机器学习在商品匹配中的实践

刘洋@1号店

2017/4/22

商品匹配

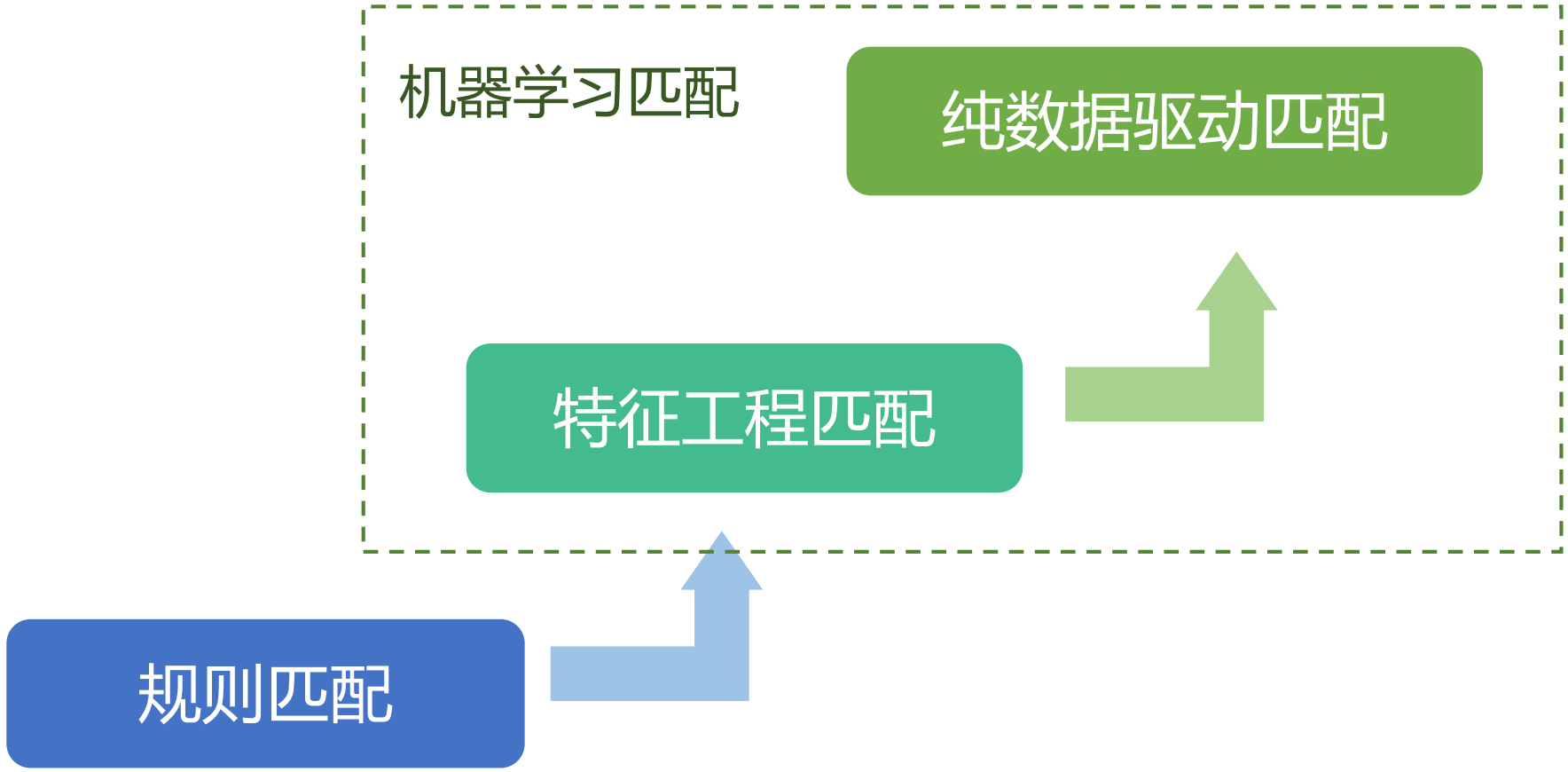
商品匹配问题特点

- 完全匹配，不是商品相似
- 短文本，对词敏感

1号店标题	千岁好 山茶油 5L/瓶
