目录

5 数据规范化

对数据作规范化处理, 是某些统计学方法的应用前提. 对于以变量尺度为重要依据的方法, 如主成分分析, 需要通过规范化使得变量尺度一致. 对于与距离有关的方法, 如k近邻和聚类分析, 若采用明氏距离(Minkowski Distance), 它基于各维度距离构建, 需要通过规范化处理获得尺度一致的各维度距离. 对于模型参数受到变量尺度影响的方法, 如神经网络和支持向量机, 一方面, 大尺度变量可能会过度影响模型, 导致模型效果欠佳; 另一方面, 拟合模型时需要自适应、迭代求解参数值. 通过规范化, 变量尺度和参数尺度变得一致, 从而可以提升模型效果, 并且有助于统一、高效地求解参数值.

变量类型不同, 数据规范化的方法亦不同. 下面, 分别针对定量和定性变量介绍几种常用的数据规范化方法.

5.1 定量变量的规范化

定量变量的规范化包含线性变换(标准化、最小值-最大值规范化、白化)和幂变换. 除了按变量作规范化, 还可以按样本观测作规范化, 即行归一化.

5.1.1 标准化

许多统计学方法要求各维度变量具有统一的中心和(或) 尺度, 因而需要对变量作位置变换和(或) 尺度变换. 标准化(standardization) 涵盖了对变量的位置和(或) 尺度作变换. 对变量的取值x作标准化变换, 计算公式为:

$$x' = \frac{x - L}{S},$$

其中L表示位置的平移量, S表示尺度的缩放量.

位置的平移量L有多种取值, 例如:

(1) 样本均值. 此时, 标准化后的变量X'的样本均值为0, X'的取值可以反映X的取值与均值的关系. 但是, 样本均值易受离群值影响, 导致标准化变换不稳健.

(2) 样本中位数. 此时, 标准化后的变量X'的中位数为0, X'的取值可以反映X的取值与中位数的关系. 样本中位数不易受离群值影响, 因而基于样本中位数的标准化变换是稳健的.

(3) 某一设定值. 例如, 取L为指标的标准值, 那么标准化后的变量X'可反映指标实际值与标准值之间的差异. 也可以取$L = 0$, 表示不作中心化.

尺度的缩放量S也有多种取值, 例如:

(1) 样本标准差. 此时, 标准化后的变量X'的样本标准差为1, 可以实现各变量尺度的统一(若以标准差衡量尺度). 但是, 样本标准差易受离群值影响, 导致标准化变换不稳健.

(2) 样本平均绝对离差, 即$\frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$, 其中$\bar{x}$是样本均值, n是样本量. 相对于标准差而言, 平均绝对离差受离群值影响的程度更小, 但仍然不是稳健的统计量.

(3) 四分位距(Inter Quartile Range, IQR), 即$Q_U - Q_L$, 其中Q_U是上四分位数, Q_L是下四分位数. IQR刻画了中间50%数据的离散程度, 是稳健的统计量.

(4) 某一设定值. 例如, 取$S = 1$, 表示不作尺度变换.

5.1.2 最小值-最大值规范化

最小值-最大值规范化(min-max normalization) 可以将变量取值变换到指定范围. 若想将变量的取值$x \in [x_{\min}, x_{\max}]$变换到指定的取值范

围$[x'_{\min}, x'_{\max}]$, 那么最小值-最大值规范化的表达式为:

$$x' = x'_{\min} + R'\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right), \tag{5.1}$$

其中$R' = x'_{\max} - x'_{\min}$.

若各个变量最小值-最大值规范化后的取值范围$[x'_{\min}, x'_{\max}]$相同, 例如, 常取作$[0, 1]$, 那么通过最小值-最大值规范化可以保证各个变量取值落入相同的值域区间. 若以极差作为尺度的衡量标准, 最小值-最大值规范化后的变量可保持一致的尺度.

