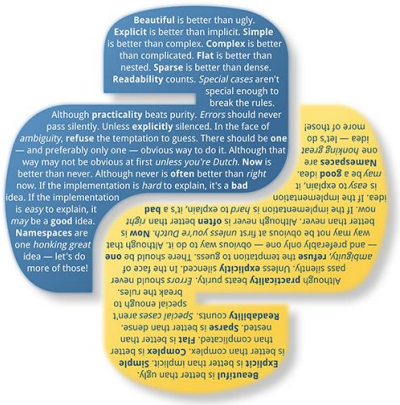
CDA 数据分析师
CERTIFIED DATA ANALYST

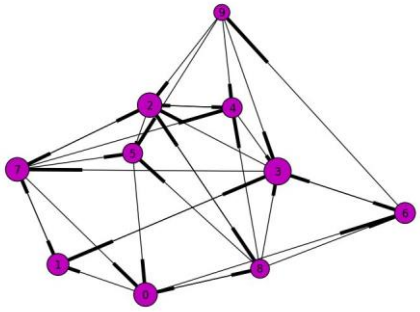
精英班系列课程

特征工程



python™

特征及特征工程

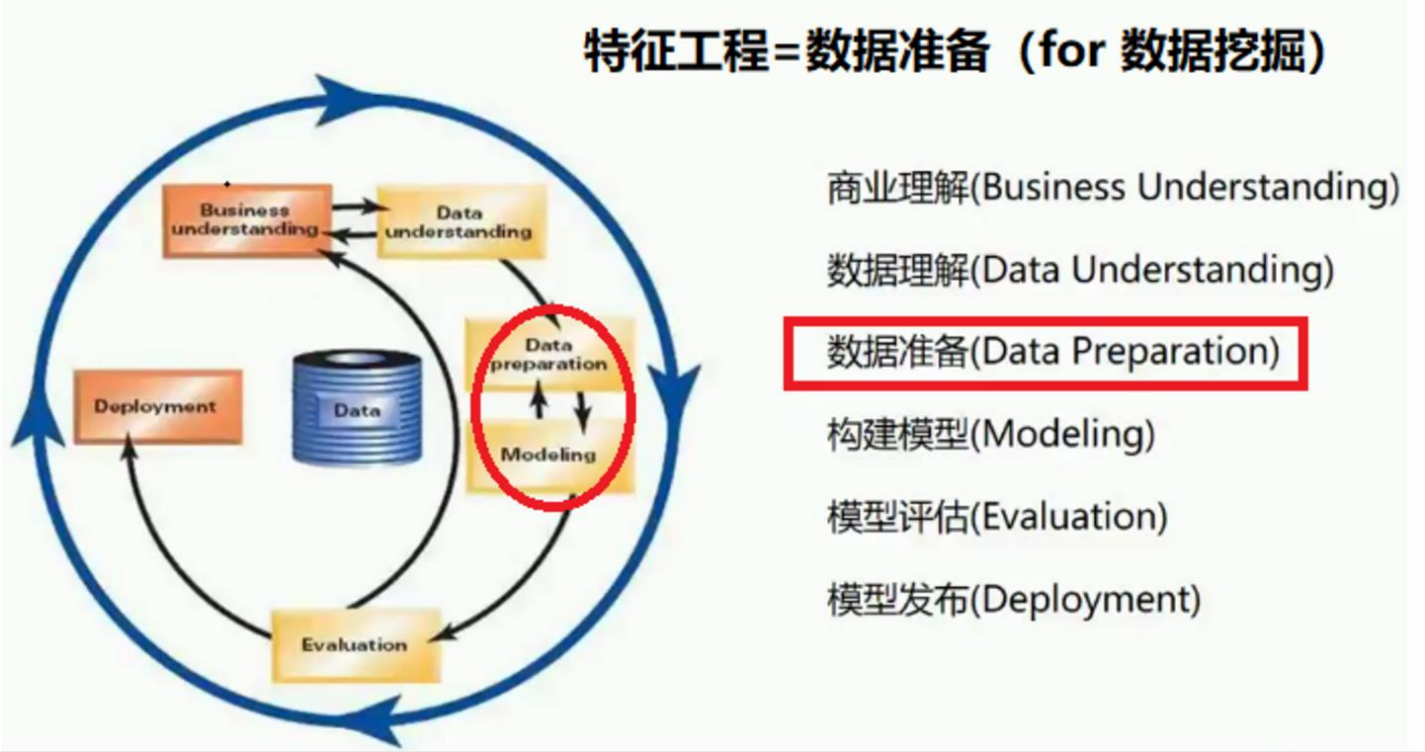


特征及特征工程

- 特征的定义：
  - ✓ 在数据科学过程中的有效属性（或字段）的形式称为特征
- 常用特征：
  - ✓ 计算机视觉：图像像素，边缘，角点
  - ✓ 语音识别：声音，噪声等
  - ✓ 自然语言处理:文档频率，互信息，信息增益等
- 特征工程的定义：