友商标题	千岁好有机山茶油5L 压榨一级油茶籽油 非转基因食用油5升

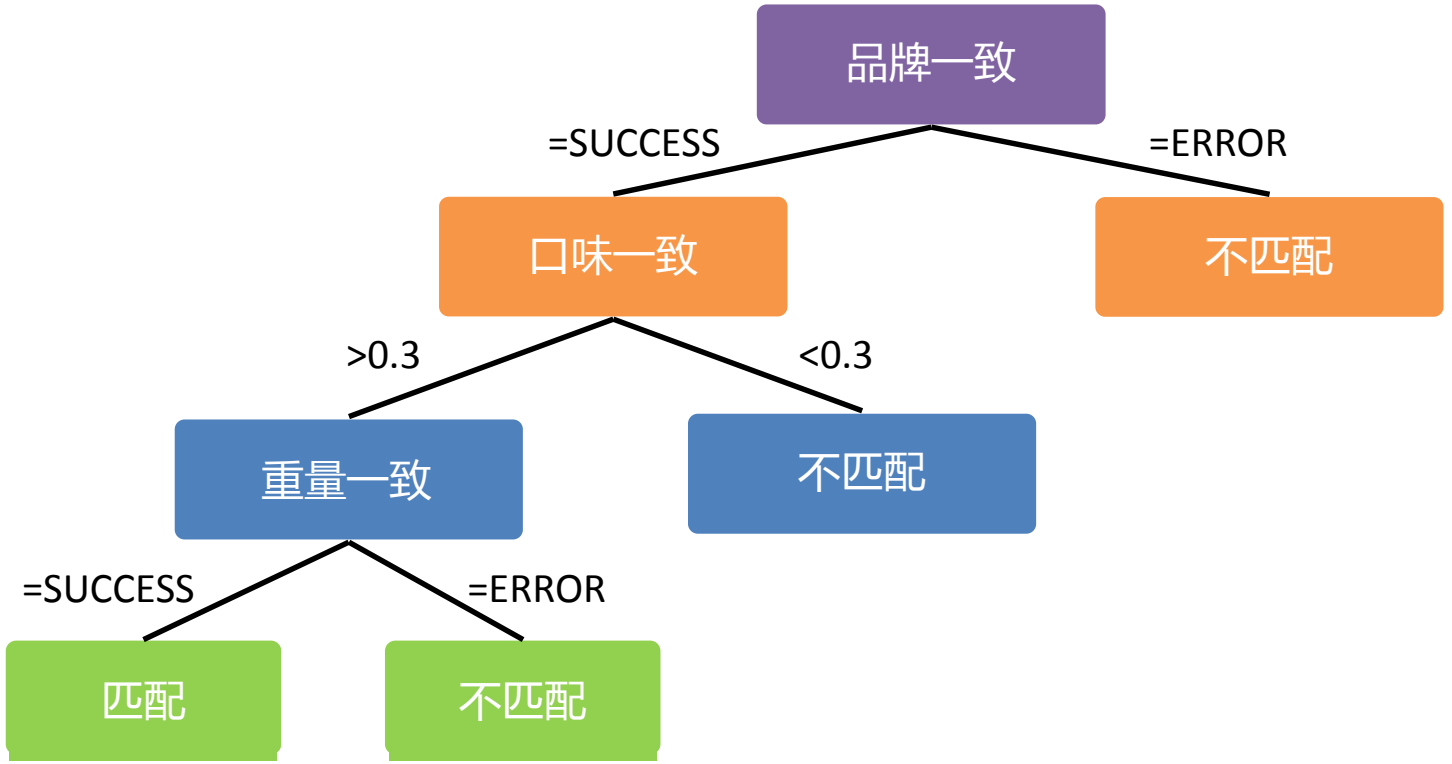
1号店标题	北纯 黑豆 360g/袋
友商标题	北纯有机黑豆360g 杂粮豆浆专用 新老包装随机

