第三章 数据预处理

目录
CONTENTS

3.1 数据预处理的必要性

3.2 数据清理

3.3 数据集成

3.4 数据归约

3.5 数据变换与离散化

3.6 本章小结

Chapter 3.1

数据预处理的必要性

数据预处理的必要性

现实世界中的数据

- 现实世界中数据常常是**不完整的**（缺少感兴趣的或需要的属性），**含有噪声的**（离群点或偏离期望的观测值），**不一致的**（相同语义的属性采取不同的命名）脏数据，无法直接进行数据挖掘，或挖掘结果差强人意。
- 例：
  - 不完整：在用户调查中，某些参与者未填写年龄或收入等关键信息；
  - 含噪声：在传感器数据中，偶尔会记录到极端温度值（如-100°C）；
  - 不一致：在不同数据库中，“客户ID”在一个数据库中被称为“Customer_ID”，而在另一个数据库中则称为“CustID”。

数据预处理的必要性

对数据质量提出的要求

- 通常从一下几个方面评估数据的质量：

- **准确性**：数据应该能准确反映所描述的事实或对象，不应包含错误或失真。医疗场景中患者生命体征数据需要准确，以确保正确的诊断和治疗。
- **完整性**：数据应包含所有必要的信息，没有遗漏。医疗场景中要有药品的出场日期、成分表、不适用人群等完整信息。
- **一致性**：描述相同对象或概念的数据应保持一致。教育场景中学生的年龄应与当前年份与其出生年份的差值一致。

数据预处理的必要性

对数据质量提出的要求

- 通常从一下几个方面评估数据的质量：
  - **合时性**：数据应该被及时更新以反映当前的情况。金融场景中股票市场数据需要实时更新，以便投资者做出及时的决策。
  - **可信性**：数据应该来自可信的来源并且经过验证。新闻报道中信息的来源应是可信的、权威的，以确保信息的可信度。
  - **可解释性**：数据应该能够被理解和解释，以便进行分析和决策。医疗场景中医生向患者解释医疗诊断时应使用易于理解的描述，以确保患者能够理解并参与决策。
- 如果我们需要的数据能够满足我们的需求，那么它就是高质量的！

数据预处理的必要性

- 数据与挖掘结果的关系
- 低质量的数据导致低质量的挖掘结果
- 高质量的挖掘结果依赖高质量的数据

数据预处理的必要性

数据预处理的主要任务

- **数据清理**（Data cleaning）
  - 包括处理缺失值、平滑噪声、识别和删除异常点或离群值，旨在提高数据质量
- **数据集成**（Data integration）
  - 属性冗余和数据对象重复问题，确保来自不同来源的数据能够提供一致的视图
- **数据归约**（Data reduction）
  - 通过简化数据集来减少数据量，从而提高处理效率，同时保留重要信息
- **数据变换**（Data transformation）
  - 将数据转换为适合数据挖掘的格式，可能包括标准化、归一化、离散化等，以便于后续分析

Chapter 3.2

数据清理

数据清理

数据清理

- 数据清理就是对数据进行重新审查和校验的过程，通过填充缺失值、光滑噪声并识别离群点、纠正数据中的不一致等方式提高数据的质量。
  - 缺失值的处理
  - 噪声的处理

数据清理

缺失值

- **缺失值**（Missing Value）是指数据集中的某个观测值或字段的数值或信息未被采集、获取或记录，从而在数据集中存在空缺或未知的值。
- 缺失值可能由如下原因产生：
  - 数据数据采集时发生错误或故障，以及人为遗漏；
  - 数据存储时丢失或损坏；
  - 隐私原因使得某些观测值被删除或修改以保护个人敏感信息等。
- 处理缺失值的方法主要分为删除和填充两种方式，通常取决于数据的性质、缺失值的原因以及数据挖掘的目标。

数据清理

缺失值的处理

- 忽略数据对象

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25	18	男	A市B小区
10002	31	24	女	A市C小区
10003	28		男	
…	…	…	…	…

- 若一个数据对象中有属性的观测值缺失，则将忽略该数据对象。
- 当数据对象缺失多个属性的观测值时常采用忽略数据对象的方法。如果存在缺失值的数据对象占比过高时，此方法会造成数据的浪费。

数据清理

缺失值的处理

- 忽略属性

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25		男	A市B小区
10002	31	24	女	A市C小区
10003	28		男	A市B小区
…	…	…	…	…

- 若一个属性的观测值大部分缺失，则忽略该属性。
- 一般情况下不采用此方法，因为可能难以判断忽略的属性的重要性。

数据清理

缺失值的处理

- 使用全局常量填充

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25	18	男	A市B小区
10002	31	24	女	A市C小区
10003	28	unknown	男	unknown
…	…	…	…	…

