特征选择方法

前言

特征工程是数据分析中最耗时间和精力的一部分工作，它不像算法和模型那样是确定的步骤，更多是工程上的经验和权衡。因此没有统一的方法。**这里只是对一些常用的方法做一个总结。**

本文关注于特征选择部分。

特征的来源

在做数据分析的时候，特征的来源一般有两块，一块是业务已经整理好各种特征数据，我们需要去找出适合我们问题需要的特征；另一块是我们从业务特征中自己去寻找高级数据特征。我们就针对这两部分来分别讨论。

选择合适的特征

我们首先看当业务已经整理好各种特征数据时，我们如何去找出适合我们问题需要的特征，此时特征数可能成百上千，哪些才是我们需要的呢？

第一步是找到该领域懂业务的专家，让他们给一些建议。比如我们需要解决一个药品疗效的分类问题，那么先找到领域专家，向他们咨询哪些因素（特征）会对该药品的疗效产生影响，较大影响的和较小影响的都要。这些特征就是我们的特征的第一候选集。

这个特征集合有时候也可能很大，在尝试降维之前，我们有必要用特征工程的方法去选择出较重要的特征结合，这些方法不会用到领域知识，而仅仅是统计学的方法。

最简单的方法就是方差筛选。方差越大的特征，那么我们可以认为它是比较有用的。如果方差较小，比如小于1，那么这个特征可能对我们的算法作用没有那么大。最极端的，如果某个特征方差为0，即所有的样本该特征的取值都是一样的，那么它对我们的模型训练没有任何作用，可以直接舍弃。在实际应用中，我们会指定一个方差的阈值，当方差小于这个阈值的特征会被我们筛掉。sklearn中的VarianceThreshold类可以很方便的完成这个工作。

特征选择方法有很多，一般分为三类：第一类过滤法比较简单，它按照特征的发散性或者相关性指标对各个特征进行评分，设定评分阈值或者待选择阈值的个数，选择合适特征。上面我们提到的方差筛选就是过滤法的一种。**第二类是包装法，根据目标函数，通常是预测效果评分，每次选择部分特征，或者排除部分特征。第三类嵌入法则稍微复杂一点，它先使用某些机器学习的算法和模型进行训练，**

得到各个特征的权值系数，根据权值系数从大到小来选择特征。类似于过滤法，但是它是通过机器学习训练来确定特征的优劣，而不是直接从特征的一些统计学指标来确定特征的优劣。

下面我们分别来看看3类方法。

过滤法选择特征

上面我们已经讲到了使用特征方差来过滤选择特征的过程。**除了特征的方差这第一种方法，还有其他一些统计学指标可以使用。**

第二个可以使用的是相关系数。这个主要用于输出连续值的监督学习算法中。我们分别计算所有训练集中各个特征与输出值之间的相关系数，设定一个阈值，选择相关系数较大的部分特征。

第三个可以使用的是假设检验，比如卡方检验。卡方检验可以检验某个特征分布和输出值分布之间的相关性。个人觉得它比比粗暴的方差法好用。如果大家对卡方检验不熟悉，可以参看这篇卡方检验原理及应用，这里就不展开了。在sklearn中，可以使用chi2这个类来做卡方检验得到所有特征的卡方值与显著性水平P临界值，我们可以给定卡方值阈值， 选择卡方值较大的部分特征。

除了卡方检验，我们还可以使用F检验和t检验，它们都是使用假设检验的方法，只是使用的统计分布不是卡方分布，而是F分布和t分布而已。在sklearn中，有F检验的函数f_classif和f_regression，分别在分类和回归特征选择时使用。

第四个是互信息，即从信息熵的角度分析各个特征和输出值之间的关系评分。在决策树算法中我们讲到过互信息（信息增益）。互信息值越大，说明该特征和输出值之间的相关性越大，越需要保留。在sklearn中，可以使用mutual_info_classif(分类)和mutual_info_regression(回归)来计算各个输入特征和输出值之间的互信息。

