机器学习常见知识汇总之特征工程

田晓彬
xiaobin9652@163.com

特征选择

从给定的特征集合中选择出相关特征子集的过程叫做特征选择。特征选择是一个重要的数据预处理过程。进行特征选择有两个很重要的原因：第一个原因是很多数据的维度非常高，进行特征选择（或者进行降维，后续会介绍）可以减少数据的维度。第二个原因是特征选择可以去除不相关的特征，降低学习任务的难度。

如果希望从原始的特征集合中找到包含最多信息的特征子集，最粗暴的方法是遍历所有可能的子集，但是这在现实中基本是不可能实现的。于是我们退而求其次，初始化一个候选子集，然后评价它的好坏并基于评价结果产生下一个候选子集，并迭代进行这个步骤直到无法找到更好的候选子集为止。

这个过程有两个关键环节：如何根据评价结果找到下一个候选子集；如何评价候选特征子集的好坏。

第一个环节是**子集搜索**问题，是一个搜索特征子集的过程。常用的算法分为三大类：完全搜索，启发式搜索和随机搜索。通常使用启发式搜索来确定特征子集。下面详细介绍启发式搜索中的前向搜索。首先从特征子集为空集开始，每次选择一个不在特征子集中的特征加入特征子集中。假设特征总数为 n，当前特征子集中特征个数为 k，每次选出一个特征，构成 $k+1$ 个特征的候选特征子集，求出候选特征子集的效果。找到比原来特征子集效果好的最优候选特征子集作为当前的特征子集。重复这个过程，直到没有候选特征子集比当前特征子集效果好时停止。

可以发现前向搜索是每次选出一个特征加入特征子集中，若从完整的特征子集中每次去掉一个特征，这种逐步减少特征的策略叫做后向搜索。还可以将前向与后向搜索结合起来，每一轮同时增加相关特征，减少无关特征，这种方法被称为前后向搜索。这三种方法全部采用了贪心策略，虽然会导致找不到最优的解，但是不进行穷举搜索时，这种情况是不可避免的。

第二个环节是**子集评价**。给定特征子集 A，根据其取值可以将所有样本集 D 划分成 V 个子集 $\{D^1, D^2, ...D^V\}$，其中每个子集中的样本在 A 上的取值相同。可以看出，特征子集 A 确定了数据集的一个划分，而样本标记 Y 可以确定数据集的真实划分。通过评估两个划分的差异，就可以对 A 进行评价。与 Y 对应划分的差异越小，说明 A 的效果越好。任何可以判断两个划分差异的机制都可以用于子集评价，如信息熵。

将子集搜索和子集评价相结合，即可以得到特征选择方法。决策树即可以看做是一种将前向搜索和信息熵结合的特征选择方法，决策树的划分属性的集合就是选择出的特征子集。

通常来说，特征选择算法可以大致分为三大类：过滤式方法、包裹式方法和嵌入式方法。

1.过滤式方法

过滤式方法首先对数据集进行特征选择，然后在训练模型，特征选择过程和后续模型学习无关。

2.包裹式方法

包裹式方法直接把最终将要使用的模型的性能作为特征子集的评价准则。可以看出，包裹式方法直接对给定的模型进行优化，因此从最后模型的性能来说，包裹式方法要比过滤式方法更好。但是因为在特征选择的过程中需要多次训练模型，所以包裹式方法的计算开销非常大。

3.嵌入式方法

嵌入式方法是将特征选择过程和模型的训练过程融为一体，两者在同一个优化过程中完成。通常引入正则化项来对模型同时进行训练和特征选择。通常使用的正则化方法有 L_1 正则化和 L_2 正则化。引入 L_2 正则化的优化目标称为**岭回归**；引入 L_1 正则化的优化目标称为**LASSO**。L_1 和 L_2 正则化都可以帮助模型降低过拟合风险，L_1 正则化还有一个好处，它可以得到更为稀疏的解，即更多的特征在模型训练中被忽视掉。

数据归一化和中心化

在机器学习中，不同的特征往往具有不同的取值范围，这样会影响到模型的结果。为了消除不同特征之间取值范围的影响，将原始数据进行归一化或者中心化。

1.中心化

中心化又叫零均值化，将数据减去数据的均值。可以将中心化看作是平移的过程，平移过后，所有数据的中心是 $(0,0)$。

2.归一化

归一化又叫标准化，对数据进行归一化可以将所有的特征都统一到一个大致相同的数值区间。对数据进行归一化就是将数据减去其均值，再除以方差，即将原始数据映射到一个均值为0，方差为1的分布上。

需要注意一点的是，归一化和中心化对于决策树模型并不适用。以C4.5为例，信息增益比与数据是否经过归一化是无关的，因为归一化不会改变样本在某一个特征上的信息增益。

数据不足或类别不平衡

一个模型所能提供的信息一般来自两个方面，一是训练数据中蕴含的信息，二是在模型的形成过程中人们提供的先验信息。当学习任务中的训练数据不足时，为了保证模型的有效性，就需要更多的先验知识。先验知识可以直接作用在模型上，例如让模型采用特定的内在结构、条件假设或者添加其他一些约束条件；先验知识还可以直接作用在数据集上，即根据特定的先验假设去调整、变换或扩展数据集，使数据集展现出更多更有用的信息。

根据上面的表述，我们可以总结出两类处理方法：一类是基于模型的方法，主要是采用降低模型过拟合风险的措施，包括简化模型（将非线性模型简化成线性模型）、添加约束项以减少假设空间（L1/L2正则化）、集成学习等；另一类是基于数据的方法，主要是对数据进行扩充，即根据先验知识在保证特定信息的前提下，对原始数据进行适当的变换以达到扩充数据集的效果。例如SMOTE、生成对抗网络。

针对于图像数据集，我们可以针对图像进行一下变换来扩充数据集：

1. 一定程度的随机旋转、平移、缩放、裁剪、填充、左右翻转等，这些变化对应着同一个目标在不同角度的观察结果；
2. 对图像中的像素添加噪声扰动，比如高斯白噪声；
3. 颜色变换；
4. 改变图像的亮度、清晰度、对比度、锐度等。

文本表示模型