山东工商学院
Shandong Technology and Business University

数学建模中的数据处理与数据分析 (下)

邓富文

计算机科学与技术学院

July 27, 2025

目录

探索性数据分析（EDA）概述

Definition

探索性数据分析（Exploratory Data Analysis, EDA）是数据分析流程中的一个关键步骤，其目的是在正式建模之前，通过各种统计图表和描述性统计量，深入了解数据的特征、结构、潜在模式、异常值以及变量之间的关系。它是一种开放式的、以好奇心驱动的分析哲学。

EDA 的核心目标：

- 发现数据中隐藏的规律和趋势；
- 识别数据质量问题（如异常值、缺失值）；
- 验证或挑战我们对数据的初步假设；
- 为后续的特征工程和模型选择提供坚实的依据和灵感；
- 更好地理解业务问题，将数据与现实世界联系起来。

描述性统计：概念 I

Definition

描述性统计（Descriptive Statistics）是用于总结、组织和描述数据集主要特征的统计方法。它不涉及对数据背后总体的推断，而是专注于对现有数据的概括，是 EDA 的基石。

描述性统计：概念 II

集中趋势度量 (Measures of Central Tendency)

用于描述数据集中数据的中心位置或典型值。

- **均值 (Mean)**: 所有数据点之和除以数据点的数量。最常用，但对异常值敏感。

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **中位数 (Median)**: 将数据按大小排序后，位于最中间位置的值。对异常值不敏感，能更好地反映普通数据的水平。
- **众数 (Mode)**: 数据集中出现频率最高的值。可用于任何类型的数据。

描述性统计：概念 III

离散程度度量 (Measures of Dispersion)

用于描述数据集中数据的分散程度或变异性。

- **方差 (Variance)**: 衡量数据点与均值之间平均偏离程度的平方。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \text{(总体)} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{(样本)}$$

- **标准差 (Standard Deviation)**: 方差的平方根，与原始数据单位相同，更易解释。
- **极差 (Range)**: 最大值与最小值之差。简单，但极易受异常值影响。
- **四分位距 (IQR)**: Q_3 (75% 分位) 与 Q_1 (25% 分位) 之差，即 $IQR = Q_3 - Q_1$。衡量中间 50% 数据的分散程度，对异常值稳健。

描述性统计：分布形态度量 I

偏度 (Skewness)

衡量数据分布的对称性。

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

- 偏度 $= 0$：数据近似对称分布（如正态分布）。
- 偏度 > 0：右偏（正偏），长尾在右，均值通常大于中位数。
- 偏度 < 0：左偏（负偏），长尾在左，均值通常小于中位数。

描述性统计：分布形态度量 II

峰度 (Kurtosis)

衡量数据分布的“尖峭”程度或尾部的厚度。通常我们使用**超峰度**（Excess kurtosis = Kurtosis - 3）来与正态分布比较。

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

- 超峰度 $= 0$：正态分布。
- 超峰度 > 0：尖峰（leptokurtic），数据集中在均值附近，尾部较厚（意味着存在更多极端值）。
- 超峰度 < 0：平峰（platykurtic），数据分布较平坦，尾部较薄（意味着极端值较少）。

描述性统计：MATLAB 实现

代码参考

对 `sample_data.csv` 文件中 `Age` 和 `Score` 列计算上述所有描述性统计量的具体实现，请查阅代码文件：**`descriptive_statistics.m`**

可视化方法 (1/2): 单变量与双变量 I

单变量可视化

用于探索单个变量的分布特征。

- **直方图 (Histogram)**：显示数值型数据的分布情况。
- **密度图 (Density Plot)**：直方图的平滑版本，更清晰地展示分布形态。
- **条形图 (Bar Chart)**：用于显示分类变量的频率或计数。
- **饼图 (Pie Chart)**：显示分类变量中每个类别所占的比例，适用于类别较少的情况。
- **箱线图 (Box Plot)**：直观地展示数值型数据的五数概括及异常值，是识别异常值的利器。