可以看到, 若数据中存在离群值, 最小值-最大值规范化可能使得大部分数据集中于一个很小的范围内, 从而削弱变量取值的区分能力. 此时, 可以考虑将(5.1)式中的$x_{\min}$和$x_{\max}$替换为分位数, 例如分别替换为1%和99%分位数, 但此时不能保证所有的数据落入区间$[x'_{\min}, x'_{\max}]$之中.

注记5.1. 应基于训练集训练最小值-最大值规范化, 即由训练集获得$x_{\min}$和$x_{\max}$. 当在验证集或测试集中应用训练好的最小值-最大值规范化时, 可能出现取值越出$[x'_{\min}, x'_{\max}]$的情况. 通常, 这是可以允许的. 若希望严格控制取值范围, 可将(5.1)式更改为

$$x' = x'_{\min} + \max\left\{x'_{\max},\ \min\left\{0,\ R'\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right)\right\}\right\}.$$

5.1.3 幂变换

幂变换主要用于改善变量的正态性、对称性和波动稳定性(方差齐性). 常用的幂变换方法包含Box-Cox变换和Yeo-Johnson变换.

Box-Cox变换(Box-Cox transformation) 由Box & Cox (1964) 提出. 当变量的取值$x > 0$时, Box-Cox变换的表达式为:

$$x' = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \ln x, & \lambda = 0, \end{cases}$$

其中λ是设置的参数值. 通常, 需要基于数据信息获得恰当的λ, 例如, 为使得变换后的数据最符合正态性, 可基于正态分布构建似然函数, 取使得似然函数达到极大值的λ.

当变量含有小于0的取值时, 可采用Yeo-Johnson变换(Yeo-Johnson transformation), 它由Yeo & Johnson (2000) 提出, 表达式为:

$$x' = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0, x \geq 0, \\ \ln(x+1), & \lambda = 0, x \geq 0, \\ \dfrac{-(-x+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2, x < 0, \\ -\ln(-x+1), & \lambda = 2, x < 0. \end{cases}$$

注记5.2. 虽然幂变换有助于改善变量的正态性、对称性和波动稳定性, 但不能保证变换后的变量一定满足这些特性. 应对变换后的变量进行诊断.

注记5.3. 幂变换可能影响模型的可解释性. 例如, 对因变量Y作幂变换$f(\cdot)$, 获得变换后的变量$Y' = f(Y)$. 以Y'为因变量, $\boldsymbol{X}$为自变量, 建立普通最小二乘回归模型. 对于幂变换$f(\cdot)$, 通常情况下$E[f(Y)] \neq f(E[Y])$, 因而无法基于模型解释$\boldsymbol{X}$对$E[Y]$的影响.

5.1.4 白化

白化(whitening transformation), 也称为球化(sphering transformation). 设$\boldsymbol{X}_{n\times p}$为数据矩阵, 将$\boldsymbol{X}$作白化的一种方式为:

$$\boldsymbol{X}' = (\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2},$$

其中$\bar{\boldsymbol{X}}$为总体(或样本) 均值, $\boldsymbol{\Sigma}$为总体(或样本) 协方差矩阵.

白化最重要的作用是使得变换后任意两个维度变量的协方差都为0, 从而可以消除变量间相关的冗余信息, 便于开展后续的统计分析(如, 独立成分分析). 同时, 白化后每一维变量的方差都为1, 尺度统一. Ranzato et al. (2010) 采用受限玻尔兹曼机处理图像时发现, 虽然白化并非必要的, 但可以加速算法收敛.

5.1.5 行归一化

行归一化是将数据按行作尺度变换, 使得变换后每行有单位范数(unit norm), 意味着每行的向量长度相同, 相当于将各样本观测都映射到单位圆上.

