基础概念

特征工程是通过对原始数据的处理和加工，将原始数据属性通过处理转换为数据特征的过程，属性是数据本身具有的维度，特征是数据中所呈现出来的某一种重要的特性，通常是通过属性的计算，组合或转换得到的。比如主成分分析就是将大量的数据属性转换为少数几个特征的过程。某种程度而言，好的数据以及特征往往是一个性能优秀模型的基础。

既然叫特征工程，自然涵盖了很多内容，而其中涉及到的比较重要的部分是特征的处理及选择。

特征处理包含：

- 数据清洗
- 数据规范化
- 特征衍生与提取

特征选择包含：

- 特征过滤
- wrapper method
- embedded method

数据清洗

数据清洗是指发现并纠正数据文件中可识别的错误以及通过处理得到建模过程需要数据的过程。

数据清洗包含：

- 缺失值处理
- 异常值检测与处理
- 调配样本比例和权重

缺失值处理

缺失值是指粗糙数据中由于缺少信息而造成的数据的聚类、分组、删失或截断。它指的是现有数据集中某个或某些属性的值是不完全的。

缺失值的处理目前主要有两种方法：删除缺失值和填充缺失值

1.删除缺失值

如果一个样本或变量中所包含的缺失值超过一定的比例，比如超过样本或变量的一半，此时这个样本或变量所含有的信息是有限的，如果我们强行对数据进行填充处理，可能会加入过大的人工信息，导致建模效果打折扣，这种情况下，我们一般选择从数据中剔除整个样本或变量，即删除缺失值。

2.缺失值填充

- **随机填充法**

从字面上理解就是找一个随机数，对缺失值进行填充，这种方法没有考虑任何的数据特性，填充后可能还是会出现异常值等情况，一般情况下不建议使用。

- **均值填充法**

寻找与缺失值变量相关性最大的那个变量把数据分成几个组，然后分别计算每个组的均值，然后把均值填入缺失的位置作为它的值，如果找不到相关性较好的变量，也可以统计变量已有数据的均值，然后把它填入缺失位置。这种方法会在一定程度上改变数据的分布。

- **最相似填充法**

在数据集中找到一个与它最相似的样本，然后用这个样本的值对缺失值进行填充。
与均值填充法有点类似，寻找与缺失值变量（比如x）相关性最大的那个变量（比如y），然后按照变量y的值进行排序，然后得到相应的x的排序，最后用缺失值所在位置的前一个值来代替缺失值。

- **回归填充法**

把缺失值变量作为一个目标变量y，把缺失值变量已有部分数据作为训练集，寻找与其高度相关的变量x建立回归方程，然后把缺失值变量y所在位置对应的x作为预测集，对缺失进行预测，用预测结果来代替缺失值。

- **k近邻填充法**

利用knn算法，选择缺失值的最近k个近邻点，然后根据缺失值所在的点离这几个点距离的远近进行加权平均来估计缺失值。

异常值检测与处理

异常值（outlier）是指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。异常值的产生一般由系统误差、人为误差或数据本身的变异引起的。

- **单变量异常值检测（格拉布斯法）**

首先，将变量按照其值从小到大进行顺序排列x1,x2.....xn

其次，计算平均值x拔和标准差S，

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

同时计算偏离值，即平均值与最大值之差和平均值与最小值之差，然后确定一个可疑值，一般是偏离平均值较大的那个。

计算统计量gi（残差与标准差的比值），i为可疑值的序列号。

$$g_i=\frac{|x_i-\bar{x}|}{s}$$

再者，将gi与格拉布斯表给出的临界值GP(n)比较，如果计算的Gi值大于表中的临界值GP(n)，则能判断该测量数据是异常值，可以剔除。这里临界值GP(n)与两个参数有关：检出水平α和测量次数n 。