  - ✓ 特征工程是将原始数据转换为更好地代表预测模型的潜在问题的特征的过程，从而提高对未知数据预测的准确性

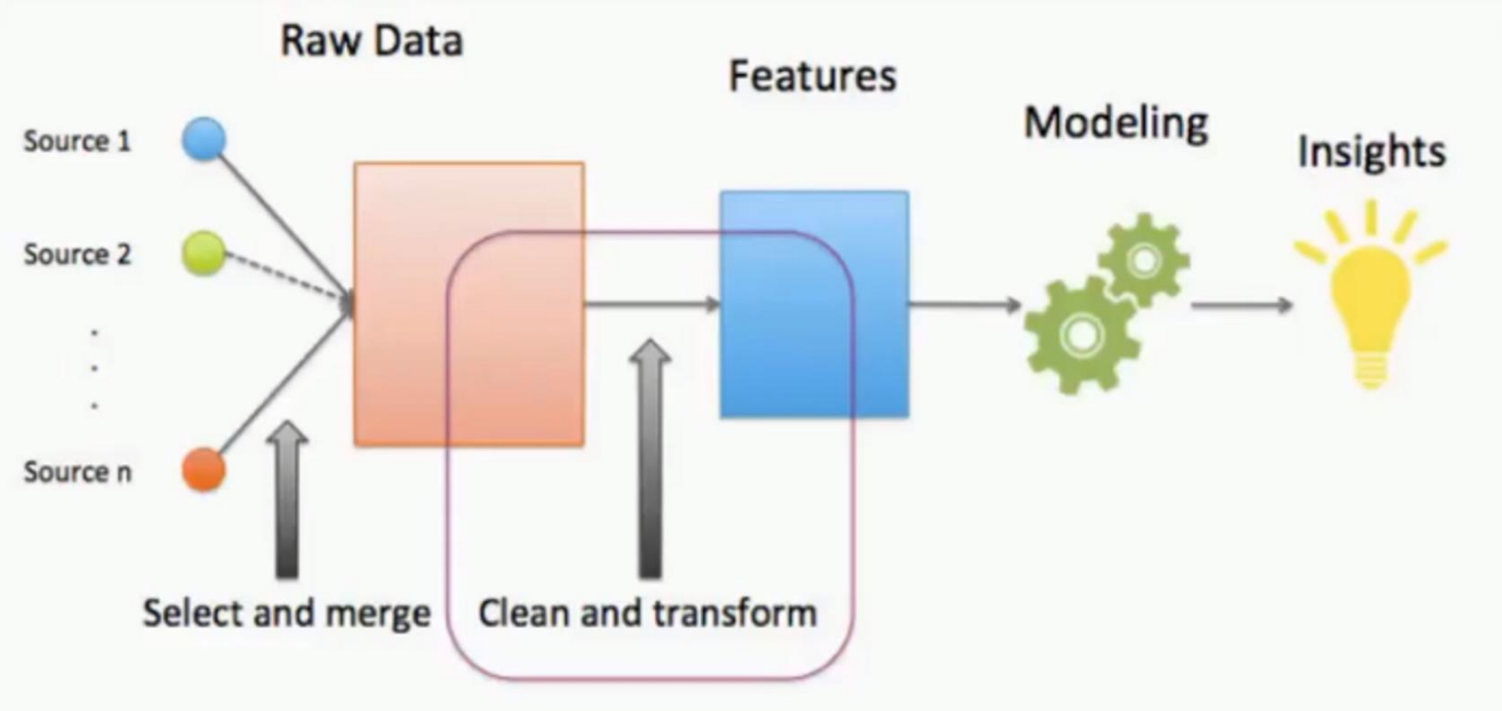
特征及特征工程

- CRISP-DM：跨行业数据挖掘标准流程



特征及特征工程

- 特征工程在机器学习流程中的位置：



特征及特征工程

- 特征工程的目的：
  - ✓ 数据是信息的载体，但是原始的数据包含了大量的噪声，信息的表达也不够简练。因此，特征工程的目的，是通过一系列的工程活动，将这些信息使用更高效的编码方式（特征）表示。使用特征表示的信息，信息损失较少，原始数据中包含的规律依然保留。此外，新的编码方式还需要尽量减少原始数据中的不确定因素（白噪声、异常数据、数据缺失等等）的影响