1号店商品匹配历程

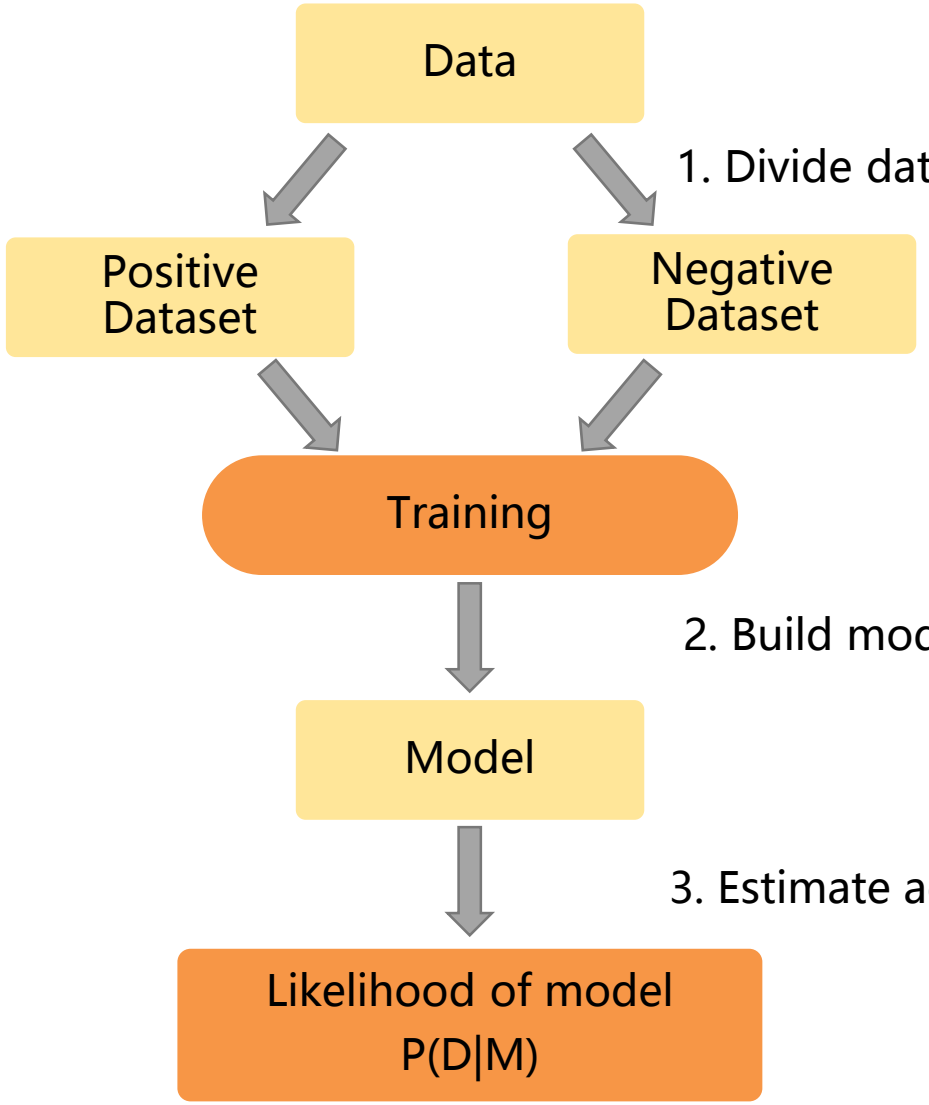


规则匹配

1号店标题	鲜得味 茄汁口味金枪鱼 180g 泰国进口
友商标题	泰国进口鲜得味金枪鱼罐头（茄汁味）无防腐剂180g



Supervised Learning

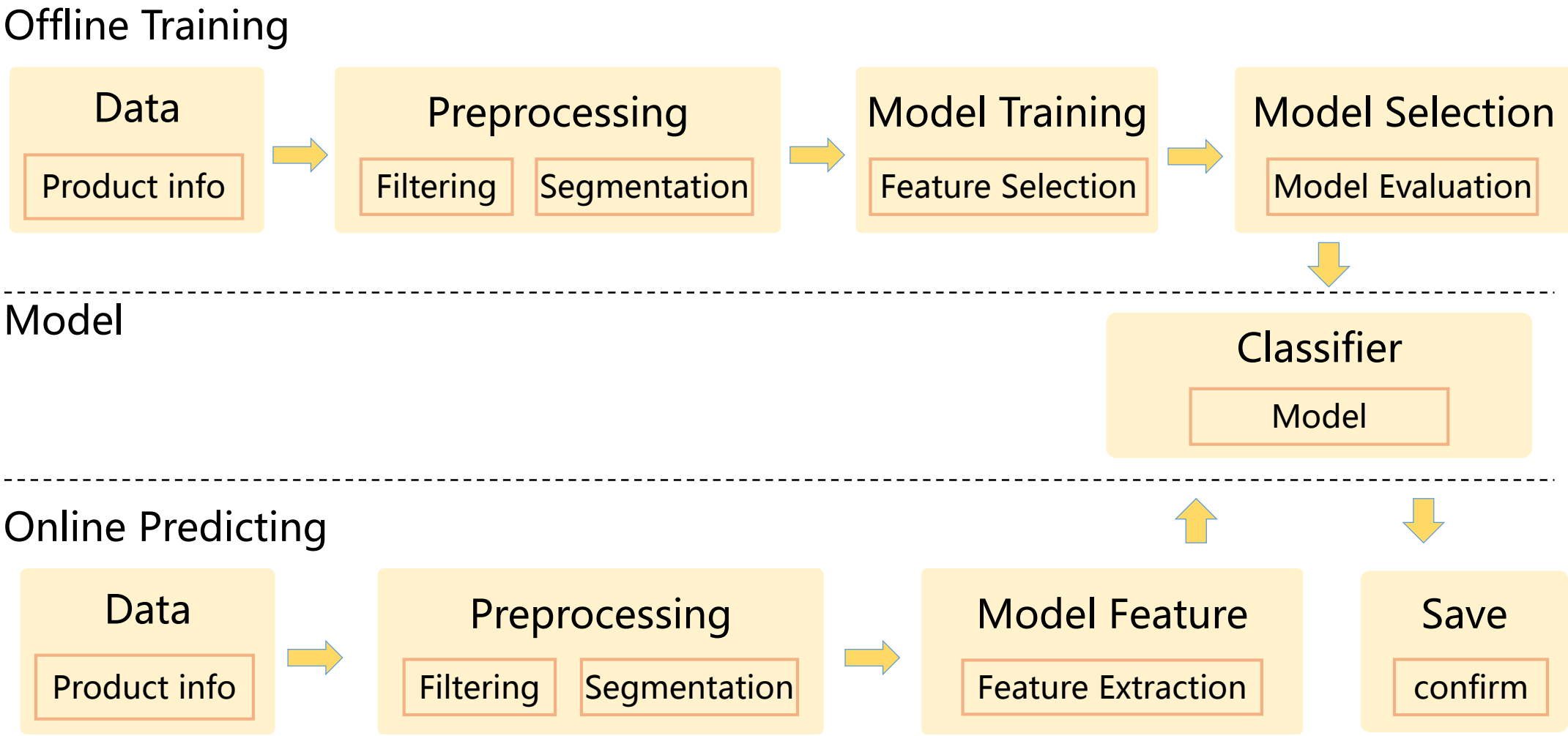


1. Divide data using *priori* known class labels

2. Build model by training

3. Estimate accuracy by calculating likelihood

机器学习匹配流程


Offline Training
Data
Product info
Preprocessing
Filtering
Segmentation
Model Training
Feature Selection
Model Selection
Model Evaluation
Model
Classifier
Model
Online Predicting
Data
Product info
Preprocessing
Filtering
Segmentation
Model Feature
Feature Extraction
Save
confirm

