牛熊市试验报告 III：特征工程

邸明轩*

08, 16, 2016

1 Explain

数据和特征决定了机器学习的上限，而统计学习等机器学习模型和算法只是逼近这个上限而已。那特征工程到底是什么呢？顾名思义，其本质是一项工程活动，目的是最大限度地从原始数据中提取特征以供算法和模型使用。

2 Result

模型：Logistic Regression分类器 + L1范数 + L2范数
目前使用了月度的指标，可以理解为每月最后一个交易日的指标，所以也是稀疏的（每月一次）的日指标。

实验一：标准化
标准化需要计算特征的均值和标准差，公式表达为：

$$x' = \frac{x - Min}{Max - Min} \tag{1}$$

F1准确率: 0.0.6278

实验二：归一化
对特征做归一化处理，利用两个最值进行缩放，公式表达为：

$$x' = \frac{x - \bar{X}}{S} \tag{2}$$

F1准确率: 0.6371

实验n：数据变换

常见的数据变换有基于多项式的、基于指数函数的、基于对数函数的. 2维特征的2阶多项式特征如下：$[1, a, b, a^2, ab, b^2]$.

*email: mingxuandi@163.com

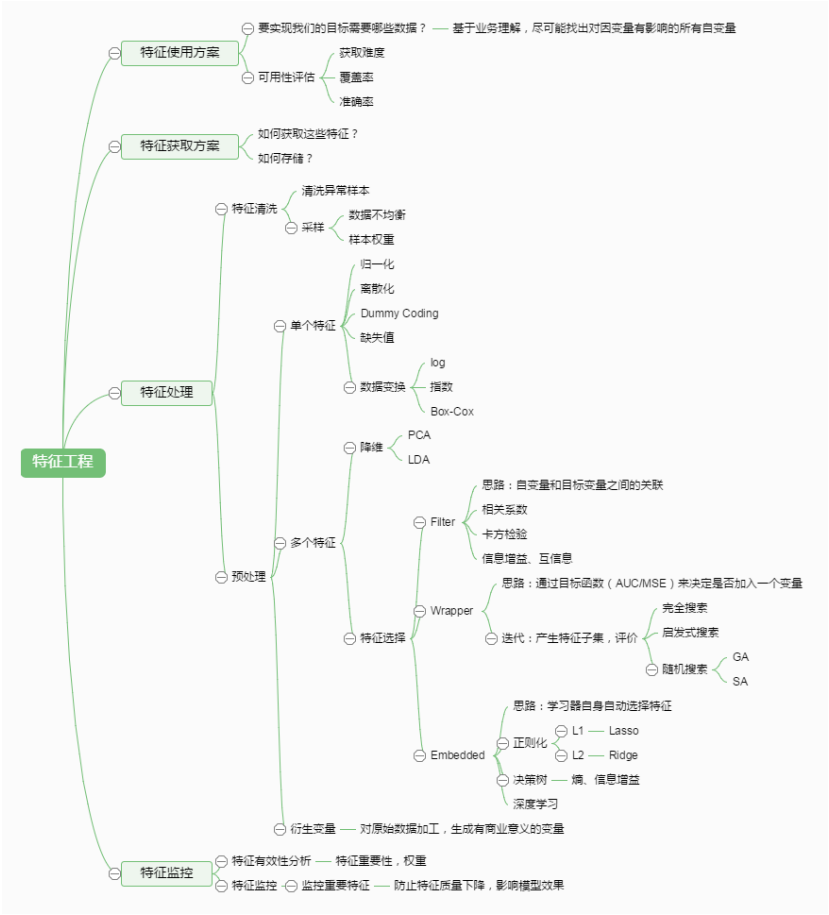

Figure 1: 特征工程的流程图，不感兴趣可以略过

实验四:
特征：实验二特征 + 前一季度归母净利 + 前一季度归母权益 + 前一季度ROE，共九个特征。(当季无法获取当季*roe*等指标，所以使用了前一季度指标)
F1准确率: 0.6312 (0.1164)

实验五:
特征：实验四特征 + 票据转贴利率 + 人民币名义有效汇率
F1准确率: 0.6366 (0.1073)