特征工程

概念

特征是数据中抽取出来的对结果预测有用的信息，可以是文本或者数据。特征工程是使用专业背景知识和技巧处理数据，使得特征能在机器学习算法上发挥更好的作用的过程。过程包含了特征提取、特征构建、特征选择等模块。 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  
　　特征工程的目的是筛选出更好的特征，获取更好的训练数据。因为好的特征具有更强的灵活性，可以用简单的模型做训练，更可以得到优秀的结果。“工欲善其事，必先利其器”，特征工程可以理解为利其器的过程。互联网公司里大部分复杂的模型都是极少数的数据科学家在做，大多数工程师们做的事情基本是在数据仓库里搬砖，不断地数据清洗，再一个是分析业务不断地找特征。

背景

特征工程重要性：

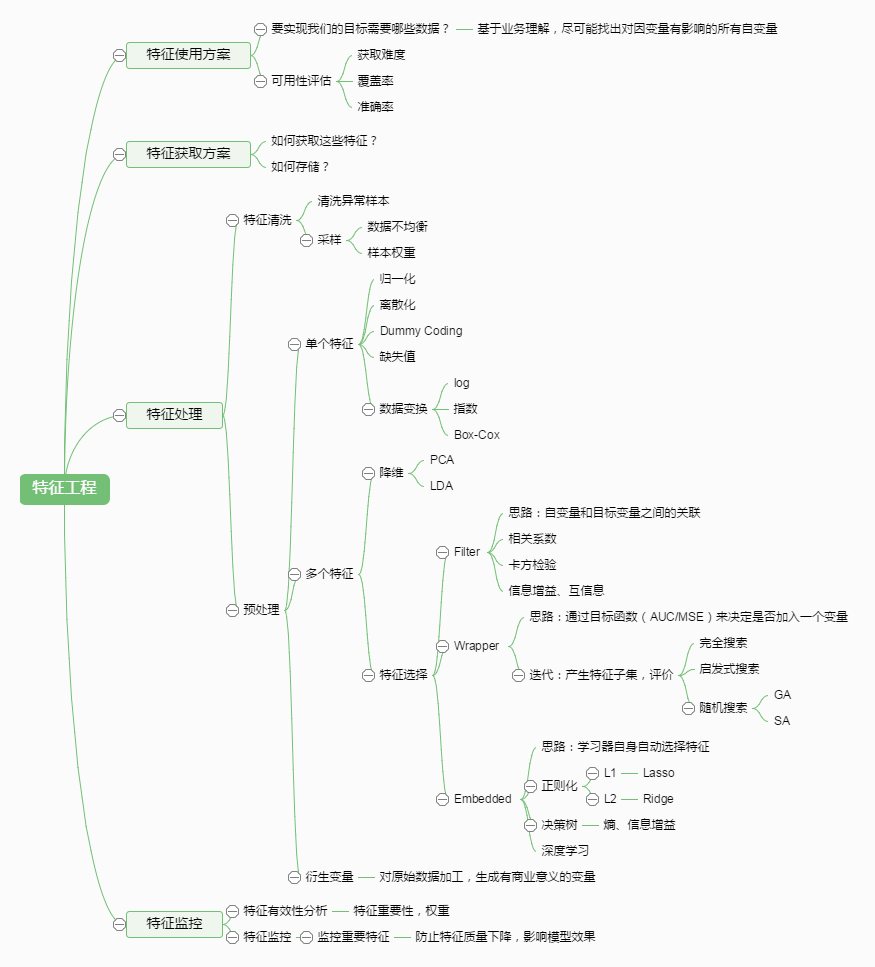
特征越好，灵活性越强；

特征越好，模型越简单；

特征越好，性能越出色；

好特征即使使用一般的模型，也能得到很好的效果！好特征的灵活性在于它允许你可以选择不复杂的模型，同时，运行速度也更快，也更容易理解和维护。  
好的特征，即使参数不是最优解，模型性能也能表现很好，因此，不需要太多时间去寻找最优参数，大大的降低了模型的复杂度，使模型趋向简单。  
模型的性能包括模型的效果，执行的效率及模型的可解释性。特征工程的最终目的就是提升模型的性能。