基于特征工程的商品匹配

Feature Engineering	
1	品牌打分
2	颜色打分
3	系列词打分
4	成分打分
5	地区打分
6	口味打分
7	型号打分
8	单位打分
9	价格打分
10	相同词比例
…	…

```
UNIT <= -1
|   MODEL <= 0: UNKNOWN (75239.0/10.0)
|   MODEL > 0
|   |   PRICE <= 0.955007
|   |   |   ELEMENT <= 0
|   |   |   |   FUNCTION <= 0
|   |   |   |   |   FUNCTION <= -1: UNKNOWN (29.0)
|   |   |   |   |   FUNCTION > -1
|   |   |   |   |   |   ELEMENT <= -1: UNKNOWN (35.0/1.0)
|   |   |   |   |   |   ELEMENT > -1
|   |   |   |   |   |   |   MODEL <= 0.8125
|   |   |   |   |   |   |   |   KEYWORD <= 0.125: CANDIDATE (10.0)
|   |   |   |   |   |   |   |   KEYWORD > 0.125
|   |   |   |   |   |   |   |   |   KEYWORD <= 0.347826: UNKNOWN (3.0)
|   |   |   |   |   |   |   |   |   KEYWORD > 0.347826: CANDIDATE (2.0)
|   |   |   |   |   |   |   MODEL > 0.8125
|   |   |   |   |   |   |   |   SIMILAROLD <= 0.8125
|   |   |   |   |   |   |   |   |   MODEL <= 0.875: UNKNOWN (42.0/2.0)
|   |   |   |   |   |   |   |   |   MODEL > 0.875
|   |   |   |   |   |   |   |   |   |   MODEL <= 2
|   |   |   |   |   |   |   |   |   |   |   SIMILAROLD <= 0.533333
|   |   |   |   |   |   |   |   |   |   |   |   KEYWORD <= 0.416667
|   |   |   |   |   |   |   |   |   |   |   |   |   KEYWORD <= 0.291667: CANDIDATE (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   KEYWORD > 0.291667: UNKNOWN (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   KEYWORD > 0.416667: CANDIDATE (5.0)
|   |   |   |   |   |   |   |   |   |   |   SIMILAROLD > 0.533333: UNKNOWN (3.0)
|   |   |   |   |   |   |   |   |   |   MODEL > 2
|   |   |   |   |   |   |   |   |   |   |   PRICE <= 0.494309
|   |   |   |   |   |   |   |   |   |   |   |   SIMILAROLD <= -1
|   |   |   |   |   |   |   |   |   |   |   |   |   KEYWORD <= -1: UNKNOWN (19.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   KEYWORD > -1: CANDIDATE (27.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   SIMILAROLD > -1: UNKNOWN (62.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   PRICE > 0.494309: UNKNOWN (42.0/9.0)
|   |   |   |   |   |   |   |   SIMILAROLD > 0.8125
|   |   |   |   |   |   |   |   |   SIMILARNEW <= 0.5
|   |   |   |   |   |   |   |   |   |   SIMILARNEW <= 0.352941: UNKNOWN (4.0/1.0)
```

纯数据驱动的商品匹配

商品标题描述词**维度高**且**稀疏**。

1号店标题　盼盼 法式软面包 奶香味 200g /袋

422284:YHD_BRAND:盼盼
17510:YHD_CATEGORY_KEYWORD:软面包
635361:YHD_UNIT_WEIGHT:200000MG
5848:YHD_UNIT_QUANTITY:1
553547:YHD_LASTWORD:法式
317607:YHD_TASTE:奶香味

友商标题　盼盼法式小面包奶香味200g 零食早餐面包蛋糕

1058153:OPPON_BRAND:盼盼
650575:OPPON_CATEGORY_KEYWORD:早餐
650996:OPPON_CATEGORY_KEYWORD:面包
658432:OPPON_CATEGORY_KEYWORD:蛋糕
659318:OPPON_CATEGORY_KEYWORD:零食
1271230:OPPON_UNIT_WEIGHT:200000MG
641717:OPPON_UNIT_QUANTITY:1
1153623:OPPON_LASTWORD:小
1189416:OPPON_LASTWORD:法式
953476:OPPON_TASTE:奶香味