检出水平α：如果要求严格，检出水平α可以定得小一些，例如定α＝0.01，那么置信概率P＝1－α＝0.99；如果要求不严格，α可以定得大一些，例如定α＝0.10，即P＝0.90；通常定α＝0.05，P＝0.95。

- **多变量异常值检测（基于距离计算）**

基于距离的多变量异常值检测类似与k近邻算法的思路，一般的思路是计算各样本点到中心点的距离，如果距离太大，则判断为异常值，这里距离的度量一般使用马氏距离(Mahalanobis Distance)。因为马氏距离不受量纲的影响，而且在多元条件下，马氏距离还考虑了变量之间的相关性，这使得它优于欧氏距离。

- **异常值处理**

单变量的情况下异常值可以考虑类似缺失值的删除法、均值填充法或回归填充法，而多变量的情况下，可以尝试用均值向量填充或者删除。

总的来说，缺失值和异常值的处理要根据实际的情况确定合适的方法，因为某些情况下异常值刚好能够反应一些现实问题。

调配样本比例和权重

当数据集中出现样本不均衡情况时，需要调配样本的比例以及权重，以便能够训练出性能更优的模型

http://www.cnblogs.com/wkslearner/p/8870673.html

数据规范化

在机器学习中，由于不同模型的需要，我们经常要多数据做不同的规范化处理，以便能够得到性能更优的模型。

在数据处理中，经常会接触到的数据规范化操作有：

- 数据无量纲化
- 连续变量离散化
- 离散变量处理

数据无量纲化

无量纲化使不同规格的数据转换到同一规格，在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。

数据无量纲化常用方法有：

- 标准化方法
- 极值化方法
- 均值化方法
- 标准差化方法

1.标准化方法

标准化方法是将变量的每个值与其平均值之差除以该变量的标准差，无量纲化后变量的平均值为0，标准差为1。使用该方法无量纲化后不同变量间的均值和标准差都相同，即同时消除了变量间变异程度上的差异。
标准化公式为：

$$x_i' = \frac{x_i - \bar{x}}{s}$$

2.极值化方法

极值化方法通常是通过变量取值的最大值和最小值将原始数据转换为特定范围内的数据，从而消除量纲和数量级的影响。这种方法十分依赖两个极端值。

通常情况下极值化方法有3种方式：

第一种方法，是将变量的值除以该变量的全距，标准化后每个变量的取值范围在[-1,1]。
公式为：

$$x_i' = \frac{x_i}{\max - \min} = \frac{x_i}{R}$$

第二种方法，是将变量值与最小值之差除以该变量的全距，标准化后取值范围在[0,1]。
公式为：

$$x_i' = \frac{x_i - \min}{\max - \min} = \frac{x_i - \min}{R}$$

第三种方法，是将变量值除以该变量的最大值，标准化后变量的最大取值为1。
公式为：

$$x_i' = \frac{x_i}{\max}$$

3.均值化方法

均值化方法是将变量值直接除以该变量的平均值，跟标准化方法不同的是，均值化方法能够保留变量间取值差异程度的信息。
均值化方法公式：

$$x_i' = \frac{x_i}{\overline{x}_i}$$

4.标准差化方法

标准差化方法是标准化方法的一种变形，标准差化方法是直接将变量值除以标准差，而不是减去均值后再除以标准差。标准差化方法无量纲化后变量的均值为原始变量均值与标准差的比值，而不是0。
公式为：

$$x_i' = \frac{x_i}{s}$$

连续变量离散化

在使用某些算法时，我们需要把连续变量转换为离散变量，在一些情况下离散变量能够简化模型计算同时能够提升模型的稳定性，比如逻辑回归经常使用离散后的变量进行训练，能够体现模型的训练速度以及提升模型的可解释性。

连**续变量离散化**大致有两类方法：

- 卡方检验方法
- 信息增益方法

1.卡方检验方法

通常情况下，将变量按照值大小进行排列，将每个值作为一个组，然后对每一对相邻的组计算卡方值，对其中最小的一对组合进行合并，接下来不断重复以上操作，直到满足我们设定的某一个条件，比如最小分组数5,即将连续变量分为5组。