特征工程主要包括以下

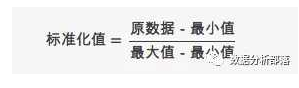


特征工程的步骤：

数据处理

数据统一化（量纲不一）

1）:0-1标准化 是对原始数据进行线性变换，将特征值映射成区间为［0，1］的标准值中：



Z标准化

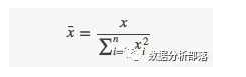
基于特征值的均值（mean）和标准差（standard deviation）进行数据的标准化。它的计算公式为：



标准化后的变量值围绕0上下波动，大于0说明高于平均水平，小于0说明低于平均水平。

Normalizer归一化

将每个样本缩放到单位范数（每个样本的范数为1），计算公式如下：



sklearn实现标准化 全名Scikit-Learn，是基于Python的机器学习模块

1.2 虚拟变量

虚拟变量：也叫哑变量和离散特征编码，可用来表示分类变量、非数据因素可能产生的影响

· 虚拟变量的两种数据类型：

离散特征的取值之间有大小的意义：例如：尺寸（L、XL、XXL）

离散特征的取值之间没有大小的意义：例如：颜色（Red、Blue、Green）

· 离散特征值有大小意义的虚拟变量处理

离散特征的取值之间有大小意义的处理函数，我们只需要把大小值以字典的方式，作为第一个参数传入即可；

(1) dict 映射的字典

pandas.Series.map(dict)

· 离散特征值没有大小意义的虚拟变量处理

离散特征的取值之间没有大小意义的处理方法，我们可以使用get\_dummies方法处理，它有6个常用的参数

(1) data 要处理的DataFrame

(2) prefix 列名的前缀，在多个列有相同的离散项时候使用

(3) prefix\_sep 前缀和离散值的分隔符，默认为下划线，默认即可

(4) dummy\_na 是否把NA值，作为一个离散值进行处理，默认不处理

(5) columns 要处理的列名，如果不指定该列，那么默认处理所有列

(6) drop\_first 是否从备选项中删第一个，建模的时候为避免共线性使用

pandas.getdummies(data,prefix=None,prefix\_sep=’‘,dummy\_na=False,columns=None,drop\_first=False)

1.3 数据处理 缺失值填充

缺失值产生原因

有些信息暂时无法获取；

有些信息被遗漏或者错误的处理了

缺失值处理方法

数据补齐

删除缺失行

不处理

特征选择

2.1特征选择 ( Feature Selection )也称特征子集选择( Feature Subset Selection , FSS ) ，或属性选择( Attribute Selection ) ，是指从全部特征中选取一个特征子集，使构造出来的模型更好。

2.2 为什么要特征选择

· 在机器学习的实际应用中，特征数量往往较多，其中可能存在不相关的特征，特征之间也可能存在相互依赖，容易导致如下的后果：

特征个数越多，分析特征、训练模型所需的时间就越长。

特征个数越多，容易引起“维度灾难”，模型也会越复杂，其推广能力会下降。

· 特征选择能剔除不相关(irrelevant)或亢余(redundant )的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一方面，选取出真正相关的特征简化了模型，使研究人员易于理解数据产生的过程。

2.3特征选择的基本原则

是否发散

是否相关

如果一个特征不发散，例如方差接近于0，也就是说样本在这个特征上基本没有差异，那我们就可以判断，这个特征对于样本的区别并没有什么用

第二个是特征与目标的相关性，与目标相关性高的特征应该优先选择

2.4特征选择的常用四种方法

2.4.1 方差选择法(过滤法)

用方差选择法，先要计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征。

2.4.2 相关系数法

先计算各个特征对目标值的相关系数，选择更加相关的特征

2.4.3 递归特征消除法

使用一个基模型来进行多轮训练，经过多轮训练后，保留指定的特征数。

2.4.4模型选择法

它是一种我们把建好的模型对象传入选择器，然后它会根据这个已经建好的模型，自动帮我吗选择最好的特征值。

维度压缩

当特征选择完成之后，就可以直接训练模型了，但是可能由于特征矩阵过大导致计算量大，训练时间长的问题；因此，降低特征矩阵维度，也是必不可少的，主成分分析就是最常用的降维方法，在减少数据集的维度的同时，保持对方差贡献最大的特征