简述

We use TextBlob model which is used in Sentiment Analysis to quantify the reviews. The result of TextBlob Sentiment Analysis is returned as a tuple in the form of(polarity, subjectivity).

…(结果(polarity, subjectivity))

The score of bias is a floating point number with a range of [-1.0, 1.0]. Positive Numbers mean positive and negative Numbers mean negative.Alfred is a floating point number with a range of [0.0, 1.0], where 0.0 is objective and 1.0 is subjective.

We use Light gradient accelerator (Light) Gradient Boosting Machine (LightGBM) which is widely used in data mining tasks to get the weights to build model. We take star\_rating as the output and the other features as the input to train the model.

In order to obtain more data correlation information, we used feature engineering to process the data set, and then put it into the model for training.Feature engineering often plays an important role in the field of data mining. Different feature engineering can obtain different feature sets. A good feature engineering can reveal more relevant information in the data set, thus making the model more accurate.

基础模型

LightGBM是对GBDT的高效实现，主要包括两个本分，一个是GOSS，另一个是EFB。

GOSS算法(算法1)排除了大部分具有小梯度的数据，只使用剩余的数据进行信息增益估计，LightGBM研究[1]表示:具有较大梯度的样本在计算信息增益的时候扮演着更加重要的角色，GOSS可以通过更加小规模的数据来获得非常精准的信息增益计算。

EFB算法(算法2和算法3)通过将互斥的特征捆绑在一起，来减少特征数目。互斥特征意味着它们几乎很少同时出现在非零值，并且LightGBM也表明:找到最优化特征捆绑是NP问题，但是贪心算法能够获得非常好的近似概率

算法1 GOSS

输入:训练数据 I,迭代次数d,大梯度数据采样率a,小梯度数据采样率b,损失函数loss,弱学习器L。

models←{ }，fact← (1 -a)/b

1) topN ← a × len(I) ,randN← b × len(I)

2) for i=1 to d do.

3) preds ← models. predict(1)

4) g ← loss(I ,preds) ,w←{1,1,⋯}

5) sorted ← GetSortedIndices( abs (g) )

6) topSet ← sorted [ 1: topN]

7) randSet ← RandomPick( sorted [ topN:len(I)] ,randN)

8) usedSet ← topSet + randSet

9) w[ randSet] ×= fact //给小梯度数据分配权重参数

10) newModel←L( I[ usedSet], - g[ usedSet] ,w[ usedSet])

11) models. append( newModel)

算法2 Greedy Bundlig。

输入：特征集”,最大冲突计数K;

输出：捆绑集bundles.

构造图G

1) searchOrder ←G.sortByDegree()

2) bundles←{},bundlesConflict←{}

3) for i in searchOrder do

4) needNew ← True ，

5) for j =1 to len (bundles) do

6) cnt ← ConflictCnt( bundles[j] , F[i])

7) if cnt + bundlesConflict[i] ≤ K then

8) bundles[j].add(F[i]),needNew ←False

9) break

10) if needNew then

11) 将F[i]作为新的捆绑bundle加入到bundles中

算法3 Merge Exclusive Features。

输入：数据的数目numData,一束关于互斥特征的捆绑特征F

输出：新的直方图newBin,直方图区间binRanges。

1) binRanges ← |0| , totalBin ← 0

2) for f in F do

3) totalBin += f. numBin

4) binRanges. append(totalBin)

5) newBin←new Bin( numData)

6) for i =1 to numData do

7) newBin[i]←0

8) for j =1 to len(F) do

9) if F[j].bin[i] ≠0 then

10) newBin[i]←F[j]. bin[i] + binRanges[j]

[1] KE G, MENG Q, FINLEY T, et al. LightCBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 2017 Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 3146 -3154.

