特征工程：特征提取、特征选择、特征构建

一、特征选择：从给定特征集合中选出相关特征的过程

**选择相关特征，抛弃不相关特征**；特征选择回影响模型的预测能力

若保留尽可能多的特征，模型的性能能提升，但是模型和计算变得复杂

若剔除较多的特征，模型的性能降低；因此留下有用的特征，剔除无用特征

1. 特征选择的原理：
2. 维度灾难：

因为属性或者特征过多，剔除无关特征或者冗余特征可以减轻维度灾难问题，降低学习任务的难度，同时让模型变得简单，降低计算复杂度（必须确保不丢失重要特征）

* 不相关特征：与当前学习任务无关的特征
* 冗余特征：该特征包含的信息可以从其他特征中推演出来，一般不起作用

1. **子集评价**

**特征子集搜索方式：**

**前向搜索：**逐渐增加相关特征

**后向搜索：**从完整特征集每次尝试去掉一个无关特征

**双向搜索：**每一轮逐渐增加选定的相关特征（这些特征在后续迭代中确定不会被去除），同时减少无关特征

**子集评价方式：**

给定数据集D，假设所有属性均为离散型。对属性子集A，假定根据其取值将D分成了V个子集: D1,D2,…,Dv，可以计算属性子集 A 的**信息增益**：

* H( )表示熵；| |表示集合大小；
* 信息增益越大，表示特征集合A包含的有助于分类的信息越多

信息熵仅是判断差异的一种方法，其他能判断这划分差异的机制都能用于特征子集的评价

* 将特征子集搜索机制与子集评价机制结合就能得到特征选择方法

1. **常见的特征选择方法**：过滤式、包裹式、嵌入式 **去掉取值变化小的特征**
2. **过滤式(filter)**：先对数据集进行特征选择，然后再用过滤后的特征训练模型
3. **包裹式(wrapper)**：实际上就是一个分类器，它是将后续的学习器的性能作为特征子集的评价标准。
4. **嵌入式(embedding)**：实际上是学习器自主选择特征
5. 过滤式选择

* 设计一些统计量来过滤特征，并不考虑后续学习器问题
* 计算时间上比较**高效**，而且对过拟合问题有较高的鲁棒性
* **倾向于选择冗余特征**，即没有考虑到特征之间的相关性

1. **Relief方法**：

设计一个**相关统计量来度量特征**的重要性 --------为二分类问题而设计

该子集的重要性由该子集中**每个特征对应的相关统计量分量之和**来决定

指定一个阈值K，选择比K值大的相关统计量分量所对应的特征即可

或指定特征个数N，选择相关统计量分量最大的N个特征

1. **方差选择法**

计算各个特征的方差，然后根据阈值K，选择方差大于K的特征

1. **相关系数法**

先计算各个特征对目标值的相关洗漱以及相关系数的P值

1. **卡方检验**

检验定性自变量对定性因变量的相关性

假设自变量有N种取值，因变量有M种取值，考虑自变量等于i且因变量等于j的样本聘书的观察值与期望的差距，构建统计量：

1. **互信息法**

评价定性自变量对定性因变量的相关性，计算公示如下：

（二）包裹式选择

包裹式特征选择直接把最终将要使用的学习器的性能作为特征子集的评价选择

其目的就是为给定学习器选择最有利与其性能、量身定做的特征子集

**通常包裹式特征选择比过滤式特征选择能训练得到一个更好性能的学习器**

**缺点：开销大**

1. **LW：Las Vegas Wrapper：**典型的包裹式特征选择方法；使用随机策略来进行子集搜索，并以最终分类器的误差作为特征子集的评价标准

（三）嵌入式选择

即学习过程中自动进行特征选择

1. 正则化：主要用于线性回归、逻辑回归、支持向量机
2. L\_1范数：降低过拟合风险，容易获得稀疏解；

L\_1正则化的特征选择过程与学习器训练过程融为一体，二者同时完成

1. 常见的嵌入式选择模型：

在Lasso中：参数控制稀疏性

* 越小，则稀疏性越小，被选择的特征就越多

在SVM和逻辑回归中，参数c控制了稀疏性

* c越小，则稀疏性越大，被选择的特征就越少

1. 决策树：决策树、随机森林、gradient boosting等

二、 特征提取：将原始数据转换成一组具有明显物理意义或者统计意义的特征

常用的方法：

1. 降维
2. PCA主成分分析 --- 线性降维方法

旨在找到数据中的主成分，并利用这些主成分来表征原始数据，从而达到降维目的

思想：通过坐标轴转换，寻找数据分布的最优子空间

PCA的解法：（将数据映射到方差比较大的方向上）

1. 对样本数据进行中心化处理
2. 对样本协方差矩阵进行特征值分析，将特征值从大到小排列，
3. 取特征值前n个最大的对应的特征向量W1,W2,…,Wn，将原来的m维降到n维