- 将缺失的观测值用同一个常量替换。
- 数据挖掘方法可能将这些全局常量归纳出一个知识，所以这种方法虽然简单但是并不可靠。

数据清理

- 缺失值的处理

- 使用属性的中心趋势度量填充

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25	18	男	A市B小区
10002	31	24	女	A市C小区
10003	28	21	男	A市B小区
…	…	…	…	…

- 可以根据属性的类型采用均值填充，或是中位数或众数填充。

数据清理

缺失值的处理

- 使用类别分组后属性的中心趋势度量填充

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址	顾客分级
10001	25	19	男	A市B小区	活跃
10002	31	24	女	A市C小区	潜在
10003	28	20	男	A市D小区	不活跃
…	…	…	…	…	…

- 选取与给定数据对象属于同一类别的所有数据对象，使用这些数据对象的属性的中心趋势度量填充给定数据对象的缺失值。

数据清理

缺失值的处理

- 使用可能的观测值填充

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25	18	男	A市B小区
10002	31	24	女	A市C小区
10003	28	21.4	男	A市B小区
…	…	…	…	…

- 使用预测模型生成缺失值，如可以利用回归、贝叶斯推理或决策树等方法，推断出该缺失值最大可能的取值。
- 这种方法面临“两难困境”：如果模型预测准确，则表明该缺失值对应的属性是冗余的；如果预测的不准确，则表明该缺失值对应的属性是独特的，但是不可靠。

数据清理

缺失值的处理

- 人工填充

顾客ID	年龄(岁)	年薪(万)	性别	家庭住址
10001	25	18	男	A市B小区
10002	31	24	女	A市C小区
10003	28	23	男	A市B小区
…	…	…	…	…

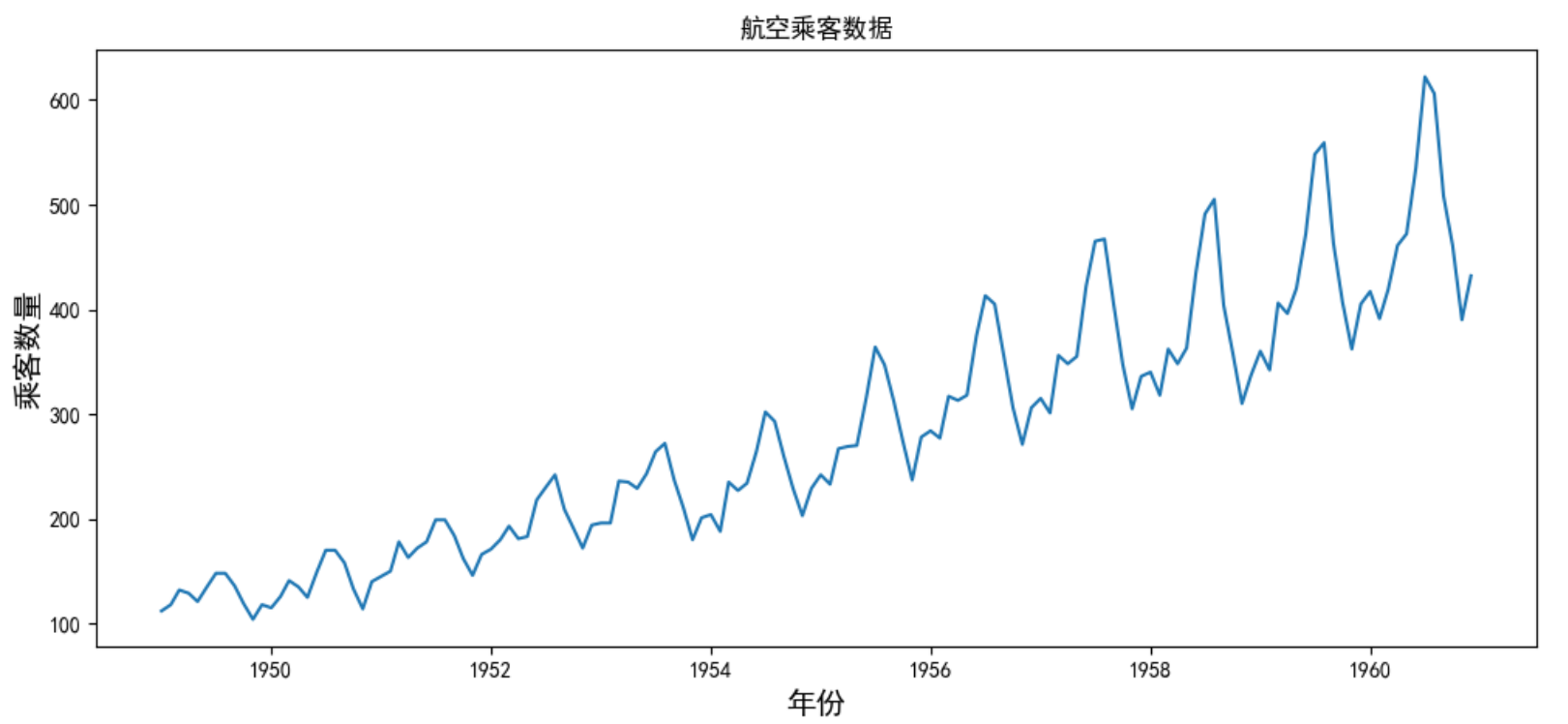
- 利用专业领域知识同时考虑其他属性和数据对象，对缺失值进行人工填写。
- 当数据集很大或缺失值较多时，该方法费时费力。