特征及特征工程

- 特征工程的重要性：数据科学家在解决问题时，会花费超过一半的时间来选择正确的特征

  - “More data beats clever algorithms, but better data beats more data.” 好数据>多数据>好算法 –Peter Norvig
  - 数据和特征决定了模型预测的上限，而算法只是逼近这个上限而已
  - “Applied machine learning” is basically feature engineering（应用机器学习基本上就是特征工程）–Andrew Ng

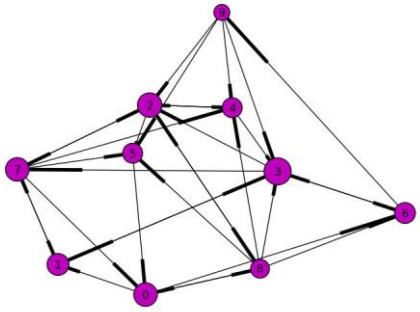
特征及特征工程

- 特征工程做什么：母婴人群标签精准细分画像

特征及特征工程

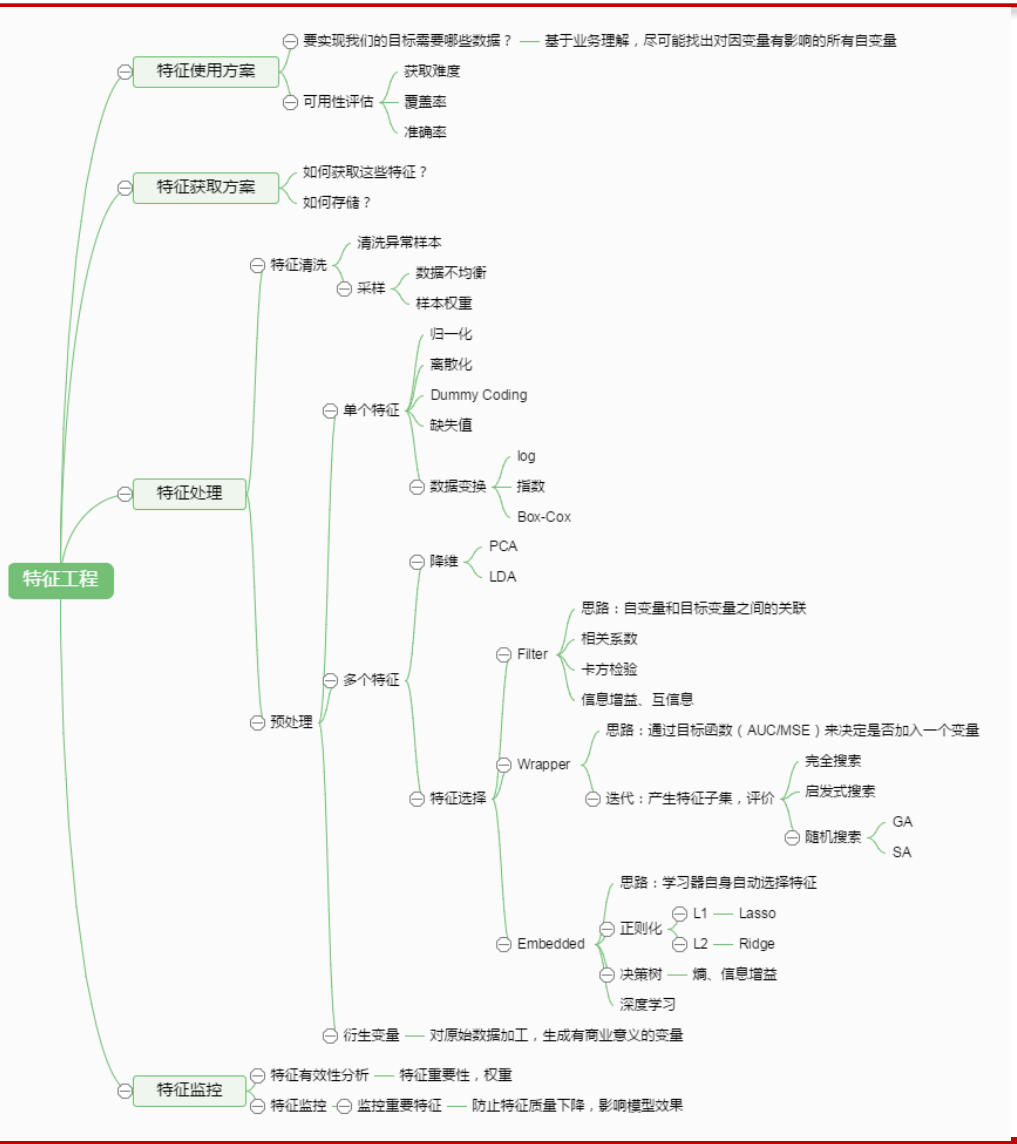
- 什么是好的特征？-少而精！
  - ✓ 模型更简单：同样的模型精度选择更简单的模型
  - ✓ 模型更精准：好的特征是数据中抽取出来的对预测结果最有用的信息