代码参考: `vis_univariate.m`

可视化方法 (1/2): 单变量与双变量 II

双变量可视化

用于探索两个变量之间的关系。

- **散点图 (Scatter Plot)**：显示两个数值型变量之间的关系，是判断相关性的首选。
- **折线图 (Line Plot)**：主要用于显示时间序列数据中一个变量随时间的变化趋势。
- **分组箱线图**：显示数值型变量在不同分类变量组中的分布情况，是比较组间差异的有效手段。
- **分组/堆叠条形图**：显示两个分类变量之间的关系。

代码参考: `vis_bivariate.m`

可视化方法 (2/2): 多变量

多变量可视化

用于探索三个或更多变量之间的关系。

- **散点图矩阵 (Scatter Plot Matrix)**：在一个视图中展示数据集中所有数值变量两两之间的散点图，快速发现变量间的总体关系。
- **热力图 (Heatmap)**：常用于可视化相关系数矩阵，用颜色深浅表示相关性强度，一目了然。
- **平行坐标图 (Parallel Coordinates Plot)**：每个变量对应一条垂直轴，每个数据点表示为连接这些轴上相应值的折线，适用于高维数据的模式发现和异常检测。

代码参考

具体实现请查阅代码文件：`vis_multivariate.m`

相关性分析 (1/2): 数值变量 I

相关性分析是研究变量之间是否存在统计关系以及关系强度和方向的方法。(重要提醒：相关性不等于因果性。)

Pearson 相关系数

衡量两个连续变量之间**线性关系**的强度和方向。取值范围 [-1,1]。

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

相关性分析 (1/2): 数值变量 II

Spearman 秩相关系数

衡量两个变量之间**单调关系**（不一定是线性的）的强度和方向。它基于变量的秩次计算，对异常值不敏感。

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Kendall 秩相关系数

另一种衡量变量之间秩次一致性的非参数统计量，适用于非线性单调关系。

$$\tau = \frac{\text{一致对数} - \text{不一致对数}}{\text{总对数}} = \frac{C - D}{\frac{1}{2}n(n-1)}$$

相关性分析 (1/2): 数值变量 III

代码参考

具体实现请查阅代码文件：`correlation_numerical.m`

相关性分析 (2/2): 分类变量

对于分类变量，我们通常通过列联表和卡方检验来评估它们之间的关联性。

- **列联表 (Contingency Table)**：也称为交叉表，用于汇总两个或多个分类变量的频数分布。
- **卡方检验 (χ^2 Test for Independence)**：用于检验两个分类变量之间是否存在显著关联。它的原假设是两个变量相互独立。如果 p 值小于显著性水平（如 0.05），则认为变量间存在关联。
- **Cramer's V**：衡量两个分类变量关联强度的指标，基于卡方值计算，取值范围 [0,1]。0 表示无关联，1 表示完全关联。

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1,\, c-1)}}$$

代码参考

具体实现请查阅代码文件：`correlation_categorical.m`

数据驱动的建模方法概述 I

回归分析 (Regression Analysis)

- **目的**: 探究变量间的因果关系，进行数值预测。
- **核心思想**: 寻找最优的曲线/曲面来拟合数据点，使得所有数据点到该线的 “总误差” 最小（最小二乘法）。
- **关键指标**: 回归系数, R^2, p 值, 残差分析。

聚类分析 (Clustering Analysis)

- **目的**: 发现数据的内在结构，进行客户分群、图像分割等。
- **核心思想**: 根据样本间的 “距离” 或 “相似度” 进行分组（无监督学习）。
- **常用算法**: K-Means 聚类, 层次聚类。

数据驱动的建模方法概述 II

分类分析 (Classification Analysis)