通过PCA，可以将方差较小的特征抛弃，特征向量理解为坐标转换中新坐标轴的方向；

特征值表示在对应特征向量上的方差，**特征值越大，方差越大，信息量也越大**

1. LDA线性判别分析

LDA是一种有监督学习算法，相对于PCA，LDA擅长处理带有类别信息的数据，

因此LDA的目的不仅是降维，还需要找到一个投影方向，使得投影后的样本尽可能按照原始类别分开，即**寻找一个可以最大化类间距离以及最小化类内距离的方向**

1. ICA独立成分分析

PCA特征转换降维提取的是不相关的部分，ICA获得的是相互独立的属性

ICA算法得本质是寻找一个线性变换Z=WX，使得z的各个分量之间的独立性最大

**通常先使用PCA对数据降维，在使用ICA来从多个维度分离出有用数据**；

PCA是ICA的数据预处理方法

* 特征提取 VS 特征选择

同：都是为了从原始特征中找出最有效的特征；都能帮助减少特征的维度、数据冗余

异：

特征提取：

* 强调通过特征转换的方式得到一组具有明显物理意义或者统计意义的特征
* 更能发现有意义的特征属性

特征选择：

* 是从特征集合中挑选一组具有明显物理意义和统计意义的特征子集

· 能表示每个特征的重要性对于构建模型的重要性

三、特征构建

从原始数据中人工的构建新的特征

需要从原始数据中找出一些具有物理意义的特征

一般可以使用混合属性或者组合属性来创建新的特征

或是分解或者切分原有的特征来创建新特征

## 数据预处理：

1. 数据分类
2. **结构化数据**：可以看作是关系型数据库的一张表，每列都有清晰的定义，包含了数值型和类别型两种基本类型；每一行数据表示一个样本信息
3. 非结构化数据：主要是文本、图像、音频和视频数据；其包含的信息无法用一个简单的数值信息表示，也没有清晰的类别定义，并且每个数据的大小互不相同
4. 结构化数据处理
5. 处理缺失值
6. 、数据缺失主要包括**记录的缺失**和记录中某个**字段信息的缺失**
7. 导致的影响：

· 丢失大量有用信息

· 包含空值的数据会使模型陷入混乱，导致不可靠的输出

·表现出不确定性，模型中蕴含的规律更难把握

1. 处理方法

·直接使用含有缺失值的特征：当仅有少量样本缺失是可以尝试使用

·删除含有缺失值的特征；该方法适用于大多数样本都缺少该特征，却仅包含少量有效值是有效的

·**插值补全缺失值**

**补全方法：**

1. 均值、中位数、众数补全
2. 同类均值、中位数、众数补全
3. 固定值补全
4. 建模预测：将缺失属性作为预测目标预测
5. 高维映射：将属性映射到高维空间，采用**独热码编码技术（one-hot）**。将包含K个离散取值范围的属性值扩展为K+1个属性值缺失，则扩展后的第K+1给属性值置为1
6. 多重插补：认为待插补的值是**随机**的，实践上通常是估出待插补的值
7. 压缩感知和矩阵补全

· 压缩感知通过信号本身所具有的稀疏性，从部分观测样本中回复原信号

·阶段一：感知测量------- 对原始信号进行处理以获得稀疏样本表示；常用手段是傅里叶变换、小波变换、字典学习、稀疏编码

·阶段二：重构恢复-------- 基于稀疏性从少量观测中恢复信号；核心阶段

1. 手动补全
2. 最近邻补全----- **寻找与该样本最接近的样本**，使用其该属性数值来补全
3. 处理异常值

异常值-----样本中的个别值，其数值明显偏离其余的观测值；也称离群点

1. **异常值分析**是检验数据是否录入错误以及含有不合常理的数据；若把异常值包括进数据的计算分析过程中对结果会产生不良影响
2. **异常值检测**
3. 简单统计 ：观察数据的统计性描述 pandas.describe()
4. 3原则：**数据需要服从正态分布**，

·在3原则下，异常值如果超过3倍标准差，则视为异常值；正负3的概率是99.7%，异常值出现属于极个别的小概率事件；

·如果数据不服从正态分布，可以用原理平均值的多少倍来描述

1. 箱线图：利用四分位距IQR对异常值及逆行检测

·超过**上四分位+1.5倍IQR距离，或者下四分位-1.5倍IQR距离**的点为异常值

1. 基于模型预测：

构建一个概率分布模型，并计算对象符合该模型的概率，将低概率的对象视为异常点

1. 基于近邻度的离群电检测

一个对象的离群点得分由到它的k-最近邻KNN的距离给定

1. 基于密度的离群点检测

法1：定义密度为k个最近邻的平均距离的倒数，若距离小，则密度高

法2：DBSCAN聚类算法，一个对象周围的密度=该对象指定距离d内对象的个数

1. 基于聚类的离群点检测
2. 专门的离群点检测
3. **异常值处理**

**删除含有异常值的记录**：直接将含有异常值的记录删除；数据集少不适用

**视为缺失值**：将异常值视为缺失值，利用缺失值处理的方法进行处理；

**平均值修正**：可用前后两个观测值的平均值修正该异常值；

**不处理**：直接在具有异常值的数据集上进行数据挖掘

1. 处理类别不平衡
2. 类别不平衡：指分类任务中存在某个或者某些类别的样本数量**远多于**其他类别的样本数量的情况
3. 解决方案
4. 扩充数据集
5. 尝试其他评价指标：混淆矩阵、精确度、召回率、ROC曲线…
6. 对数据集重采样
7. 尝试人工生成数据样本