常用特征工程方法



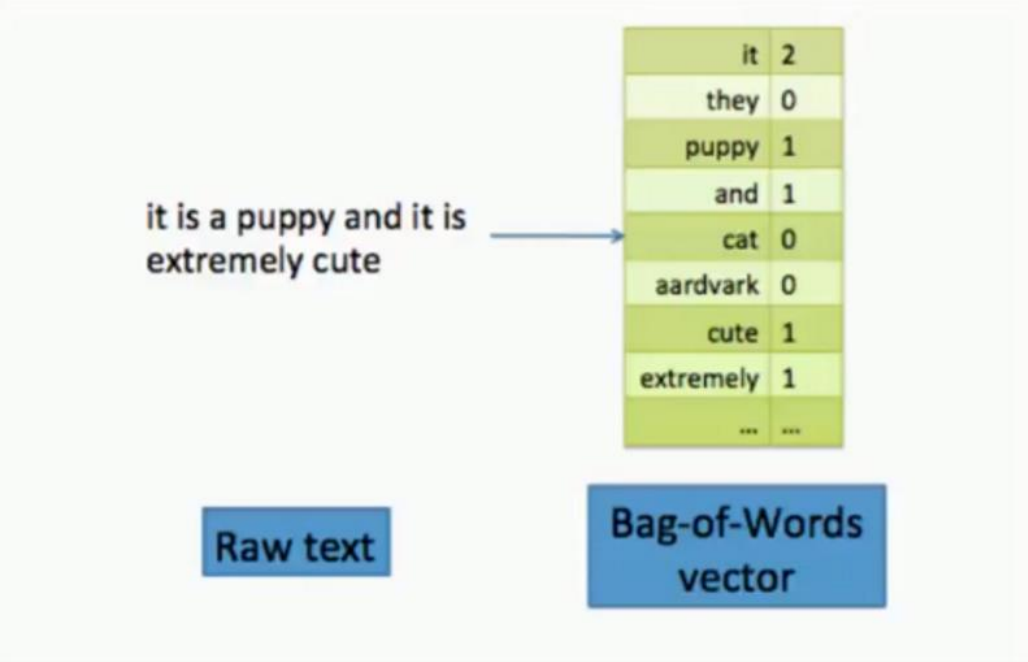
常用特征工程方法

特征工程包括



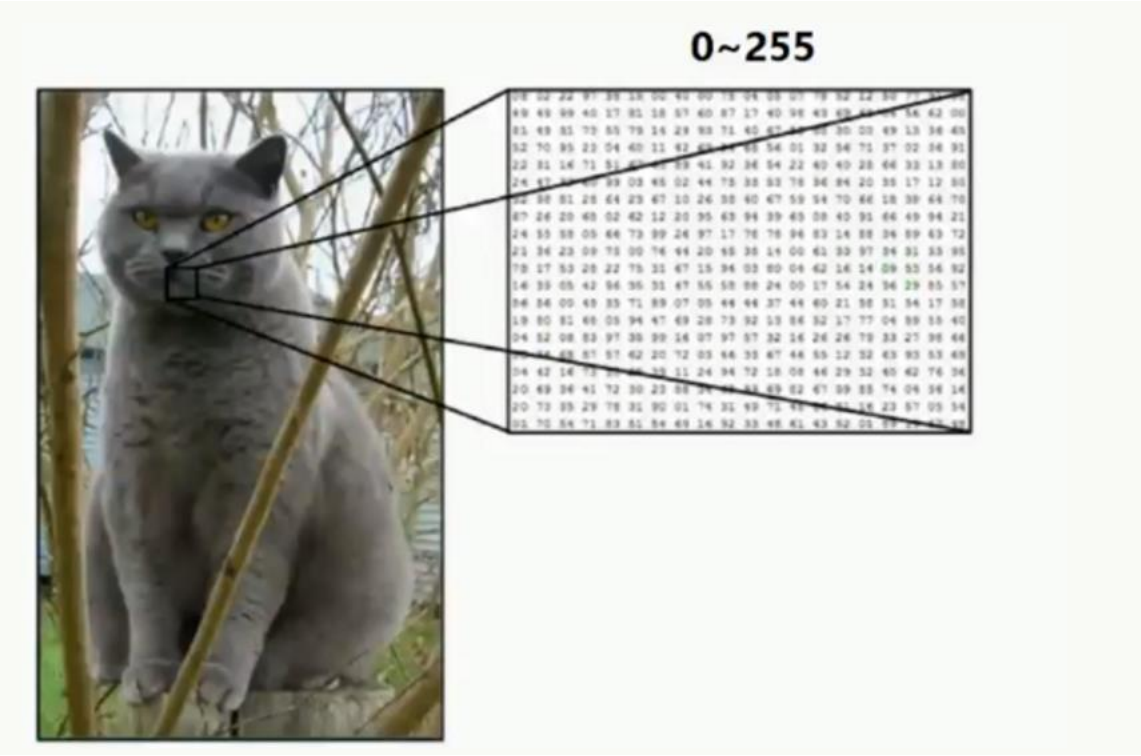
常用特征工程方法-特征提取

- 特征提取：文本数据的特征提取（词袋模型，TF-IDF等）



常用特征工程方法-特征提取

- 特征提取：图像数据的特征提取
  - ✓ 图像构成：像素+颜色（RGB，每个通道0~255）
  - ✓ 图像的每个像素点：RGB的值



常用特征工程方法-特征提取

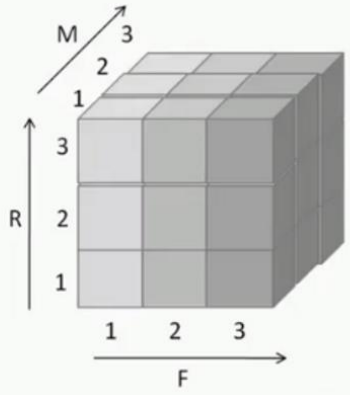
- 特征提取：用户行为特征（RFM）
  - ✓ 提取最近一次消费时间、消费频率、总的购买金额
  - ✓ 购买商品类型、价格

用户交易数据

用户ID，交易时间，交易金额

CardID,	Date,	Amount
"C0100000199",	2011/08/20,	229.000000
"C0100000199",	2011/06/28,	139.000000
"C0100000199",	2011/12/29,	229.000000
"C0100000343",	2011/07/27,	49.000000
"C0100000343",	2011/02/02,	169.990000
"C0100000343",	2011/07/12,	299.000000
"C0100000343",	2011/02/02,	34.950000
"C0100000343",	2011/09/07,	99.000000
"C0100000343",	2011/05/13,	49.000000
"C0100000375",	2011/09/22,	99.990000
"C0100000375",	2011/05/02,	5.990000
"C0100000375",	2011/11/01,	49.000000
"C0100000375",	2011/10/16,	69.000000
"C0100000482",	2011/08/12,	84.000000
"C0100000482",	2011/03/28,	69.000000
"C0100000482",	2011/04/03,	24.990000
"C0100000482",	2011/12/10,	19.990000
"C0100000689",	2011/05/23,	79.000000

- 最近一次消费时间(Recency)
- 消费频率(Frequency)
- 购买金额(Monetary)