- **目的**: 进行类别预测。
- **核心思想**: 从已带标签的数据集学习，构建模型预测新样本类别（有监督学习）。
- **常用算法**: 决策树, 支持向量机 (SVM), 逻辑回归, K 近邻 (KNN)。

时间序列分析 (Time Series Analysis)

- **目的**: 进行时间相关的预测。
- **核心概念**: 趋势 (Trend), 季节性 (Seasonality), 平稳性 (Stationarity)。
- **常用算法**: 移动平均, 指数平滑, ARIMA 模型。

引言：维度灾难与多属性决策

维度灾难

当面对含有众多变量（指标）的数据集时，我们常常会遇到“维度灾难”：

- 分析变得极其复杂
- 计算量指数级增长
- 变量之间高度相关，信息冗余

降维是解决这一问题的关键。

多属性决策

数学建模的核心任务之一就是对多个备选方案进行优劣排序。这需要一套科学的评价体系来综合多个（往往是冲突的）评价指标。**综合评价模型**为此提供了系统化的解决方案。

主成分分析 (PCA) I

核心思想与目标

PCA 是最常用的线性降维方法。其思想是将多个相关的原始变量，通过线性变换，重组成少数几个**互不相关**的综合变量，即 “主成分”。PCA 的目标是让第一个主成分尽可能多地解释原始数据的总方差，后续主成分则依次解释剩余方差。

处理步骤

1. **数据标准化**：对原始数据矩阵 X 进行 Z-score 标准化，得到 Z。
2. **计算协方差矩阵**：$R = \frac{1}{m-1} Z^T Z$。
3. **计算特征值和特征向量**：求解协方差矩阵 R 的特征值 λ_k 和特征向量 $\mathbf{v}_k$。
4. **选择主成分**：计算各主成分的方差贡献率 $\eta_k = \lambda_k / \sum \lambda_i$ 和累计贡献率 $C_p = \sum_{k=1}^{p} \eta_k$。通常选择累计贡献率超过 85% 的前 p 个主成分。
5. **计算主成分得分**：第 i 个样本的得分为 $F_i = Z_i \cdot [\mathbf{v}_1, \ldots, \mathbf{v}_p]$。

案例：城市综合经济实力评价 (PCA) I

问题背景与数据

收集 10 个城市关于 5 个经济指标的数据，希望通过 PCA 将这 5 个指标压缩为少数几个综合经济指数，并对城市进行排名。

Table 1: 城市经济指标原始数据

城市	GDP(X1)	人均收入 (X2)	工业产值 (X3)	消费总额 (X4)	财政收入 (X5)
A	30000	75000	28000	12000	3500
B	25000	68000	23000	10000	3000
C	22000	65000	20000	9000	2800
D	18000	58000	17000	7500	2200
E	15000	52000	14000	6000	1800
F	12000	48000	11000	5000	1500
G	9000	42000	8000	3500	1100
H	7000	38000	6500	2800	900
I	5000	32000	4500	2000	600
J	3000	28000	2500	1200	400

案例：城市综合经济实力评价 (PCA) II

代码参考

具体实现请查阅代码文件：`pca_city.m`

因子分析 (Factor Analysis) I

核心思想与区别

因子分析与 PCA 关系密切，但出发点不同。PCA 旨在概括总方差，而因子分析旨在寻找支配原始变量的、少数不可观测的**潜在因子** (Latent Factors)，并用这些因子来解释原始变量之间的**相关关系**。它更侧重于探索变量背后的内在结构。

处理步骤

1. **因子模型**：$X_i = a_{i1}F_1 + \cdots + a_{ik}F_k + \epsilon_i$
   - F_j: 公共因子, a_{ij}: 因子载荷, ϵ_i: 特殊因子
2. **因子载荷矩阵估计**：使用主成分法或最大似然法。
3. **因子旋转**：如方差最大化旋转 (Varimax)，使因子更易解释，这是因子分析的关键步骤。
4. **因子得分计算**：计算每个样本在公共因子上的得分。