例5.1. 对表5.1中的变量X_1和X_2, 使用L_2范数归一化. 对于第1行数据, X_1与X_2取值平方和为$3^2 + 4^2 = 25$, 因此, X_1与X_2的行归一化值分别为$\sqrt{3^2/25} = 0.6$和$\sqrt{4^2/25} = 0.8$. 类似可得到第2行数据的行归一化值(见表5.1中行归一化后变量X_1'和X_2'的取值).

表 5.1: 行归一化示例

序号	X_1	X_2	X_1'	X_2'
1	3	4	0.6	0.8
2	30	40	0.6	0.8

除了使用L_2范数, 还可以使用L_1范数和最大值等进行归一化.

当我们关心对象的**相对量**时, 可应用行归一化. 例如, 若我们并不关心每个省各产业GDP的绝对量, 而是关心各产业GDP在GDP中所占的比重, 也就是经济结构性问题时, 可使用行归一化.

5.2 定性变量的规范化

定性变量的取值仅是区分类别的编号或符号, 不具有量化的信息, 难以直接放入模型之中. 一般需要通过变换, 以量化的形式表达定性变量. 一种常用的处理方式是将一个定性变量表达为若干个取值为0或1的虚拟变量(dummy variables), 从而将虚拟变量作为定量变量处理, 这就是独热编码. 此外, 一部分模型针对定性变量进行设计, 具备处理定性变量的能力, 因而不需要作独热编码. 但是, 软件中实现这类模型的函数往往不能接受字符串类型的定性变量, 此时需要对定性变量作有序编码.

5.2.1 独热编码

独热编码(one-hot encoding) 是一种常用的定性变量规范化方法. 它将一个定性变量变换为若干个虚拟变量, 每一个虚拟变量对应于定性变量的某一个类别, 当定性变量取值为对应该类别时, 该虚拟变量取值为1, 否则取值为0. 这就是“独热”的含义, 即取值1代表“热”, 表示原变量取值为对应类别; 取值0则代表“冷”, 表示原变量取值并非对应类别.

例5.2. 定性变量Outlook含有3个取值: Overcast, Rainy和Sunny. 设有3个样本观测, 其取值列于表5.2第2列中. 对Outlook作独热编码, 可以获得3个虚拟变量(见表5.2第3-5列), 其中每一个虚拟变量的名称是Outlook的一个取值. 第1个样本观测取值为Overcast, 因此, 它在变量Overcast的取值为1, 在其余两个虚拟变量的取值为0. 另外两个样本观测的取值可类似获得.

注意到, 若虚拟变量的数量等于定性变量的取值个数, 则对于每一个样本观测, 这些虚拟变量之和都等于1, 也就是说虚拟变量具有共线性. 对于以变量的线性组合为基础的模型, 例如普通最小二乘回归模型和Logtistic回归模

表 5.2: 独热编码示例

序号	Outlook	Overcast	Rainy	Sunny
1	Overcast	1	0	0
2	Rainy	0	1	0
3	Sunny	0	0	1

型, 共线性会导致模型不稳定、误差增大. 为避免共线性, 应该去掉其中任意一个虚拟变量[1].

但是, 在不涉及共线性的场合, 不能轻易删除独热编码中的虚拟变量. 例如, 在k近邻等算法中, 需要计算样本观测之间的距离. 对于以定类变量刻画距离的情况, 通常认为取值相同的样本观测距离为0, 取值不同的样本观测之间的距离大于0且为恒定值. 下面的例子将探讨独热编码中虚拟变量的数量对距离运算的影响.

例5.3. 采用明氏距离

$$d(\boldsymbol{a},\boldsymbol{b}) = \left(\sum_{i=1}^{p} |a_i - b_i|^m\right)^{1/m},$$

其中$\boldsymbol{a}$和$\boldsymbol{b}$都是p维向量, $m(>0)$是明氏距离中的参数. 若仅取表5.2中的两个虚拟变量, 例如取Overcast和Rainy, 则第1与第2个样本观测之间的距离为:

$$(|1-0|^m + |0-1|^m)^{1/m} = 2^{1/m}.$$

第1与第3个样本观测之间的距离为:

$$(|1-0|^m + |0-0|^m)^{1/m} = 1.$$

表5.2中的3个样本观测取值都不同, 但两两之间的距离应该相同. 由于去掉了一个虚拟变量, 导致距离值有所差异, 这不符合我们对距离运算的要求. 若保留所有的三个虚拟变量, 则每两个样本观测之间的距离都为1, 符合要求.

