目录

3 特征衍生

特征衍生(feature derivation) [1] 是特征工程中重要的一部分, 其典型的应用场景是所获得的原始数据不能直接用于模型构建, 需要通过特征衍生获得可用于建模的数据. 3.3节给出了一个案例.

此外, 即便是在可以直接建模的数据中, 特征衍生也有用武之地. 例如, 利用原始数据集中用户的留存天数和消费总金额, 可以衍生用户在留存期间平均每天的消费金额. 相对于消费总金额, 所衍生的平均每天消费金额可以更客观地衡量用户的消费能力或对商家的忠诚程度. 根据业务需求和知识背景进行特征衍生, 可能对提升模型性能大有益处, 特别是对于那些不具有自动特征学习功能的模型, 因为特征衍生增加了建模信息.

3.1 一阶特征衍生

特征衍生应根据实际需求开展. 常见的一阶衍生特征有多种, 例如:

(1) 数量特征. 如, 计算总和、乘积、除数、差值、最大值、最小值、平均值、标准差、极差等. 例如, 在天猫用户重复购买案例中, 可以计算用户在平台的购物总金额, 单次消费的最高金额等.

(2) 类别特征. 通过分箱或分类产生. 例如, 按照用户平均每月消费金额, 将用户划分为深度用户和浅层用户两类.

(3) 时间特征. 例如, 可以衍生用户两次购物之间的平均时间间隔, 反映其购物频次.

[1]特征衍生也称为属性构建(attribute construction), 可以看作是数据变换的一种. 由于它可能衍生出全新的数据集(与原始数据集中的变量完全不同), 因此本书将特征衍生放入数据集成.

(4) 空间特征. 例如, 若数据中包含商家的地理位置, 则可以获得商家所在的省份和城市等空间信息.

(5) 文本特征. 例如, 从用户对商品的文字评价中提取用户满意度等信息.

3.2 二阶与高阶特征衍生

除了衍生一阶特征, 还可以衍生交叉特征, 即将某些特征进行组合. 交叉特征可以反映变量的交互作用, 有助于探索自变量与因变量之间更为复杂的作用关系, 并提升模型效果. 一般, 优先考虑二阶特征组合, 其衍生方法应依据特征的类型:

(1) 对于定量特征与定量特征, 应将两个特征相乘.

(2) 对于定性特征与定性特征, 应对两个特征作笛卡尔积, 即采用两个定性特征的所有取值组合.

(3) 对定性特征与定量特征, 应先对其中一个特征做变换, 再根据前述两种情况衍生交叉特征.

在二阶特征组合基础上, 可以获得三阶特征组合. 以此类推, 可以得到更多高阶组合. 所衍生的高阶特征越多, 可能发掘出数据中更多的非线性结构. 但是, 过多的特征极易造成维数灾难, 影响建模的速度和效果. 一种较好地控制维数的高阶特征衍生方法是通过核函数实现, 例如, 多项式核函数

$$\kappa(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a}^T \boldsymbol{b} + c)^d,$$

其中c和d是需要设置的参数, c是平移量, d是阶数. 对于具有二维特征的数据$\boldsymbol{X} = (X_1, X_2)$, 若取$d = 1$, 可得到二阶特征组合:

$$\kappa(\boldsymbol{X}, \boldsymbol{X}) = X_1^2 + X_2^2 + c.$$

若取$d = 2$, 可得到四阶特征组合:

$$\kappa(\boldsymbol{X}, \boldsymbol{X}) = X_1^4 + X_2^4 + 2X_1^2X_2^2 + 2cX_1^2 + 2cX_2^2 + c^2.$$

3.3 案例分析

在天猫用户重复购买预测案例[2]中, 研究目标是: 对于给定的商家, 预测“双十一”新用户在未来会成为忠实客户的可能性, 即预测新用户在未

[2]案例来源于阿里云天池的学习赛, 具体介绍可参看https://tianchi.aliyun.com/competition/entrance/231576/introduction, 天池平台(2020) 分析了该案例.