案例：课程满意度调查 (因子分析)

问题背景

对学生进行课程满意度调查，包含 6 个问题（变量）：课程内容 (X1)、教师教学 (X2)、教材质量 (X3)、作业难度 (X4)、考核方式 (X5)、学习收获 (X6)。希望通过因子分析，找出影响满意度的几个核心潜在因子。

代码参考

具体实现请查阅代码文件：`factor_analysis.m` (注：此为示意文件名)

层次分析法 (AHP) I

核心思想与适用场景

AHP 是一种将复杂的决策问题分解为目标、准则、方案等层次，并通过定性与定量相结合的方式进行系统化分析的决策方法。它特别适用于那些难以完全用定量指标衡量、依赖专家主观判断的决策问题。

处理步骤

1. **建立层次结构模型**：目标层、准则层、方案层。
2. **构造判断矩阵**：使用 1-9 标度法进行两两比较。
3. **计算权重与一致性检验**：
   - 计算判断矩阵最大特征值 $\lambda_{\max}$ 及对应特征向量（归一化后即为权重 W）。
   - 计算一致性比例 $CR = \frac{CI}{RI}$，其中 $CI = \frac{\lambda_{\max} - n}{n-1}$。
   - 若 $CR < 0.1$，则通过一致性检验，判断有效。
4. **计算组合权重**：逐层计算，得到最终得分并排序。

案例：AHP 辅助旅游目的地决策

问题背景与判断矩阵

小明计划去旅游，备选地有杭州、厦门、成都。他主要考虑三个因素：景色、花费、交通。他根据自己的偏好给出了以下判断矩阵：

准则层 (A)

A	景色	花费	交通
景色	1	2	4
花费	1/2	1	3
交通	1/4	1/3	1

对景色 (B1)

B1	杭	厦	成
杭	1	2	4
厦	1/2	1	3
成	1/4	1/3	1

对花费 (B2)

B2	杭	厦	成
杭	1	1/3	1/5
厦	3	1	1/2
成	5	2	1

对交通 (B3)

B3	杭	厦	成
杭	1	3	5
厦	1/3	1	2
成	1/5	1/2	1

代码参考

具体实现请查阅代码文件：**ahp.m**

TOPSIS 法（优劣解距离法） I

核心思想与适用场景

TOPSIS 是一种基于 “理想解” 的排序方法，适用于具有多个量化指标的多方案评价问题。其核心思想是，一个最优的方案应该离 “正理想解”（所有指标都达到最优值的虚拟方案）最近，同时离 “负理想解”（所有指标都达到最劣值的虚拟方案）最远。

TOPSIS 法（优劣解距离法） II

处理步骤

1. **数据正向化**：将所有指标统一为“效益型”。
2. **数据标准化**：向量归一化 $Z_{ij} = X_{ij} / \sqrt{\sum_{i=1}^{m} X_{ij}^2}$。
3. **计算加权标准化矩阵**：$V_{ij} = w_j \cdot Z_{ij}$ (权重可由 AHP 或熵权法确定)。
4. **确定正、负理想解**：$V_j^+ = \max_i(V_{ij})$, $V_j^- = \min_i(V_{ij})$。
5. **计算距离**：$D_i^+ = \sqrt{\sum_{j=1}^{n}(V_{ij} - V_j^+)^2}$, $D_i^- = \sqrt{\sum_{j=1}^{n}(V_{ij} - V_j^-)^2}$。
6. **计算相对贴近度**：$C_i = \frac{D_i^-}{D_i^+ + D_i^-}$。$C_i$ 越接近 1，表示方案越优。

案例：TOPSIS 选择供应商

问题背景与数据

一家公司需要从 4 家供应商中选择一家。评价指标包括：价格（成本型）、质量（效益型）、供货周期（成本型）。专家打分得到权重为 $W = [0.4, 0.3, 0.3]$。