特征工程

将通过分析数据集中的各个特征在LightGBM中重要性表现来进行特征工程操作.首先，将预处理后的数据集放入到LightGBM训练的结果如表1，表2，表3

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| customer\_id | 0 |
| review\_id | 0 |
| product\_id | 0.0365 |
| helpful\_votes | 0.0412 |
| total\_votes | 0.1380 |
| vine | 0 |
| verified\_purchase | 0.0285 |
| review\_body | 0.2238 |
| review\_date | 0.2 |
| year | 0.0031 |
| month | 0.0539 |
| rate | 0.2746 |

表1. hair\_dryer的特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| customer\_id | 0 |
| review\_id | 0 |
| product\_id | 0 |
| helpful\_votes | 0.0389 |
| total\_votes | 0.1229 |
| vine | 0 |
| verified\_purchase | 0.0286 |
| review\_body | 0.3299 |
| review\_date | 0.2131 |
| year | 0 |
| month | 0.0819 |
| rate | 0.1844 |

表2. microwave的特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| customer\_id | 0 |
| review\_id | 0 |
| product\_id | 0.0095 |
| helpful\_votes | 0.0444 |
| total\_votes | 0.1396 |
| vine | 0.0047 |
| verified\_purchase | 0.0253 |
| review\_body | 0.2857 |
| review\_date | 0.2015 |
| year | 0 |
| month | 0.0555 |
| rate | 0.2333 |

表3. pacifie的特征值权值

从三类产品在LightGBM模型中学习情况可以看到，有部分的特征权值几乎为0，这些对于模型来说，可能是多余的,综合三种产品情况，删除customer\_id，review\_id，product\_id，year，vine。再次使用模型进行训练，得到表4，表5，表6

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0365 |
| total\_votes | 0.1412 |
| verified\_purchase | 0.0253 |
| review\_body | 0.2253 |
| review\_date | 0.2063 |
| month | 0.0523 |
| rate | 0.3126 |

表4. hair\_dryer的处理后特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0389 |
| total\_votes | 0.1229 |
| verified\_purchase | 0.0286 |
| review\_body | 0.3299 |
| review\_date | 0.2131 |
| month | 0.0819 |
| rate | 0.1844 |

表5. microwave的处理后特征值权值

|  |  |
| --- | --- |
| 特征名称 | 权值 |
| helpful\_votes | 0.0428 |
| total\_votes | 0.1428 |
| verified\_purchase | 0.0301 |
| review\_body | 0.2936 |
| review\_date | 0.2 |
| month | 0.0571 |
| rate | 0.2333 |

表6. pacifie的处理后特征值权值

删除后,模型的特征权值变化不大，并且F1值提升不明显。然后我们观察数据发现，有些评论存在一定错误性，如给好评，却打了一星，给差评，却给了五星。

|  |  |  |
| --- | --- | --- |
| review\_id | Start\_rating | review\_body |
| R134FUK2D9TQU6 | 1 | I have used the dryer several times and it works great. I had questions which were answered promptly by other customers which was helpful in making my decision. Definitely recommend. |
| R32375P4DJ9UH1 | 4 | After one year the turntable plate cracked while we were making a bag of popcorn. Very strange indeed! Now trying to find a replacement as I really miss the food items turning while cooking! Other than that it is still working in top notch condition. UPDATE: This microwave died after only 2 years. Was hoping to get more out of it then we did. |
| RQ73ZP5XYKGME | 2 | Great pacifiers. Selection wasn't so great though. Gave me all sports themed and I have a little girl and we don't do sports lol. Guess we will just use these at home |
| r1ikvd52i8f8xp | 2 | if given the choice, i'd probably go with a different unit. while the bells and whistles are really cool when they work, the price is probably not worth your hassle. we are on our third unit, since the first two have failed. the first one stopped connecting entirely, and in the second, the charging housing on the unit has come lose. to their credit, withings has great customer service and despite this being a gift (hence no proof of purchase) they've replaced our unit (twice now). |
| rvj638trfhuj4 | 2 | works great! baby slept perfect. ! |
| R1HI3QGXJQ2RUT | 5 | We owned these from the store and they are exactly the same. Too bad my grandson decided he was done with pacifiers one week later |
| ⋯ | ⋯ | ⋯ |

我们删除完那些”一星好评”和“五星差评”后，三类产品F1值分别从0.8766,0.8044,0.9032提升至0.8921,0.8369,0.9351.