**Tast3 特征工程**

Tast2进行了数据的分析探索，现在要对数据进行处理。

常见的特征工程：数据清洗，分桶，特征标准化/归一化，特征筛选，特征构造，降维。

1. 数据清洗
   1. 异常值
2. 简单统计：如 describe() 的统计描述；散点图等
3. 通过箱线图（或3-Sigma）分析删除异常值
4. BOX-COX转换
5. 截尾
6. 利用模型进行离群点检测：聚类、K近邻、One Class SVM、Isolation Forest
   1. 缺省值
7. 不处理（一是缺省较少，二是训练模型为boost）
8. 删除（缺失数据过多）
9. 插值补全（均值，中位数，众数，建模预测，多重插补，矩阵补全等）
10. 分箱（缺失值一个箱）
11. 分桶
12. 等频分桶
13. 等距分桶
14. Best-KS分桶
15. 卡方分桶
16. 特征标准化/归一化
17. 标准化（转换为标准正态分布）
18. 归一化（所有值在[0，1]之间）
19. 针对幂律分布，可以采用公式：log
20. 特征筛选
21. 过滤式：先对数据进行特征选择，然后在训练学习器，常见的方法有 Relief/方差选择发/相关系数法/卡方检验法/互信息法
22. 直接把最终将要使用的学习器的性能作为特征子集的评价准则，常见方法有 LVM（Las Vegas Wrapper）
23. 嵌入式（embedding）：结合过滤式和包裹式，学习器训练过程中自动进行了特征选择，常见的有 lasso 回归
24. 特征构造
25. 构造统计量特征，报告计数/求和/比例/标准差等
26. 时间特征，包括相对时间和绝对时间，节假日，双休日等
27. 地理信息，包括分箱，分布编码等方法
28. 非线性变换，包括log/平方/开根等
29. 特征组合
30. 降维

PCA(主成分分析法)：符合高斯分布的样本点比较有效，过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性。

LDA(线性判别分析)：一种有监督的（supervised）线性降维算法。与PCA保持数据信息不同，LDA是为了使得降维后的数据点尽可能地容易被区分！

ICA(独立成分分析法)；对于高斯分布的样本点无效，对于其他分布的有效。

参考：[Datawhale 零基础入门数据挖掘-Task3 特征工程](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.9.1cd81b43sczqra&postId=95501)

[非常有价值的分享：零基础数据挖掘入门系列(四) - 特征工程](https://blog.csdn.net/wuzhongqiang/article/details/105012656)

另外还有很多大佬的提交很好，[这里插个眼](https://jinshuju.net/f/2eu4Gw/r/r1QTkX/share_entries?page=2)