·对该类下的所有样本每个属性特征的取值空间中随机选取一个组成新的样本，即属性值随机采样

1. 尝试不同分类算法：决策树
2. 尝试对模型进行惩罚
3. 尝试一个新角度理解问题
4. 尝试创新：是否可以将问题划分为多个更小的问题

·将你的大类压缩成小类；

·使用 One Class 分类器（将小类作为异常点）；

·使用集成方式，训练多个分类器，然后联合这些分类器进行分类；

## 特征缩放：归一化、正则化

1. 归一化

·也称标准化；将特征或者数据都缩放到一个指定的大致相同的数值区间

·归一化的原因：

1、某些算法要求样本数据或者特征的数值具有零均值和单位方差

2、为了消除样本数据或者特征之间的量纲影响，即消除数量级的影响

3、数量级的差异将导致数量级较大的属性占据主导地位

4、数量级的差异会导致迭代收敛速度减慢

5、所有依赖于样本距离的算法对于数据的数量级否非常敏感

·常见归一化方法：

1. 线性函数归一化（Min-Max Scaling）,使得结果映射到[0,1]，实现等比缩放

零均值归一化：将原始数据映射到均值为0标准差为1的分布上

训练集、验证集和测试集要采用相同的归一化参数，数值都是通过训练集计算得到的

通过梯度下降法求解的模型时需要归一化的，包括线性回归、逻辑回归、支持向量机、神经网络等模型；决策树不需要归一化

1. 正则化 ------ 将样本或者特征的某个范数L1\L2缩放到单位1

正则化的结果是：每个属性值除以其Lp范数

·正则化——针对单个样本，对每个样本将它缩放到单位范围

·归一化——针对单个属性，需要用到所有样本在该属性上的值

·通常如果使用二次型或者其他核方法计算两个样本之间的相似性时，该方法有用

1. 特征编码
2. 序号编码——用于处理类别间具有大小关系的数据
3. 独热编码（one-hot encoding）——处理类别间不具有大小关系的数据

采用N位状态位来对N个可能的取值进行编码；比如血型A B AB O转换成一个4维稀疏向量：

A——(1,0,0,0) B——(0,1,0,0) AB——(0,0,1,0) O——(0,0,0,1)

·独热编码可以处理非数值属性

·编码后是**稀疏向量**，只有一位1；可节省存储空间

·能**处理缺失值**：当所有位都为0，表示发生了缺失；此时可以采用高维映射，用第N+1位来表示缺失值

1. 二进制编码

——先采用序号编码给每个类别赋予一个类别ID

——接着将类别ID对应的二进制编码作为结果



二进制编码本质上是利用**二进制对类别 ID 进行哈希映射，最终得到 0/1 特征向量，并且特征维度小于独热编码，更加节省存储空间**

1. 二元化

——特征二元化是将数值型的属性转化为布尔型的属性；通常用于假设属性来取值扽不是伯努利分布的情形

·对于属性j指定一个阈值m：

·若样本在属性j上的值大于等于m；则二元化后位1；

若样本在属性j上的值小于m，则二元化为0

--- 根据以上定义，m是一个超参数，它的取值需要结合模型和具体的任务来选择

1. 离散化 —— 将连续的数值属性转换为离散一系列0/1的数值属性

“海量离散特征+简单模型”——对于线性模型

“少量连续特征+复杂模型”——对于非线性模型

·离散化方法——**分桶**

1. 将所有样本在连续的数值属性 j 的取值从小到大排列。
2. 从小到大依次选择分桶边界。其中分桶的数量以及每个桶的大小都是超参数，需要人工指定。每个桶的编号为 0，1，...，M，即总共有 M 个桶。
3. 给定属性 j 的取值 a，判断 a 在哪个分桶的取值范围内，将其划分到对应编号 k 的分桶内，并且属性取值变为 k

·分桶的数量和边界通常由人工指定的方法：

·根据业务领域的经验来指定

·根据模型指定

·分桶的大小：

·足够小

·足够大；使得每个桶内都有足够的样本

·每个桶内的样本尽量分布均匀

·优点

**特征离散化简化了逻辑回归模型，同时降低模型过拟合的风险**

特征缩放是非常常用的方法，特别是归一化处理特征数据，对于利用梯度下降来训练学习模型参数的算法，有助于**提高训练收敛的速度**；而特征编码，特别是独热编码，也常用于对**结构化数据的数据预处理**