纯数据驱动匹配采用的模型

Factorization Machine（因子分解机）

- 把标题词作为特征，通过隐向量进行表示。
- 适合解决稀疏矩阵特征组合问题。
- 模型高效，可在线性时间训练和预测。
- 应用面广，还可以用于：
  - 计算搜索查询词和商品间的相关性
  - 推荐相似商品

Factorization Machine 模型

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

w_0 ： 全局偏置

w_i ： 线性项权重

v_i ： 隐向量

二阶多项式模型

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} x_i x_j$$

训练样本示例

1 480773:1 13333:1 633837:1 5848:1 55446:1 317591:1 317520:1 8829:1 1116642:1 649202:1 650575:1 648085:1 636763:1 1269706:1 641717:1 691315:1 953460:1 644698:1

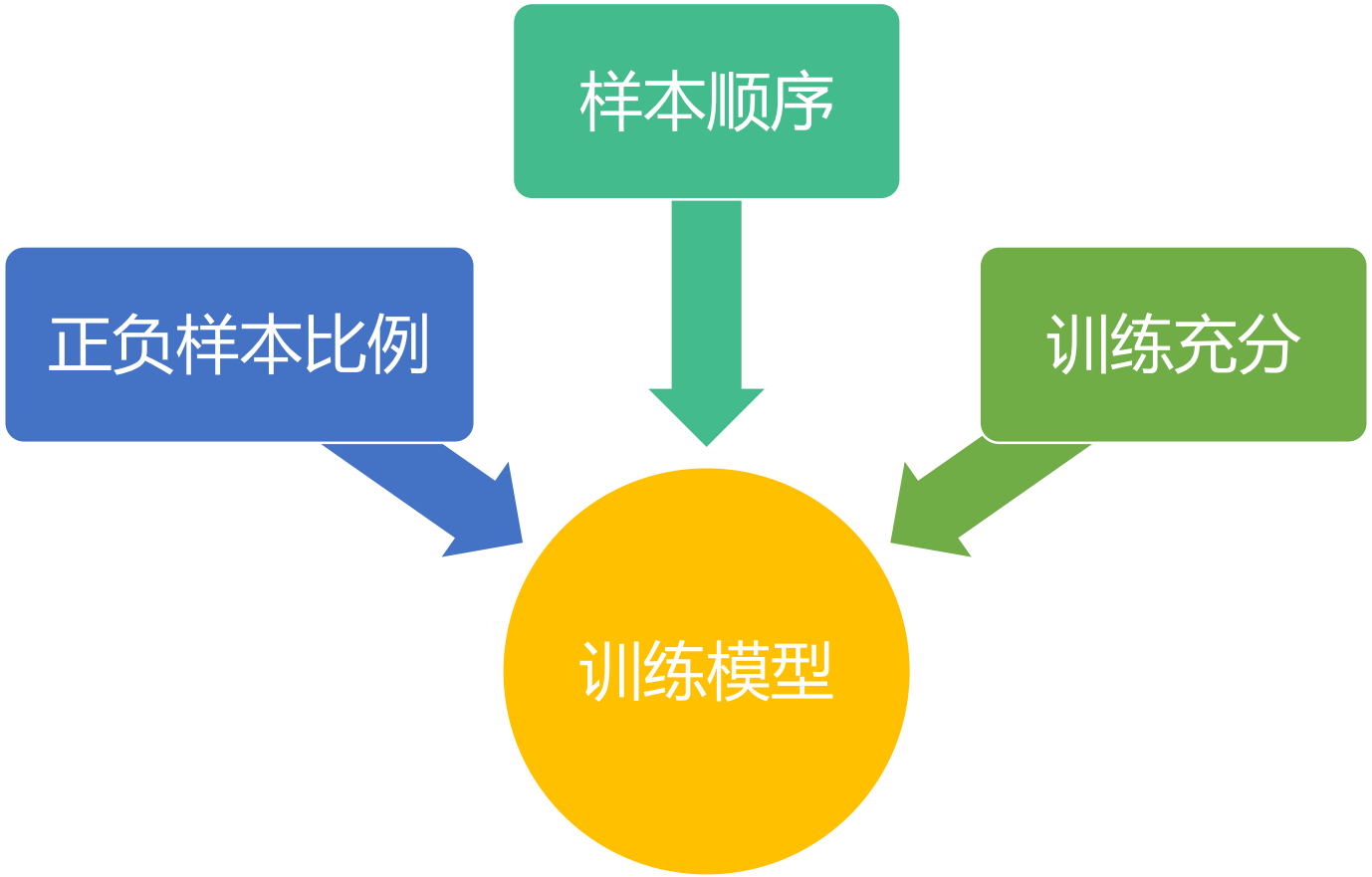
480773:YHD_BRAND:康师傅
13333:YHD_CATEGORY_KEYWORD:苏打
633837:YHD_UNIT_WEIGHT:125000MG
5848:YHD_UNIT_QUANTITY:1
55446:YHD_MODEL:3+2
317591:YHD_TASTE:柠檬味
317520:YHD_TASTE:甜
8829:YHD_FUNCTION_SERIES:清新

