Data Cleaning

数 据 嗨 客

代码中的数据变量为：

df | 一个DataFrame对象.　　dy | 一个ndarray对象.

s | 一个Series对象.　　x | 数据样本

导入第三方库：

```
import pandas as pd   import numpy as np
import sklearn.svm as svm
import scipy.stats as stats
import statsmodels.api as sm
import sklearn.ensemble as ens
import sklearn.neighbors as neigh
import sklearn.preprocessing as prep
import sklearn.linear_model as lm
import imblearn.over_sampling as os
import imblearn.under_sampling as ius
from matplotlib import pyplot as plt
```

基本操作

np.median(dy) | 获取dy的中位数.

(dy/df).min()/max() | 获取dy/df的最值.

(dy/df).mean() | 获取dy/df的均值.

(dy/df).std() | 获取dy/df的标准差.

stats.mode(dy) | 获取dy的众数.

df['name'].map(str.upper/lower/title)
把df的特征name的取值转换为大写/小写/首字母大写.

s.value_counts() | 统计s中各个取值的频数，默认忽略缺失值，通过参数dropna进行调节.

df.info() | 查看df的详细信息，包含是否含有空值，取值类型，占用空间等.

df.describe() | 查看df各个特征的统计信息，包含均值，方差，最值等，自动忽略离散型特征.

正态性检验

sm.qqplot(x) | 绘制分位数-分位数图.

stats.skewtest(x) | 进行偏度检验.

stats.kurtosistest(x) | 进行峰度检验.

stats.normaltest(x) | 进行正态性检验.

stats.kstest(x) | 进行KS检验.

stats.shapiro(x) | 进行W检验.

非正态性检验

np.log(x) | 进行自然对数变换.

np.log10(x) | 进行以10为底的对数变换.

stats.boxcox(x) | 进行box-cox变换.

数据去重

df.duplicated() | 返回一个布尔型对象，用来检测重复的行或列.

df.drop_duplicated() | 返回删除重复行或者列的DataFrame对象.

数据整合

pd.merge() | merge函数通过一个或多个索引 将数据集的行连接起来,键值可以为列标签或索引,类似于数据库的JOIN操作； 该函数的主要应用场景是针对同一个主键存在两张包含不同特征的表， 通过该主键的连接， 将两张表进行合并.

pd.concat() | pandas库以及它的Series和DataFrame等数据结构实现了带编号的轴， concat()函数可以沿着一条轴将多个对象堆叠到一起.

df.combine_first() | 调用该方法组合两个DataFrame对象，适用于索引全部或部分重叠的两个数据集，合并后的索引和列是两个对象的并集.

数据组合

df['person'].groupby(df['score']) | 对标签score分组，根据sum()函数对特征person进行聚合，返回一个groupby对象 .

df.agg(custom_func) | 使用自定义聚合函数custom_func对df进行分组.

数据不均衡

os.SMOTE() | SMOTE过采样.

os.ADASYN() | 基于自适应合成方法的过采样.

os.RandomOverSampler() | 随机过采样，通过随机添加少类样本达到数据平衡.

ius.RandomUnderSampler() | 随机欠采样，通过随机删除多类样本达到数据平衡.

ius.ClusterCentroids()|基于聚类的欠采样方法.

缺失值表示

NA | R语言中缺失值的表示方式.

na | Matlab中缺失值的表示方式.

None | Python语言中缺失值(空值)的表示方式，其类型为object.

np.nan/np.NaN | Numpy和Pandas中缺失值的表示方式，其类型为浮点(float)型.

缺失值处理

df.isnull() | 返回一个由布尔值组成的对象，判断哪些数据元素是缺失值，True表示为缺失值.

df.notnull() | 返回一个由布尔值组成的对象，判断哪些数据元素不是缺失值，True表示不是缺失值.

df.dropna() | 根据选定的轴标签axis，删除含有缺失值的行或者列，可通过how参数调节删除数据的范围.

df.fillna() | 通过method属性(如ffill或bfill)填充缺失数据；也可以通过value指定填充的值或者字典对象，如零值，均值，众数，中位数等.

异常值检测

neigh.LocalOutierFactor() | 使用局部异常因子算法进行异常值检测，给x的每一样本计算局部离群分数，要求scikit-learn版本在0.19以上.

svm.OneClassSVM() | 使用one class SVM算法进行异常值检测，其属于无监督，基于密度的异常值检测方法，只需输入数据样本x即可.

ens.IsolationForest() | 使用孤立森林算法进行异常值检测，当多颗决策树共同为特定样本产生较短的路径长度时，该样本极可能是异常值，要求scikit-learn版本在0.19以上.

x[abs(x-x.mean())>3*np.std(x)] | 使用拉依达准则检测x中的异常值.

plt.boxplot(x) | 通过绘制盒图对x进行可视化，检测异常值.

数据类型转换

df.dtypes() | 查看df各个特征的类型，常见类型有object，int，float和bool型.

df.get_dytpe_counts() | 统计df各个特征的类型数量.

df['age'].astype('float') | 把df的age特征的类型改为float型.

特征编码

prep.Binarizer(threshold=num) | 特征二值化编码，特征取值大于num编码为1，小于num取值为0.

prep.OnehotEncoder() | One-hot特征编码，输入变量必须是2维ndarry或者DataFrame.

prep.LabelEncoder() | 对数据样本的标签特征进行数字编码，编码后的标签取值范围为[0, nclass -1].

prep.LabelBinarizer() | 对数据样本的标签特征进行二值化编码，编码后的标签取值为{0, 1}，默认正类标签为1，负类标签为0.

pd.get_dummies(x) | 使用Pandas模块对x进行One-hot特征编码.

特征标准化

prep.MinMaxScaler() | Min-Max标准化.

prep.StandardScaler() | Z-score标准化.

特征离散化

pd.cut(x, bins) | 对x进行离散化，子区间端点集合为数组bins.

pd.cut(x, num) | 对x进行等距离散化，子区间个数为num.

pd.qcut(x, 4) | 按照四分位数进行离散化.