以上就是过滤法的主要方法，个人经验是，在没有什么思路的时候，可以优先使用卡方检验和互信息来做特征选择。

包装法选择特征

包装法的解决思路没有过滤法这么直接，它会选择一个目标函数来一步步的筛选特征。

最常用的包装法是递归消除特征法(recursive feature elimination,以下简称RFE)。递归消除特征法使用一个机器学习模型来进行多轮训练，每轮训练后，消除若干权值系数的对应的特征，再基于新的特征集进行下一轮训练。在sklearn中，可以使用RFE函数来选择特征。

我们下面以经典的SVM-RFE算法来讨论这个特征选择的思路。这个算法以支持向量机来做RFE的机器学习模型选择特征。它在第一轮训练的时候，会选择所有的特征来训练，得到了分类的超平面 $wx + b = 0$ 后，如果有n个特征，那么RFE-SVM会选择出w中分量的平方值ω_i^2最小的那个序号i对应的特征，将其排除，在第二类的时候，特征数就剩下n-1个了，我们继续用这n-1个特征和输出值来训练SVM，同样的，去掉最小ω_i^2的那个序号i对应的特征。以此类推，直到剩下的特征数满足我们的需求为止。

嵌入法选择特征

嵌入法也是用机器学习的方法来选择特征，但是它和RFE的区别是它不是通过不停的筛掉特征来进行训练，而是使用的都是特征全集。在sklearn中，使用SelectFromModel函数来选择特征。

最常用的是使用L1正则化和L2正则化来选择特征。在之前讲到的用scikit-learn和pandas学习Ridge回归第6节中，我们讲到正则化惩罚项越大，那么模型的系数就会越小。当正则化惩罚项大到一定的程度的时候，部分特征系数会变成0，当正则化惩罚项继续增大到一定程度时，所有的特征系数都会趋于0.但是我们会发现一部分特征系数会更容易先变成0，这部分系数就是可以筛掉的。也就是说，我们选择

特征系数较大的特征。常用的L1正则化和L2正则化来选择特征的基学习器是逻辑回归。

此外也可以使用决策树或者GBDT。那么是不是所有的机器学习方法都可以作为嵌入法的基学习器呢？也不是，一般来说，可以得到特征系数coef或者可以得到特征重要度(feature importances)的算法才可以做为嵌入法的基学习器。

寻找高级特征

在我们拿到已有的特征后，我们还可以根据需要寻找到更多的高级特征。比如有车的路程特征和时间间隔特征，我们就可以得到车的平均速度这个二级特征。根据车的速度特征，我们就可以得到车的加速度这个三级特征，根据车的加速度特征，我们就可以得到车的加加速度这个四级特征。。。也就是说，高级特征可以一直寻找下去。

在Kaggle之类的算法竞赛中，高分团队主要使用的方法除了集成学习算法，剩下的主要就是在高级特征上面做文章。所以寻找高级特征是模型优化的必要步骤之一。当然，在第一次建立模型的时候，我们可以先不寻找高级特征，得到以后基准模型后，再寻找高级特征进行优化。

寻找高级特征最常用的方法有：

1.若干项特征加和： 我们假设你希望根据每日销售额得到一周销售额的特征。你可以将最近的7天的销售额相加得到。

2.若干项特征之差： 假设你已经拥有每周销售额以及每月销售额两项特征，可以求一周前一月内的销售额。

3.若干项特征乘积： 假设你有商品价格和商品销量的特征，那么就可以得到销售额的特征。

4.若干项特征除商： 假设你有每个用户的销售额和购买的商品件数，那么就是得到该用户平均每件商品的销售额。

当然，寻找高级特征的方法远不止于此，它需要你根据你的业务和模型需要而得，而不是随便的两两组合形成高级特征，这样容易导致特征爆炸，反而没有办法得到较好的模型。个人经验是，聚类的时候高级特征尽量少一点，分类回归的时候高级特征适度的多一点。

特征选择小结

特征选择是特征工程的第一步，它关系到我们机器学习算法的上限。因此原则是尽量不错过一个可能有用的特征，但是也不滥用太多的特征。