由例5.3可以看出, 当涉及明氏距离等运算时, 独热编码中的所有虚拟变量都应保留. 实际上, 对于(广义) 线性模型以外的大部分模型, 都应该保留完整的虚拟变量.

注记5.4. 虽然独热编码使得定性变量以量化的形式参与分析, 但不代表它与定量变量具有完全相同的处理方式.

[1]这被称为虚拟编码(dummy coding), 参见Zheng & Casari (2018) 第5章

(1) 直观地看, 虚拟变量的取值不连续, 可能无法应用数学运算的结果, 例如, 以均值插补缺失值.

(2) 由于一个定性变量所衍生的各虚拟变量具有一定的关联性(在一个样本观测中, 有且仅有一个虚拟变量取值为1), 独立地处理各虚拟变量可能导致问题. 例如, 在缺失值处理的k近邻插补方法中, 基于k个近邻插补各维度变量(不考虑变量之间的关联性), 可能会出现违背逻辑的插补值, 例如, 不满足虚拟变量间的关联性的插补值.

独热编码的一个缺陷是增大了变量的维度, 为数据分析方法的运行和解释带来了困难. 当模型或算法具备处理定性变量的能力时, 应避免使用独热编码[2].

5.2.2 有序编码

一部分模型或算法, 例如LightGBM算法和4.3节中的SMOTENC算法, 具备处理定性变量的能力. 但是, 它们在软件实现中往往要求以数值形式表达定性变量, 此时需要对以字符串形式表达的定性变量作变换, 转换为数值形式. 例如, 以取值$0, 1, 2, \cdots, C-1$表达含有C个类别的定性变量, 这就是有序编码(ordinal encoding).

若将有序编码应用于定类变量, 我们仅使用其数值形式, 忽略数值的大小顺序; 若将有序编码应用于定序变量, 可以使用数值的顺序信息. 有序编码的优点是不增加变量的维度, 并且可以保留定序变量中取值的顺序关系.

注记5.5. 在软件实现中, 一般需要在模型函数中声明定性变量, 否则这些经有序编码后的定性变量会被作为定量变量处理[3].

[2]此时, 使用独热编码可能会改变模型或算法的结果, 例如决策树CART算法(参见附录1).

[3]一般, 函数中设置了参数, 用以声明定性变量, 例如, 第4章SMOTENC函数中的categorical_features和第9章MissForest函数中的cat_vars. LightGBM则提供两种方式: (1) 将定性变量设置为'category'类型; (2)在模型的fit函数中由参数categorical_feature声明定性变量.

本章小结

(1) 思维导图

- 规范化思维导图
  - 1 定量变量的规范化
    - 标准化
    - 最小值-最大值规范化
    - 幂变换(Box-Cox变换, Yeo-Johnson变换)
    - 白化
    - 行归一化
  - 2 定性变量的规范化
    - 独热编码
    - 有序编码

(3) Python实现

表5.3列出了数据规范化相关的Python函数. 在scikit-learn中，我们没有找到实现白化的函数，可自行编程实现。但在需要白化处理的方法中，如主成分分析，有白化处理的参数whiten

习题

可使用scikit-learn提供的Normalizer函数实现，它也提供使用L1范数和最大值进行归一化。

CART模型构建中, 定性变量的虚拟变量取值个数为?

References

[1] Yeo I.K., Johnson R.A. (2000) A new family of power transformations to improve normality or symmetry. Biometrika, 87(4): 954-959.