Table 2: 供应商评价原始数据

供应商	价格 (X1)	质量 (X2)	供货周期 (X3)
S1	10.5	98.5	15
S2	12.0	99.2	12
S3	11.2	97.8	10
S4	10.8	98.8	18

代码参考

具体实现请查阅代码文件：`topsis.m`

熵权法 (Entropy Weight Method) I

核心思想与适用场景

当评价模型中指标权重难以主观确定时，熵权法提供了一种完全基于数据本身的客观赋权方法。其基本思想是：指标的信息熵越小，说明该指标值的变异程度越大，提供的信息量就越多，在综合评价中应占有更高的权重。

熵权法 (Entropy Weight Method) II

处理步骤

1. **数据标准化**：对原始矩阵 X 进行归一化，以消除量纲影响。
2. **计算信息熵**：
   - 计算比重：$p_{ij} = X_{ij} / \sum_{i=1}^{m} X_{ij}$。
   - 计算信息熵：$e_j = -\frac{1}{\ln(m)} \sum_{i=1}^{m} p_{ij} \ln(p_{ij})$。
3. **计算权重**：
   - 计算信息熵冗余度 (差异性)：$d_j = 1 - e_j$。
   - 计算权重：$w_j = d_j / \sum_{j=1}^{n} d_j$。

代码参考

沿用供应商评价案例，使用熵权法计算权重。
具体实现请查阅代码文件：`entropy_weight.m`

练习 2.1：探索性数据分析（EDA） I

背景

现有一份包含二手车信息的样本数据集（包括品牌、车龄、里程数、价格等），需要通过描述性统计和数据可视化手段，探索影响二手车价格的关键因素。

任务要求

1. **描述性统计分析**: 计算 Age, Mileage, Price 三个变量的集中趋势、离散程度和分布形态度量。
2. **单变量可视化**: 绘制 Price 的直方图和箱线图；绘制 Brand 的条形图。
3. **双变量可视化**: 绘制 Age 与 Price 的散点图；绘制按 Brand 分组的 Price 的箱线图。
4. **多变量可视化**: 创建 Age, Mileage, Price 的散点图矩阵和相关系数热力图。

练习 2.1：探索性数据分析（EDA） II

提示

在分析价格时，注意观察其分布是否为正偏（右偏）。在许多经济学模型中，对价格等正值变量取对数后再进行分析，可以使数据更接近正态分布，有助于后续建模。

练习 2.2：评价模型综合应用 I

背景

某区域发展研究中心希望对区域内的 8 个主要城市进行综合发展水平的评估，他们收集了每个城市关于经济、社会和环境三个维度的六项关键指标。由于专家无法就各项指标的权重达成一致，请你建立一个客观、科学的评价模型。

练习 2.2：评价模型综合应用 II

任务要求

1. **客观赋权 (熵权法)**：使用熵权法计算六项指标的客观权重。
2. **综合评价 (TOPSIS 法)**：利用熵权法得到的权重，使用 TOPSIS 法对 8 个城市进行排名。
3. **降维分析 (主成分分析)**：
   - 使用 PCA 对数据降维，确定主成分数量（如累计贡献率 >85%）。
   - 分析主成分的实际意义（例如，经济发展因子、环境宜居因子等）。
   - 以各主成分的方差贡献率为权重，计算每个城市的 PCA 综合得分，并进行排名。
4. **结果对比与分析**：对比 TOPSIS 法与 PCA 的排名结果，分析异同及原因，讨论哪种方法在本问题中更具说服力。

练习 2.2：评价模型综合应用 III

提示

对比分析时，可以思考：PCA 是基于方差最大化的数据重构，而 TOPSIS 是基于与理想解的距离。哪种方法的内在逻辑更符合“综合评价”的目标？PCA 的综合得分是否容易解释？TOPSIS 的结果是否更直观？

感谢观看！