卡方统计量是指数据的分布与所选择的预期或假设分布之间的差异的度量。它是由各项实际观测次数（fo ）与理论分布次数（fe ）之差的平方除以理论次数，然后再求和而得出的，其计算公式为：

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

卡方值包含两个信息：

- 实际值与理论值偏差的绝对大小。
- 差异程度与理论值的相对大小。

2.信息增益方法

信息增益方法是使用信息计算确定分割点的自上而下的分裂技术。

首先是把每个值看成分割点，将数据分成两个部分，在多种可能的分法中选择产生最小信息熵的分法。然后在分成的两个区间中，寻找最大熵区间，继续进行按前面的方法进行分割，直到满足条件为止，比如满足指定个数时结束过程。

数据的信息属性是与任务相关的，对于分类任务, 标签值y包含的信息量为:

$$info(y) = -lnp(y)$$

其中, p(y)为y出现的概率. p(y)越小, y包含的信息量越大. 这是符合直觉的.

熵定义为信息的期望值.
一个可以分为m类的数据集S, 它的信息熵为随机得到的一个label包含的信息量的期望值:

$$E(S) = -\sum_{i=1}^{m} p(y_i) lnp(y_i)$$

数据集的信息熵代表这个数据集的混乱程度. 熵越大, 越混乱.

$$gain = E(S) - \sum_{i=1}^{n} E(S_i)$$

若按照某种特定的方式, 例如按照某一属性的值对S进行划分, 得到n个子集。新的子集们都有自己的信息熵, 它们的熵的和与原S的熵的差值就是这个划分操作带来的信息熵增益.

离散变量处理

在某些情况下，比如回归建模时，我们通常需要将分类变量量化处理或离散变量哑变量化。

分类变量分为两种：有序分类变量和无序分类变量，在引入模型是我们通常需要对其进行量化处理，转化为离散变量，比如疾病的严重程度轻微，中度，重度，量化后用1，2，3来代替。但无序的分类变量，比如血型A，B，O型，如果我们也用1,2,3表示就不合理了，因为血型之间并不存在递进的关系。

此时我们需要对离散变量进行进一步的处理，即哑变量化。

哑变量（Dummy Variable），又称为虚拟变量、虚设变量或名义变量，通常取值为0或1，来反映某个变量的不同属性。对于有n个分类属性的自变量，通常需要选取1个分类作为参照，因此可以产生n-1个哑变量。

哑变量化后，特征就变成了稀疏的了。这有两个好处，一是解决了模型不好处理属性数据的问题，二是在一定程度上也起到了扩充特征的作用。

特征衍生与提取

在建模过程中，我们通常会遇到一些问题，现有特征的显著性不高或者特定算法的需要，我们需要从现有数据中构造一些特征，有时又可能因为特征过多，而需要降维处理，一般的方法是从众多特征中提取出特征的共性，然后进行建模。

特征衍生

特征衍生一般是对原有的特征进行转换，计算以及组合而产生的新的特征。

1. 单一变量的基础转换，比如通过对单一变量进行平方，开根号，log转换等。
2. 变量通过添加时间维度进行衍生，比如3个月交易数据，6个月交易数据等。
3. 多变量的运算，比如两个变量相加，相乘或变量间计算一个比率后得到新变量。

当然特征衍生的方式各种各样，具体还是要看业务场景的需要，然后做相应的处理。

特征提取

特征提取是从原始特征中找出最有效的特征，这种做法的目的是降低数据冗余，减少模型计算，发现更有意义的特征等。

特征提取分为：

- 线性特征提取
- 非线性特征提取

线性特征提取

线性特征提取一般方法有PCA-主成分分析，LDA-线性判别分析，ICA-独立成分分析等

1.PCA-主成分分析

主成分分析的原理是将一个高维向量x,通过一个特殊的特征向量矩阵U，投影到一个低维的向量空间中，表征为一个低维向量y，并且仅仅损失了一些次要信息。也就是说，通过低维表征的向量和特征向量矩阵，可以基本重构出所对应的原始高维向量。