用户M特征扩展（分item）

用户id	item1_M	item2_M	item3_M	item4_M	item5_M
1	245	801	88	45	
2		180	128		480

用户F特征扩展（分item）

用户id	item1_F	item2_F	item3_F	item4_F	item5_F
1	1	2	1	1	
2		1	1		2

常用特征工程方法

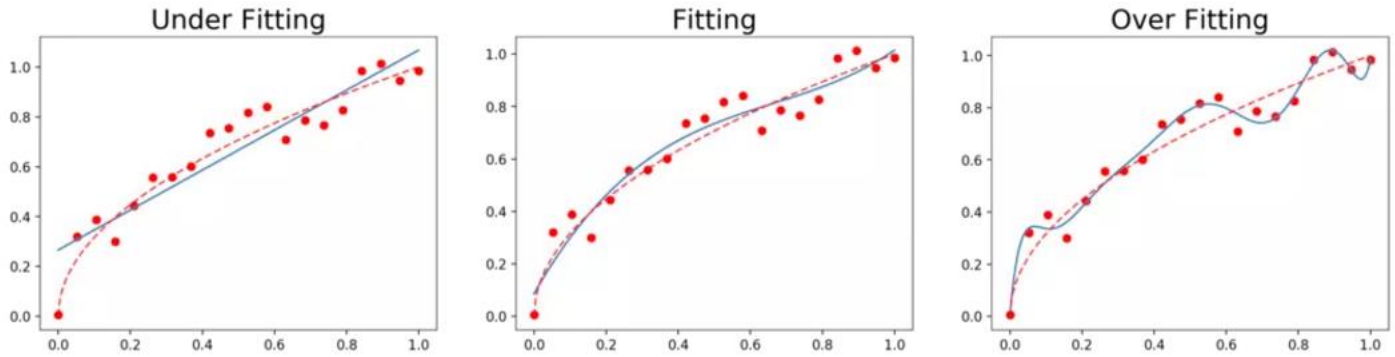
- 通过特征提取我们能得到未经处理的特征，这时的特征可能有以下问题：
  - ✓ 不属于同一量纲：即特征的规格不一样，不能够放在一起比较。无量纲化可以解决这一问题
  - ✓ 信息冗余：对于某些定量特征，其包含的有效信息为区间划分，例如学习成绩，假若只关心“及格”或不“及格”，那么需要将定量的考分，转换成“1”和“0”表示及格和未及格。二值化可以解决这一问题

常用特征工程方法

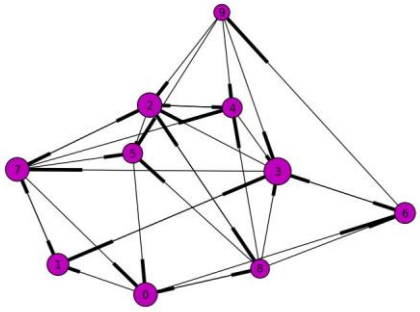
- 通过特征提取我们能得到未经处理的特征，这时的特征可能有以下问题：
  - ✓ 定性特征不能直接使用：某些机器学习算法和模型只能接受定量特征的输入，那么需要将定性特征转换为定量特征。最简单的方式是为每一种定性值指定一个定量值，但是这种方式过于灵活，增加了调参的工作。通常使用哑编码的方式将定性特征转换为定量特征：假设有N种定性值，则将这一个特征扩展为N种特征，当原始特征值为第i种定性值时，第i个扩展特征赋值为1，其他扩展特征赋值为0。哑编码的方式相比直接指定的方式，不用增加调参的工作，对于线性模型来说，使用哑编码后的特征可达到非线性的效果

常用特征工程方法

- 通过特征提取我们能得到未经处理的特征，这时的特征可能有以下问题：
  - ✓ 信息利用率低：不同的机器学习算法和模型对数据中信息的利用是不同的，之前提到在线性模型中，使用对定性特征哑编码可以达到非线性的效果。类似地，对定量变量多项式化，或者进行其他的转换，都能达到非线性的效果



特征处理



常用特征工程方法-特征处理

- 特征处理：
  - ✓ 特征处理是特征工程的核心部分，sklearn提供了较为完整的特征处理方法，包括数据预处理，特征选择，降维等

常用特征工程方法-特征处理

- 特征处理：缺失值数据处理
  - ✓ 删除包含缺失数据的特征（列）
  - ✓ 删除包含缺失数据的行
  - ✓ 用重要的数据进行填充：平均值，中值，零，众数