	user_id	item_id	cat_id	seller_id	brand_id	time_stamp	action_type
0	328862	323294	833	2882	2661.0	829	0
1	328862	844400	1271	2882	2661.0	829	0
2	328862	575153	1271	2882	2661.0	829	0
3	328862	996875	1271	2882	2661.0	829	0
4	328862	1086186	1271	1253	1049.0	829	0

Figure 3.1: 用户行为日志数据集示意图

表 3.1: 用户行为日志数据集变量信息

序号	变量名称	变量含义	变量类型	取值(范围) 及其含义
1	user_id	用户ID	定类	-
2	item_id	商品ID	定类	-
3	cat_id	商品所属品类ID	定类	-
4	seller_id	商家ID	定类	-
5	brand_id	商品品牌ID	定类	-
6	time_tamp	行为发生时间	定序	0511至1112 (格式为mmdd)
7	action_type	用户行为类型	定类	0:点击; 1:加入购物车; 2:购买; 3:收藏

来6个月内再次在该商家购物的概率. 该案例提供了3个数据集, 其中用户行为日志数据集为预测用户是否复购提供了最主要的信息. 图3.1显示了前5行用户行为日志数据. 每一列数据为一个变量, 表3.1给出了变量的相关信息. 每一行数据对应一位用户在某一个时间、进入某一商家、对某一商品所进行的点击、加购物车、购买和收藏等行为. 可以看到, 一个用户和一个商家的组合在用户行为日志数据集的许多行中都出现了. 为了预测建模, 应将一个用户和一个商家的组合作为一个样本观测, 此时, 需要提取该组合在多行数据中的信息, 即作特征衍生.

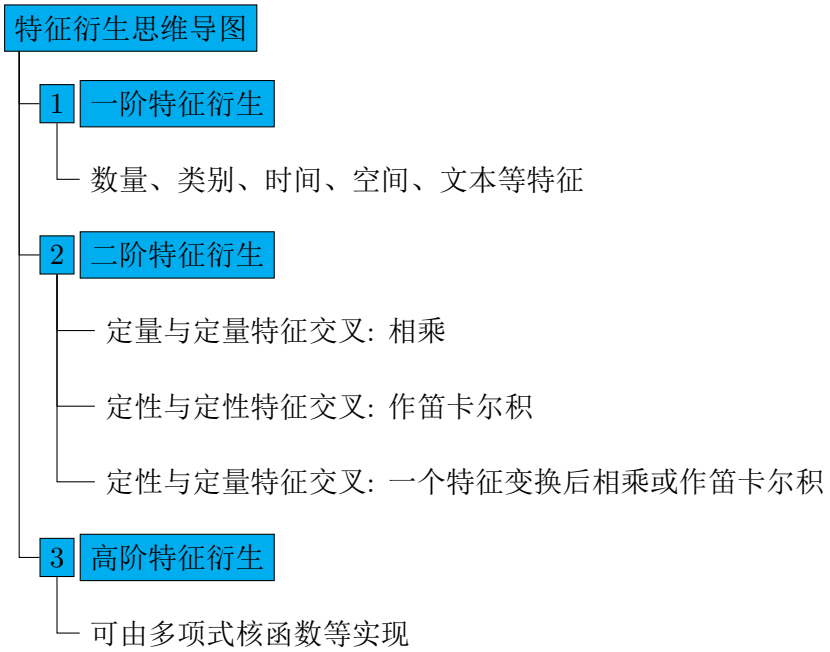
首先衍生一阶特征. 一种思路是: (未完待续)

表 3.2: 特征衍生Python函数

方法	函数名称与所属模块	重要参数	应用提示
由多项式核函数衍生特征	PolynomialFeatures (scikit-learn库preprocessing)	(1) degree: 设置最高阶数 (2) interaction_only: 是否仅衍生交互特征, 默认值为False. 若设置为True, 则高阶项中仅出现两个特征的交互项 (3) include_bias: 是否包含常数项, 默认值为True. 若设置为False, 则不会出现0次方项(常数项)	函数不允许缺失值, 应处理缺失值

本章小结

(1) 思维导图



(2) Python函数

表3.2列出了由多项式核函数衍生高阶特征的Python函数.

References

[1] 天池平台(2020) 天池大赛赛题解析–机器学习篇. 北京: 电子工业出版社.