PCA的算法步骤：

- 去平均值，对每一个特征减去各自的平均值
- 计算协方差矩阵
- 计算协方差矩阵的特征值及对应的特征向量
- 将特征向量按对应特征值从大到小进行排序，取靠前的k个特征向量
- 将数据转换到k个特征向量构建的新空间中，即为降维到k维后的数据

2.LDA-线性判别分析

LDA是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的。这点和PCA不同。PCA是不考虑样本类别输出的无监督降维技术。LDA是通过将数据在低纬度上进行投影，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据的类别中心之间的距离尽可能的大。

LDA的算法步骤：

- 计算类内散度矩阵
- 计算类间散度矩阵
- 计算类内散度矩阵的逆与类间散度矩阵的乘积
- 计算乘积结果的最大特征值及其对应的特征向量，得到投影矩阵
- 将数据集中的每一个样本特征转换为新样本
- 输出得到新数据集

特征选择

特征选择的一般过程是这样的，首先是从特征全集中产生出一个特征子集，筛选过程采用某种评价标准，把符合标准的特征筛选出来，同时对筛选出来的特征进行有效性验证。

产生特征子集一般是一个搜索的过程，搜索空间中的每个状态就是一个特征子集，搜索算法分为完全搜索，启发式搜索和随机搜索。
特征选择的过程可分为，特征过滤，wrapper及embedded。

特征过滤

特征过滤是选定一个指标来评估特征，根据指标值来对特征进行重要性排序，去掉达不到指标的特征，评价指标包含方差，相关性，信息增益等。

- 基于方差的特征过滤：计算每个特征的方差大小进行排序，然后按照特定的阈值或者特征个数进行筛选，方差的大小实际上表示的是变量所含有的信息量，方差较小可能的表现是变量的取值比较单一，对于我们区分目标变量的用处不大，因而可以选择剔除。
- 相关性特征过滤：计算各个特征对目标特征的相关系数以及相关系数的P值，选择显著性高的特征。
- 基于信息增益的特征过滤：计算包含每个特征带来信息增益的大小，并以此来判断每个特征对于我们分类器所提供信息量的大小，信息增益越大，说明该特征对于分类结果正确提供的帮助越大。

Wrapper Method

Wrapper方与特征过滤不同，它不单看特征和目标直接的关联性，而是从添加这个特征后模型最终的表现来评估特征的好坏。而在一个特征空间中，产生特征子集的过程可以看成是一个搜索问题。目前主要用的一个Wrapper方法是递归特征消除法。

递归特征消除的主要思想是不断使用从特征空间中抽取出来的特征子集构建模型，然后选出最好的的特征，把选出来的特征放到一遍，然后在剩余的特征上重复这个过程，直到所有特征都遍历了。这个过程中特征被消除的次序就是特征的排序。这是一种寻找最优特征子集的贪心算法。

Embedded Method

Embedded方法是在模型构建的同时选择最好的特征。最为常用的一个Embedded方法就是：正则化。正则化就是把额外的约束或者惩罚项加到已有模型的损失函数上，以防止过拟合并提高泛化能力。正则化分为L1正则化(Lasso)和L2正则化(Ridge回归)。

L1正则化是将所有系数的绝对值之和乘以一个系数作为惩罚项加到损失函数上，现在模型寻找最优解的过程中，需要考虑正则项的影响，即如何在正则项的约束下找到最小损失函数。同样的L2正则化也是将一个惩罚项加到损失函数上，不过惩罚项是参数的平方和。其他还有基于树的特征选择等。