍对未来销量的预测准确度。

具体而言，预测误差评测公式为

$$L = \sum_{ij} L_{ij}(q_i)$$

$$L_{ij}(q_i) = \frac{\sum_t \left[q_i \left(y_{ijt} - \hat{y}_{ijt}^{q_i} \right)^+ + (1 - q_i) \left(\hat{y}_{ijt}^{q_i} - y_{ijt} \right)^+ \right]}{T \sum_t y_{ijt}}$$

这里$L_{ij}(q_i)$为对商品i在销量区域j的q_i分位点的预测误差（q_i事先给定），y_{ijt}是商品i第t天在销量区域j的实际销量，$\hat{y}_{ijt}^{q_i}$是对于q_i分位点相应的预测销量值，T为

[1] 请注意我们没有提供 6 月与 11 月的销售数据，因为大促期间的销售情况与平时非常不同。最终评测中我们也选取了非大促时段。

需要预测的总天数，这里使用$\frac{1}{\sum_t y_{ijt}}$作为第i个商品在销售区域j衡量时的权重[2]，减小销量大小对目标函数的影响。

数据描述

训练数据集

1.sku_info.csv

SKU 基本信息

Column_name	Sample_data	Description
item_sku_id	68	item unique identification number
item_first_cate_cd	5	item first level category code, e.g. shoes
item_second_cate_cd	21	item second level category code, e.g. sports shoes
item_third_cate_cd	405	item third level category code, e.g. running shoes
brand_code	1697	item brand code

[2]为了避免预测算法向销量高的产品倾斜，我们将使用销量的倒数作为权重，增加销量较低的商品在最后评测中的重要程度。 $\sum_t y_{ijt}$不会为零。

2. sku_attr.csv

SKU 属性信息

Column_name	Sample_data	Description
item_sku_id	434	item SKU ID
attr_cd	345	attribute code, e.g. color
attr_value_cd	1681	attribute value code, e.g. red. If an item is colorful, there would be multiple color attribute values.

3.sku_prom.csv

促销信息

Column_name	Sample_data	Description
item_sku_id	118	item SKU ID. If a promotion is applicable for all SKUs in a 3rd level category, then the item_sku_id is -999
item_third_cate_cd	18	item 3rd level category id
date	2016/01/01	date of promotion
promotion_type	6	a specific promotion type: e.g. direct discount, coupon, etc.

Notes: The promotion data is from Jan 1st 2016 to Dec 31st 2017. Data in

June and November are excluded.

4.sku_sales.csv

销售信息

Column_name	Sample_data	Description
item_sku_id	36	item SKU ID
dc_id	1	distribution center ID
date	2017/2/13	date
quantity	1	sales quantity of the day
vendibility	0	stock availability at the end of the day, 0 means no inventory left, otherwise 1
original_price	0.0373797	original price
discount	10	daily average discount= daily average transaction price/original price. Range is from 0 to 10, 9.5 means 5% discount.

Notes:

i. The sales data is from Jan 1st 2016 to Dec 31st 2017. Data in June and November are excluded.

ii. Null values of original_price and discount represent that there is no

sales or the price information is not accurately recorded.

iii. As transaction price is aggregated on DC level, the daily average discounts could be different among DCs.

测试数据集

1.sku_prom_testing_2018Jan.csv

2018 年 1 月促销信息

Column_name	Sample_data	Description
item_sku_id	118	item SKU ID. If a promotion is applicable for all SKUs in a 3rd level category, then the item_sku_id is -999
item_third_cate_cd	18	item 3rd level category id
date	2018/01/01	date of promotion
promotion_type	6	a specific promotion type: e.g. direct discount, coupon, etc.

Notes:

i. The promotion data is from Jan 1st 2018 to Jan 31st 2018.

ii. Average discount information is not provided for 2018 January.

2.sku_quantile.csv

目标预测分位点

Column_name	Sample_data	Description
item_sku_id	978	item SKU ID
target_quantile	96%	targeted quantile in the loss function

提交规则 Submission

参赛者需对 2018 年 1 月所有 1000 个 SKU 每天在每个 DC 的销量进行预测。提交的数据文件应为 csv 文件，文件大小不超过 5MB,英文逗号分隔，无 BOM 的 utf8 编码，包含列名，字段如下：

1. date，未来 31 天的日期，数据范围 1~31

2. dc_id，配送中心 ID，数据范围 0~ 5

3. item_sku_id，商品集合中的商品 ID，数据范围 1~1000

4. quantity，预测销量≥ 0。如预测值为 0，也需要在结果中包括。对于包含小数的预测结果，评测中仅保留小数点后两位，例如:预测结果为 3.146，将以 3.14 作为评分依据。

请勿在结果文件中包含重复的 date - dc_id - item_sku_id 记录。结果文件应该包含 31(day) *6(dc)*1000(sku) =186000 行记录。

表：提交结果数据格式

date	dc_id	item_sku_id	quantity
1	0	217	0
2	0	217	1.23
…	…	…	…