特征工程（上）—特征选择

原创 stephenDC 大数据与人工智能 2019-07-17

作者丨stephenDC

这是作者的第13篇文章

机器学习问题，始于构建特征。

特征质量的好坏，直接影响到最终的模型结果。

构建特征是一个很大的工程，总体来讲包括“**特征选择**”、“**特征表达**”和“**特征评估**”3个部分。我们也按这3个部分，并结合自己的具体实践，用3篇文章来和大家聊一下特征工程的相关问题。

本篇文章，我们讨论一下特征选择。**特征选择指的是，在全部的特征中，挑选出对最终的机器学习任务有用的特征。**

整体来讲，从特征选择的过程中有没有模型的参与，可以将特征选择的方法分为，基于统计量的选择和基于模型的选择。

（在本文的讨论中，默认所有的特征取值已经去除量纲的影响，或者说已经做过归一化处理。）

基于统计量的特征选择

如果把每个特征看做一个随机变量，在不同的样本点处该随机变量可能会取到不同的值。可以用统计的方法，基于样本集的统计结果，对特征做出选择。

选择的标准主要有两个，一是特征本身取值的分散程度；二是该特征与要预测的结果之间的相关程度。

常用的几个统计量和方法包括，方差、相关系数、假设检验和互信息。下面依次说明。

方差

方差衡量的是一个随机变量取值的分散程度。如果一个随机变量的方差非常小，那这个变量作为输入，是很难对输出有什么影响的。在进行特征选择时，可以丢弃那些方差特别小的特征。

> **例子：**
>
> 如果你手上有5个offer，年收入水平分别是100万、100.01万、100.02万、100.03万和100.04万，我想你最终会选择哪个offer，年收入这个因素基本对你没太大影响吧。

相关系数

相关系数取值在-1到1之间，表征的是两个随机变量之间的线性相关关系。相关系数为0，表明两个变量之间线性无关；相关系数大于0，说明两个变量之间是正相关；相关系数小于0，代表两个变量之间负相关。

特征与输出的相关系数的绝对值越大，说明对输出的影响越大，应该优先选择。

例子：

1. 收入和学历有关系吗？高学历完全不能保证高收入。但从统计总体来看，学历较高，收入也会相对较高。如果要你对一个陌生人的收入做预测，那么学历肯定是要关注的因素之一。

2. 健康状况和吸烟多少有关系吗？我想，绝大多数人会认同，吸烟对健康是负相关，即有害的。

3. 个人成功和家庭背景有关系吗？“当然没有”，前总理的女儿曾说过，“能力之外的资本等于0”！

假设检验

假设检验是一种统计推断方法，简单来说就是先做一个假设，然后再基于某个量来判断该假设是否成立。比如，可以假设某个特征和输出是有显著相关性的，如果假设成立，即选择该特征；反之，丢弃该特征。

例子：

淑女品茶是一个有关假设检验的著名例子，这里换一下描述。

如果你有一个同事，宣称他对咖啡非常有研究，可以喝出来是先加的奶还是先加的糖。

你当然不信，所以你的假设是他没有这种判断能力。

检验方式是，给他10杯咖啡，不告诉他制作过程，让他通过喝来判断。设他判断正确的杯数为N，如果N超过了9，你可能就要拒绝当初的假设了，他可能真的有这个能力。

——来自维基百科

互信息

互信息，也叫信息增益，用过决策树模型的同学，对这个应该都不陌生。

简单来说，如果一个系统的信息熵为A，在某一个特征的已知的情况下，系统的信息熵变成B，则信息增益为A-B。互信息越大，证明这个信息对系统的分类越有帮助，相应的特征应优先选择。（P.S. 决策树用于回归问题时，互信息最大的标准变成了平方误差损失最小）

咦？不是说基于统计量的方法吗，怎么这里用到树模型了？

决策树模型分为树的生成和树的剪枝两个阶段，在树的生成阶段采用的是贪心策略，可以看做是基于统计量的。而“模型学习”的过程，更多的是树的剪枝。

当然，如果把这种方法看做是基于模型的特征选择，也完全没有问题。

基于模型的特征选择

基于模型的特征选择，可以直接根据模型参数来选择，也可用子集选择的思路选出特征的最优组合。

模型参数

对具有线性结构的模型，如线性模型（如Linear Regression）和对数线性模型（Logistic Regression，最大熵、线性链条件随机场等）等，都可以直接根据权重参数的大小来衡量对应特征的重要程度。

因为模型的线性结构，某个维度上的特征如果对应的参数绝对值大，这个维度的特征就相对重要；反之，参数绝对值小，则特征相对不重要。

对基于树结构的模型，如决策树、梯度提升树、随机森林和XGBoost等，每一颗树的生成过程，都对应了一个特征选择的过程。如上面关于信息增益一段的描述，可以对模型中涉及的树求平均，来表示特征的重要程度。与其他模型比，树模型的方差较大，因此选出来的特征也相对更不稳定。

因此，用树模型选择特征时，建议综合多次的模型训练结果。

如果我们想要得到稀疏特征或者说是对特征进行降维，可以在模型上主动使用正则化技术。使用L1正则，调整正则项的权重，基本可以得到任意维度的稀疏特征。

子集选择

基于模型，我们也可以用子集选择的思路来选取特征。假设特征的维度为N，要从中选出n个（n < N）特征，目标是让模型在选出的特征上效果最好。显然，n可以取不超过N的任意整数值，这就带来了组合爆炸的问题，总共要考虑的情况多到无法计算。解决组合爆炸问题，最常用的思路就是贪心策略（比如，决策树的生成过程中要选择切分特征和切分点，也是组合爆炸问题），常见的有前向搜索和反向搜索两种思路。

如果我们先从N个特征中选出一个最好的特征，然后让其余的N-1个特征分别与第一次选出的特征进行组合，从N-1个二元特征组合中选出最优组合。之后，再次在上次的基础上，添加一个新的特征，考虑3个特征的组合。这种思路有很多种叫法，可以被称为“递归式特征添加”、“前向搜索”或“自下向上的搜索”等。

反之，如果我们的目标是每次从已有特征中去掉一个特征，并从这些组合中选出最优组合。可以称为“递归式特征消除”、“反向搜索”或“自上向下的搜索”等。

显然，子集选择是需要很大的计算量的，因为每种特征组合都要用模型训练一遍。因此，这种方法基本不适合大规模使用，可以用于最后的精挑细选。

小结

本文从基于统计量和基于模型两个角度，笼统地介绍了特征选择的各种方法和思路。

基于统计量的特征选择，因为没有模型的参与，计算起来相对简单，可以作为一个特征预处理的过程。

基于模型的方法，要对模型参数进行学习，因此计算量相对较大；但也更接近于最终目标，即训练出一个泛化能力很好的模型。因此，这两种方法可以结合起来使用。

最后，提出一个问题，供大家一起交流。

如果用线性回归这样的简单模型来选择特征，最后用诸如FM或者GBDT这样的复杂模型来训练，会有什么问题吗?

-end-