极客大学

极客大学机器学习训练营
常见机器学习模型

王然

众微科技 AI Lab 负责人

二〇二一年二月二十日

大纲

大纲

核心内容

极客大学

- 构造变量的三种方法；
  - 基于业务理解；
  - 基于常见的构建模式；
  - 根据 EDA 和 Bad Case 分析。
- 选择变量的方法：
  - 算法内置的选择法；
  - 单变量相关的选择方法；
  - Permutation Loss 为基础的选择方法。

在我们讨论构建变量之前…

- 变量的构建和选择与我们采用的模型有很大关系；
- 以树为基础的模型和包含线性组合的形式有很大不同。
- 思考题：假设我们将某离散变量按照其出现频率进行排列（最常出现的给 0，次常出现的给 1 等等）。这样的编码形式是否适合以树为基础的模型？是否适合以线性数学表达形式为基础的模型？

为什么特征工程非常重要

- 如果树模型或者神经网络模型可以认为是自动的特征提取器，为什么我们还需要手动构建特征呢？
- 可能原因：
  - 模型未必能够把正确的关系找到；
  - 减少模型的估计复杂度。

数据质量的一些说明

极客大学

- 在比赛中，一般来说数据集已经是清理好的；
- 但是在实际应用中，数据质量常常是非常糟糕的；
- 发现数据质量问题常常需要结合业务理解、EDA 和多源头检查进行；
- 处理数据质量问题往往只能依靠数据源头配合；
- 幸运的是：在一些情况下，即使数据质量很差，预测性建模仍然可以找到一些有效的模型；
- 注意部署的问题：防止新输入模型的变量出现奇怪的异常值。

构造变量和选择变量

极客大学

- 不论构造变量还是选择变量，都要通过**公平**的比较才可以进行；
- 通常来说，构造变量是比较容易的，但是如何构造出有效的变量是非常困难的；
- 在比赛中，构造变量通常是一个个进行的（除非时间有限制）；但是在实际工作中，构造变量常常是成块成块进行的；
- 最常见的模式是，尝试对一系列具有类似业务解释的变量进行构建，然后对于整体有效果的变量进行集中构建；在变量数量达到一定数量时候，进行一定的变量选择；
- 很遗憾的是，目前还无法总结出一个共性的变量构建流程。

大纲

核心原则

- 即使是之前建模中总结出来的有效的变量，在换了一个场景之后，仍然可能是无效的；
- 所以预先去通过“业务理解”找到黄金变量是不大可能的；
- 业务理解更多的是通过多种角度出发，构建可能的变量；

大纲

大纲

one-hot

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

one-hot 注意事项

- 有一些 one-hot 的变种，这些对于预测性建模影响不大；
- 对于树模型，一般不需要去掉其中一列；对于线性模型，一般来说需要去掉其中一列；
- 有时当类别极多时候，可以合并一些出现比较少的类别。

Target Encoding

极客大学

- 核心思想：通过预测值的平均值替代该变量；
- 核心问题：target leakage；类别内的异常值；
- 解决方法：leave-one-out encoder；将类别均值和整体均值进行加权估计；catboost encoder 等等
- 对于高维稀疏类别常常非常有效。

Count Encoding 和 Freq Encoding

极客大学

- 用该类别出现次数（频率）替换该变量;
- 合并训练集和测试集之后进行计算（不要单独计算）。

关于 Ordinal Encoding

- 不同实现对于 Ordinal Encoding 有不同的说法；但是一般来说，Ordinal Encoding 仅仅把类别映射成不同的整数而已，并且谁映射成对应整数是随机的；
- 从这个角度来说，单独的 Ordinal Encoding 不宜直接应用在建模当中。

大纲

对于树模型和线性模型的一些说明

极客大学

- 对于 XgBoost 和 LightGBM，由于选择树的分割点的时候，模型只关心顺序，所以理论上来说，保序或者倒序的变换不会对模型拟合结果产生影响；
- 对于线性模型，情况比较复杂：
  - 对于线性回归和逻辑回归而言，线性变换理论不会改变其模型效果，但是从数值解法的角度来说，不需要变量中出现过大或者过小的现象；
  - 如果带有惩罚项，则要保证量纲的一致；
  - 对于 SVM，需要进行标准化。
- KNN 显然需要进行标准化（除非采用不受标准化影响的距离）。

一些常见的变换

- 多项式（n 次方）；
- 指数和对数；
- 倒数（注意 0 和接近 0 的值）。

异常值的发现和处理

- 这里所说异常值，往往指特别大和特别小的值；
- 在一些情况下，这些值的出现可能是合理的，但是他对于我们构建模型却没有什么帮助；
- 问题往往在于无法对于异常值做非常具体的确定；
- 几种处理异常值的方法：
  - 将该观测去除；
  - 截断（尝试采用不同的截断方法）；
  - 赋值为缺失值。

缺失值处理

- 对于树模型，模型本身就可以处理缺失值；
- 对于线性模型，如果不希望将该观测丢弃，则需要采取填充的方法；
- 简单的填充方法（如中位数填充）往往已经够了；
- 一些复杂的填充方法如 MICE 可以在 R 当中实现；
- 对于非常重要的含有缺失值的变量，可以采用 lightgbm 进行预测。

正态化（?）

- 一些人称将变量变化为正态变量是有必要的；一些人认为这么做是非常愚蠢的；实际应用中，我没有发现任何的理论支持任何一种声称；
- 思考题：如何将任何一种观测都变成正态分布；
- 其他方法：Box-Cox 变换和 Johnson 变换。

离散化

极客大学

- 一定会丢失信息；
- 常用方法：
  - 均分；
  - 分位数；
  - 基于树模型；
  - 基于聚类分析（效果往往不好）。

大纲

Group-by

- 可以看作是两个变量之间的交叉效应；
- groupby 的操作可以是任何的 summary statistics（如 mean，median，max，min，range，moment 等等）

大纲

- 一般来说，这种处理都会有业务逻辑支撑；
- 最常见处理：乘和除。

大纲

唯一方法：直接构建交叉效应

大纲

一些其他方法

- row summary statistics;
- 将其他模型的结果输入到其他模型中;
- isolation forest;
- ...

大纲

大纲