常用特征工程方法-特征处理

CDA 数据分析师 CERTIFIED DATA ANALYST

- 特征处理：无量纲化
  - ✓ 无量纲化使不同规格的数据转换到同一规格。常见的无量纲化方法有标准化，区间缩放法，正则化
  - ✓ 标准化的前提是特征值服从正态分布，标准化后，其转换成标准正态分布

$$x' = \frac{x - \overline{X}}{S}$$

  - ✓ 区间缩放法利用了边界值信息，将特征的取值区间缩放到某个特点的范围，例如[0,1]等。常见的一种为利用两个最值进行缩放

$$x' = \frac{x - Min}{Max - Min}$$

常用特征工程方法-特征处理

特征处理：无量纲化

- 正则化是依照特征矩阵的行处理数据，其目的在于样本向量在点乘运算或其他核函数计算相似性时，拥有统一的标准，也就是说都转化为“单位向量”。规则为l2的正则化公式如下：

$$x' = \frac{x}{\sqrt{\sum_j^m x[j]^2}}$$

常用特征工程方法-特征处理

- 特征处理：对定量特征二值化
  - ✓ 定量特征二值化的核心在于设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0，公式表达如下：

$$x' = \begin{cases} 1, x > threshold \\ 0, x \leq threshold \end{cases}$$

常用特征工程方法-特征处理

- 特征处理：对定性特征（分类特征）LabelEncoder编码
  - ✓ 用每个数字代替每个值，对不连续的数字或者文本进行编号

常用特征工程方法-特征处理

- 特征处理：对定性特征哑编码
  - ✓ 哑编码主要是针对定性的特征进行处理然后得到可以用来训练的特征
  - ✓ 使用preproccessing库的OneHotEncoder类对数据进行哑编码

常用特征工程方法-特征处理

CDA 数据分析师 CERTIFIED DATA ANALYST

- 特征处理：数据变换
  - ✓ 常见的数据变换有基于多项式的、基于指数函数的、基于对数函数的
  - ✓ 多项式特征变换（增维），目标是将特征两两组合起来，使得特征和目标变量之间的的关系更接近线性，从而提高预测的效果
  - ✓ 4个特征，度为2的多项式转换公式如下：

$$(x'_1, x'_2, x'_3, x'_4, x'_5, x'_6, x'_7, x'_8, x'_9, x'_{10}, x'_{11}, x'_{12}, x'_{13}, x'_{14}, x'_{15}) = (1, x_1, x_2, x_3, x_4, x_1^2, x_1 * x_2, x_1 * x_3, x_1 * x_4, x_2^2, x_2 * x_3, x_2 * x_4, x_3^2 x_3 * x_4, x_4^2)$$

常用特征工程方法-特征处理

CDA 数据分析师 CERTIFIED DATA ANALYST

- 特征处理：降维
  - ✓ 当特征矩阵过大时，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
  - ✓ 常见的降维方法有主成分分析法（PCA）和线性判别分析（LDA）
  - ✓ PCA和LDA的映射目标不一样：PCA是为了让映射后的样本具有最大的发散性；而LDA是为了让映射后的样本有最好的分类性能。所以说PCA是一种无监督的降维方法，而LDA是一种有监督的降维方法

常用特征工程方法-特征选择

- 特征选择：当数据预处理完成后，我们需要选择有意义的特征输入机器学习的算法和模型进行训练。通常来说，从两个方面考虑来选择特征：
  - ✓ 特征是否发散：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上没有差异，这个特征对于样本的区分并没有什么用
  - ✓ 特征与目标的相关性：与目标相关性高的特征，应当优选选择

常用特征工程方法-特征选择

- 特征选择：根据特征选择的形式又可以将特征选择方法分为3种：
  - ✓ Filter：过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值的个数，选择特征
  - ✓ Wrapper：包装法，根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征
  - ✓ Embedded：嵌入法，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，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣

常用特征工程方法-特征选择

- 特征选择：使用sklearn中的feature_selection库来进行特征选择

类	所属方式	说明
VarianceThreshold	Filter	方差选择法
SelectKBest	Filter	可选关联系数、卡方校验、最大信息系数作为得分计算的方法
RFE	Wrapper	递归地训练基模型，将权值系数较小的特征从特征集合中消除
SelectFromModel	Embedded	训练基模型，选择权值系数较高的特征

Talk is cheap

Show me the

CODE