1116642:OPPON_BRAND:康师傅
649202:OPPON_CATEGORY_KEYWORD:苏打
650575:OPPON_CATEGORY_KEYWORD:早餐
648085:OPPON_CATEGORY_KEYWORD:早点
636763:OPPON_INGREDIENT:夹心饼干
1269706:OPPON_UNIT_WEIGHT:125000MG
641717:OPPON_UNIT_QUANTITY:1
691315:OPPON_MODEL:3+2
953460:OPPON_TASTE:柠檬味
644698:OPPON_FUNCTION_SERIES:清新

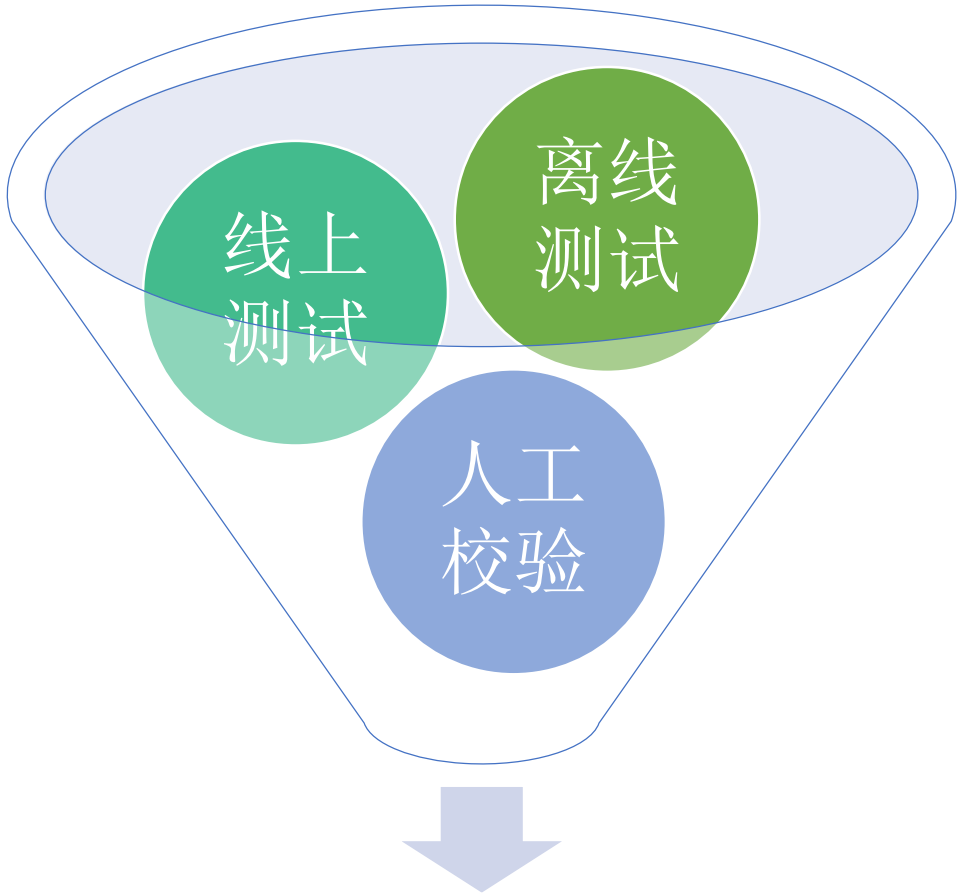
训练样本校验

- 切分词正确。
- 确认正样本、负样本。
- 负样本是否和正样本有冲突。
- 拿到样本全部特征。
- 商品类目分布情况。

训练技巧



结果评判



模型性能

FM预测结果

	准确率	召回率	F1
初步预测	0.3723	0.2448	0.2954

思考

FM模型真的适合解决这个问题吗？

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

理解商品匹配问题

纯数据驱动商品匹配优化（1）

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

去除**模型线性项**部分

- 单从一个商品的特征无法判断两商品是否匹配。

	准确率	召回率	F1
初步预测	0.3723	0.2448	0.2954
优化 （1）	0.3639	0.2587	0.3025

纯数据驱动商品匹配优化（2）

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

交叉项限定只在两个商品的特征间进行组合

- 单独一个商品的特征组合无法判断两商品是否匹配。

	准确率	召回率	F1
初步预测	0.3723	0.2448	0.2954
优化 （1）	0.3639	0.2587	0.3025
优化 （2）	0.4848	0.3706	0.4201

纯数据驱动商品匹配优化（3）

