定量分析和定性分析

对比分析 → 两个互相联系的指标进行比较

绝对数比较（相减） / 相对数比较（相除）

结构分析、比例分析、空间比较分析、动态对比分析

统计分析

统计指标对定量数据进行统计描述，常从集中趋势和离中趋势两个方面进行分析

集中趋势度量 / 离中趋势度量

帕累托分析（贡献度分析）

相关性分析：图示初判 / Pearson相关系数（皮尔逊相关系数） / Sperman秩相关系数（斯皮尔曼相关系数）

# 缺失值插补# 几种思路：均值/中位数/众数插补、临近值插补、插值法ffill临近值插补,前值插值法 —— 拉格朗日插值法print(lagrange(x,y))

异常值分析（1）3σ原则：如果数据服从正态分布，异常值被定义为一组测定值中与平均值的偏差超过3倍的值 → p(|x - μ| > 3σ) ≤ 0.003数据标准化（1）0-1标准化将数据的最大最小值记录下来，并通过Max-Min作为基数（即Min=0，Max=1）进行数据的归一化处理 x = (x - Min) / (Max - Min) （2）Z-score标准化 Z分数（z-score）,是一个分数与平均数的差再除以标准差的过程 → z=(x-μ)/σ，其中x为某一具体分数，μ为平均数，σ为标准差 Z值的量代表着原始分数和母体平均值之间的距离，是以标准差为单位计算。在原始分数低于平均值时Z则为负数，反之则为正数 数据连续属性离散化 等宽法 → 将数据均匀划分成n等份，每份的间距相等； cut方法 bins = [18,25,35,60,100]cats = pd.cut(ages,bins)print(cats)等频法 → 以相同数量的记录放进每个区间； qcut方法

**因子分析(factor analysis)**

是指研究从变量群中提取共性因子的统计技术。 因子分析是简化、分析高维数据的一种统计方法。

因子分析又存在两个方向，一个是探索性因子分析（exploratory factor analysis）。另一个是验证性因子分析（confirmatory factor analysis）。

**探索性因子分析**是先不假定一堆自变量背后到底有几个因子以及关系，而是我们通过这个方法去寻找因子及关系。

**验证性因子分析**是假设一堆自变量背后有几个因子，试图验证这种假设是否正确。

因子分析有两个核心问题，一是如何构造因子变量，二是如何对因子变量进行命名解释。

**因子分析的一般步骤**

将原始数据标准化处理 X

计算相关矩阵C

计算相关矩阵C的特征值 r 和特征向量 U

确定公共因子个数k

构造初始因子载荷矩阵,其中U为r的特征向量

建立因子模型

对初始因子载荷矩阵A进行旋转变换，旋转变换是使初始因子载荷矩阵结构简化，关系明确，使得因子变量更具有可解释性，如果初始因子不相关，可以用方差极大正交旋转，如果初始因子间相关，可以用斜交旋转，进过旋转后得到比较理想的新的因子载荷矩阵A'.

将因子表示成变量的线性组合，其中的系数可以通过最小二乘法得到.

计算因子得分

eig\_value,eig\_vector=nlg.eig(C) #计算特征值和特征向量eig=pd.DataFrame() #利用变量名和特征值建立一个数据框eig['names']=data.columns#列名eig['eig\_value']=eig\_value#特征值if eig['eig\_value'][:k].sum()/eig['eig\_value'].sum()>=0.8: #如果解释度达到80%解释度：建模h=np.zeros(10) #变量共同度，反映变量对共同因子的依赖程度，越接近1，说明公共因子解释程度越高，因子分析效果越好D=np.mat(np.eye(10))#特殊因子方差，因子的方差贡献度 ，反映公共因子对变量的贡献，衡量公共因子的相对重要性A=np.mat(A) #将因子载荷阵A矩阵化

过拟合(overfitting)：创建的模型与训练数据过于匹配，以致于模型无法根据新数据做出正确的预测。

欠拟合(underfitting)：创建的模型与训练数据匹配度较差。

K值较小，则模型复杂度较高，容易发生过拟合，学习的估计误差会增大，预测结果对近邻的实例点非常敏感。

K值较大，则模型复杂度较低，可以减少学习的估计误差，但是学习的近似误差会增大，与输入实例较远的训练实例也会对预测起作用，使预测发生错误。