|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | 变量 | 公式 | 应用情况 | 图例 |
| 平方根变换 | 进行变换，不变 |  | 变换前数据分布，偏中前 |  |
| 平方变换 | 进行变换，不变 |  | 变换前数据分布，偏中后 |  |
| 倒数变换 | 进行变换，不变 |  | 变换前数据分布，集中前面 |  |
| 取对数变换(最常用) | 进行变换，不变 |  | 变换前数据分布，偏 |  |
| 幂函数变换 | 进行变换，不变 |  | 变换前数据分布，偏后 |  |
| Box-Cox变换 | 进行变换，不变 | https://ask.qcloudimg.com/http-save/yehe-1051732/s8n3yuyz1m.jpeg?imageView2/2/w/1620 |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **方法** | **描述** | **公式** | **特点** |
| **最小-最大归一化** | **线性函数将原始数据线性化的方法转换到[0, 1]的范围** |  | **在不涉及距离度量、协方差计算、数据不符合正太分布的时候，使用该方法比较好。** |
| **0均值标准化** | **均值为0、方差1的数据集** |  | **在分类、聚类算法中，需要使用距离来度量相似性的时候、或者使用PCA技术进行降维、涉及到正态分布的时候使用该方法较好。** |
| **对数变换** | **类似对数变换** |  | **在实际工程中，经常会有类似点击次数/浏览次数的特征，这类特征是长尾分布的，可以将其用对数函数进行压缩。特别的，在特征相除时，可以用对数压缩之后的特征相减得到。** |
| **三角函数** | **三角函数的值在[-1, 1]之间** |  |  |
| Sigmoid函数 | **当数据趋向于正无穷和负无穷的时候，映射出来的值就会无限趋向于1和0** |  |  |
| **小数定标规范化** | **通过移动小数点的位置，进行规范化** |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **分类** | **方法** | **描述** | **总结** | **图例** |
| 无监督的方法 | **等宽划分** | 按照相同宽度将数据分成几等份。 | 对异常点比较敏感，倾向于不均匀地把实例分布到各个箱中，严重地损坏特征建立好的决策结构的能力。可能将具有相同类标签的相同特征值分入不同的箱中以满足箱中数据的固定个数的条件。方法简单，易于操作，但都需要人为地规定划分区间的个数这个参数。 |  |
| **等频划分** | 将数据分成几等份，每等份数据里面的个数是一样的。 |
| **聚类划分** | 使用聚类算法将数据聚成几类，每一个类为一个划分。 | 聚类分析的离散化方法也需要用户指定簇的个数，从而决定离散产生的区间数在没有类信息的情况下，要得到好的离散化结果比较困难，并且离散化的结果也比较难衡量。 |  |
| 有监督学习方法 | **1R方法** | 1R是一种使用分箱的有监督的方法。它把连续的区间分成小的区间，然后再使用类标签对小区间的边界进行调整。 | 1R离散化方法也是分箱方法，操作仍然比较方便，但又不再需要用户人为指定箱的个数，也克服了无监督的分箱方法的不使用类信息的缺陷，并且能避免把具有相同特征值相同类标签的实例分入不同的小区间中。 | |
| **基于信息熵的方法** | 决策树时用熵来分裂连续特征的准则在实际中运行得很好，考虑将这种思想扩展到更通常的离散化中，通过反复地分裂小区间直到满足停止的条件。 | 它的缺点在于使用的停止准则的原则化合理性是有待考虑的。分类和离散化毕竟是两个不能完全等同的处理过程。 |  |
| **基于卡方的方法** | 采用自底向上的策略，首先将数据取值范围内的所有数据值列为一个单独的区间，再递归地找出最佳邻近可合并的区间，然后合并他们，进而形成较大的区间。在判定最佳邻近可合并的区间时，会用到卡方统计量来检测两个对象间的相关度。 | 应用卡方统计量检验两个对象是否相关时，需要人为设定置信水平参数，由统计学知  识算出一个与计算量相比较的阈值。对于置信水平的设置要合理，过高会导致过分离散化，过低又会导致离散化不足。 |  |