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

交叉项限定只在两个商品同一词性的特征间进行组合

- 两个商品不同词性的特征组合无法判断两商品是否匹配。

	准确率	召回率	F1
初步预测	0.3723	0.2448	0.2954
优化 （1）	0.3639	0.2587	0.3025
优化 （2）	0.4848	0.3706	0.4201
优化 （3）	0.6412	0.5291	0.5798

纯数据驱动商品匹配优化（4）

1号店商品特征

YHD_BRAND:康师傅
YHD_CATEGORY_KEYWORD:苏打
YHD_UNIT_WEIGHT:125000MG
YHD_MODEL:3+2
YHD_TASTE:柠檬味
YHD_FUNCTION_SERIES:清新

友商商品特征

OPPON_BRAND:康师傅
OPPON_INGREDIENT:夹心饼干
OPPON_UNIT_WEIGHT:125000MG
OPPON_MODEL:3+2
OPPON_TASTE:柠檬味
OPPON_FUNCTION_SERIES:清新

OPPON_BRAND:康师傅
OPPON_INGREDIENT:夹心饼干
OPPON_UNIT_WEIGHT:125000MG
OPPON_MODEL:3+2
OPPON_TASTE:巧克力味
OPPON_FUNCTION_SERIES:清新

纯数据驱动商品匹配优化（4）

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

两个商品的相同特征使用同一隐向量进行表示

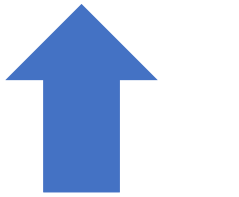
- 两商品中相同特征构成的组合项打分要高。

	准确率	召回率	F1
初步预测	0.3723	0.2448	0.2954
优化 （1）	0.3639	0.2587	0.3025
优化 （2）	0.4848	0.3706	0.4201
优化 （3）	0.6412	0.5291	0.5798
优化 （4）	0.6862	0.6270	0.6553

纯数据驱动匹配的优化过程

人工加了先验知识，使模型更有针对性的去解决问题。

纯数据驱动匹配



特征词→特征向量

干预难

THANK YOU！