表 5.3: 数据规范化Python函数

方法	函数名称与所属模块	重要参数	应用提示
标准化	StandardScaler (scikit-learn库preprocessing)	(1) with_mean: 是否中心化. 默认值为True, 即进行中心化 (2) with_std: 是否尺度归一化. 默认值为True, 即除以标准差, 实现尺度归一化	-
标准化	RobustScaler (scikit-learn库preprocessing)	(1) with_centering: 是否中心化. 默认值为True, 以中位数为中心进行中心化 (2) quantile_range: 尺度计算的范围. 为二元向量, 默认值为(25,75), 即以25%分位数作为下限, 75%分位数作为上限 (3) with_scaling: 是否进行尺度变换. 默认值为True, 以设置的quantile_range作为尺度进行变换	-
最小值-最大值规范化	MinMaxScaler (scikit-learn库preprocessing)	(1) feature_range: 规范化后的取值范围, 默认值为$(0,1)$, 对应区间$[0,1]$ (2) clip: 是否对规范化后的数值实施截尾. 默认值为False, 即不作截尾; 若设置为True, 可实现截尾, 即保证变换后的数据落入设置的取值范围	-
行归一化	Normalizer (scikit-learn库preprocessing)	norm: 所采用的归一化范数. 默认值为'l2', 即以L_2范数归一化; 还可以设置为'l1'和'max', 分别对应以L_1范数和最大的绝对值归一化	函数不允许缺失值, 需处理缺失值
幂变换	PowerTransformer (scikit-learn库preprocessing)	(1) method: 所采用的变换方法. 默认值为'yeo-johnson', 即采用Yeo-Johnson变换. 另一取值为'box-cox', 即采用Box-Cox变换. (2) standardize: 是否进行标准化. 默认值为True, 输出幂变换结果的标准化值(均值为0, 方差为1). 若取值为False, 则输出结果为幂变换后的数值(3) copy: 是否产生复制数据. 默认值为True, 产生并展示复制的数据集(含幂变换后的数值), 但不会改变原有数据集. 若取值为False, 则不产生复制数据集, 将原始数据集中的变量值替换为幂变换后的数值	-
独热编码	OneHotEncoder (scikit-learn库preprocessing)	(1) drop: 指定删除的变量. 默认值为None, 即不删除任何虚拟变量. 为避免共线性问题, 最简单的方法设置为'first', 表示删除每个定性变量变换得到的第1个虚拟变量. 也可以通过数组, 指定删除某一些虚拟变量. (2) handle_unknown: 当出现训练集中不存在的类别(称为未知类别) 时的处理方式. 默认值为'error', 即给出报错信息; 若取值为'ignore', 表明忽略该问题, 此时未知类别对应的虚拟变量取值均为0 (3) sparse: 返回值的类型. 默认值为True, 即返回结果为稀疏矩阵, 若要将其转换为数据框类型, 需要在返回值后加上.toarray(). 若设置为False, 则返回结果为数组类型	函数允许缺失值, 会将缺失值看作一个类别, 增加一个虚拟变量
顺序编码	OrdinalEncoder (scikit-learn库preprocessing)	(1) encoded_missing_value: 对缺失数据赋值. 默认值为np.nan, 即原始缺失数据经编码后仍为缺失值; 若取值为某个整数值, 则以该值为缺失数据赋值 (2) handle_unknown: 当出现未知类别时的处理方式. 默认值为'error', 即给出报错信息; 若取值为'use_encoded_value', 则按照参数unknown_value中设置的取值对未知类别赋值	-

[2] Box G.E.P., Cox D.R. An analysis of transformations. Journal of the Royal Statistical Society B, 26: 211-252.

[3] Zheng A., Casari A. (2018) Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol: O'Reilly Media, Inc..

[4] Ranzato M., Krizhevsky A., Hinton G.E. (2010) Factored 3-way restricted boltzmann machines for modeling natural images. Journal of Machine Learning Research, 9: 621-628.