数据清理

- 缺失值的处理
  - 忽略数据对象
  - 忽略属性
  - 使用全局常量填充
  - 使用属性的中心趋势度量填充
  - 使用类别分组后属性的中心趋势度量填充
  - 使用可能的观测值填充
  - 人工填充

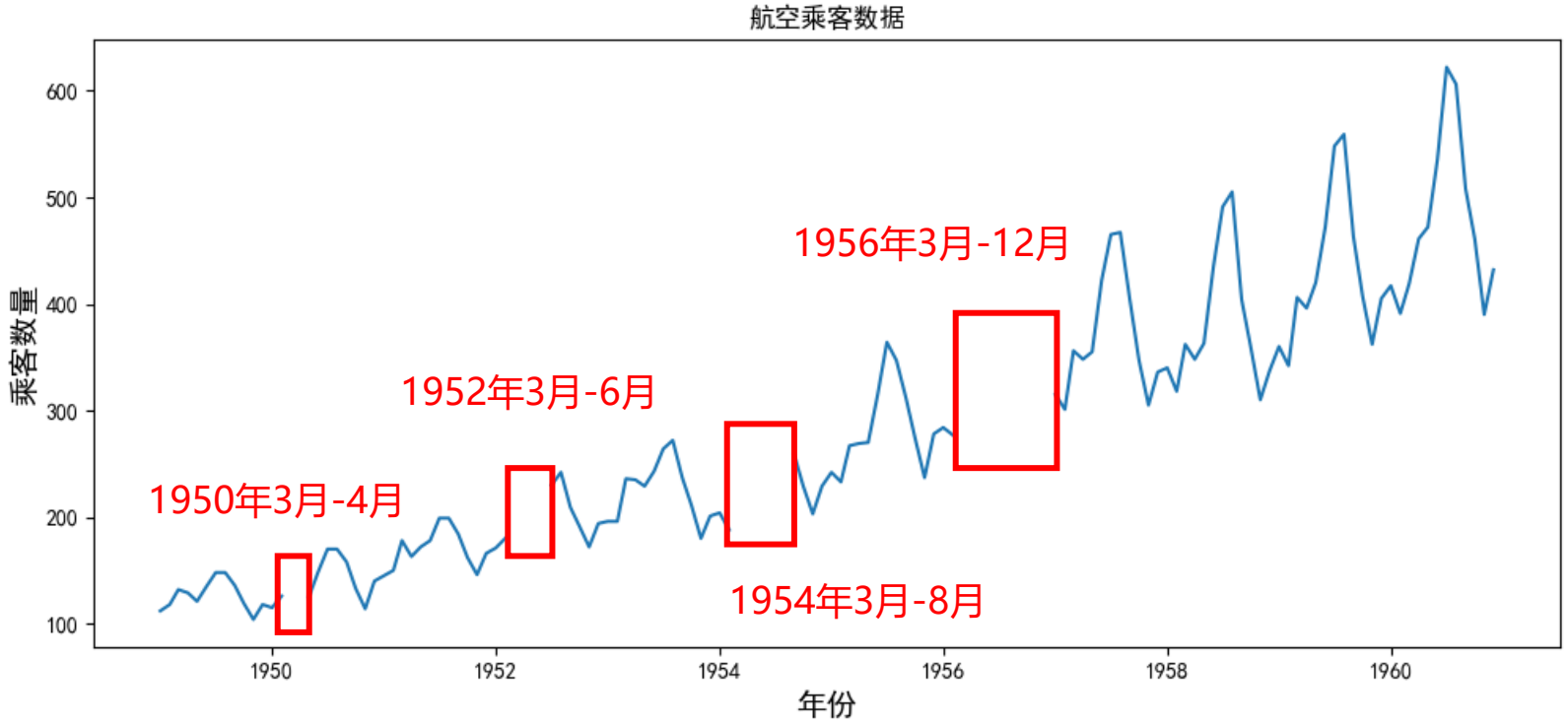
时序数据的缺失值处理

- Air Passengers数据集中描述了1949年1月至1960年12月期间每个月的国际航空乘客数量。



数据清理

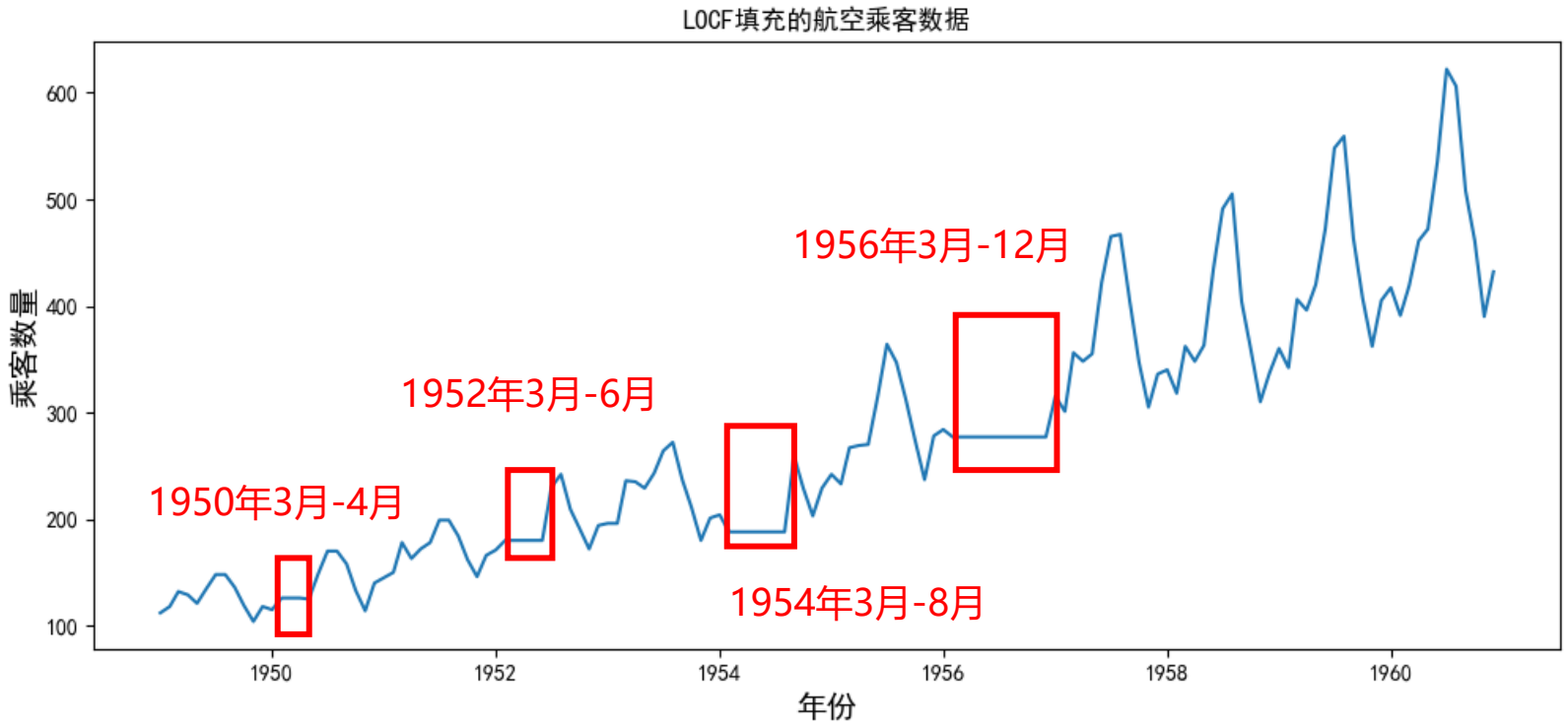
- 时序数据的缺失值处理
  - 带缺失值的Air Passengers数据集



时序数据的缺失值处理

- **前推法**（Last Observation Carried Forward，LOCF）
  - 如果在时间序列中某个时间点有缺失值，那么就将缺失值替换为缺失**之前**的**最后一个非缺失**观测值。
- 例：
  - 原始时序数据：[1, 2, None, None, 3, 3, 3, 4, None, None, 4, 4, 5, 5, 6, None, 7, 8, 9]
  - LOCF填充后的时序数据：[1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 8, 9]

- 时序数据的缺失值处理
  - 使用前推法填充缺失值

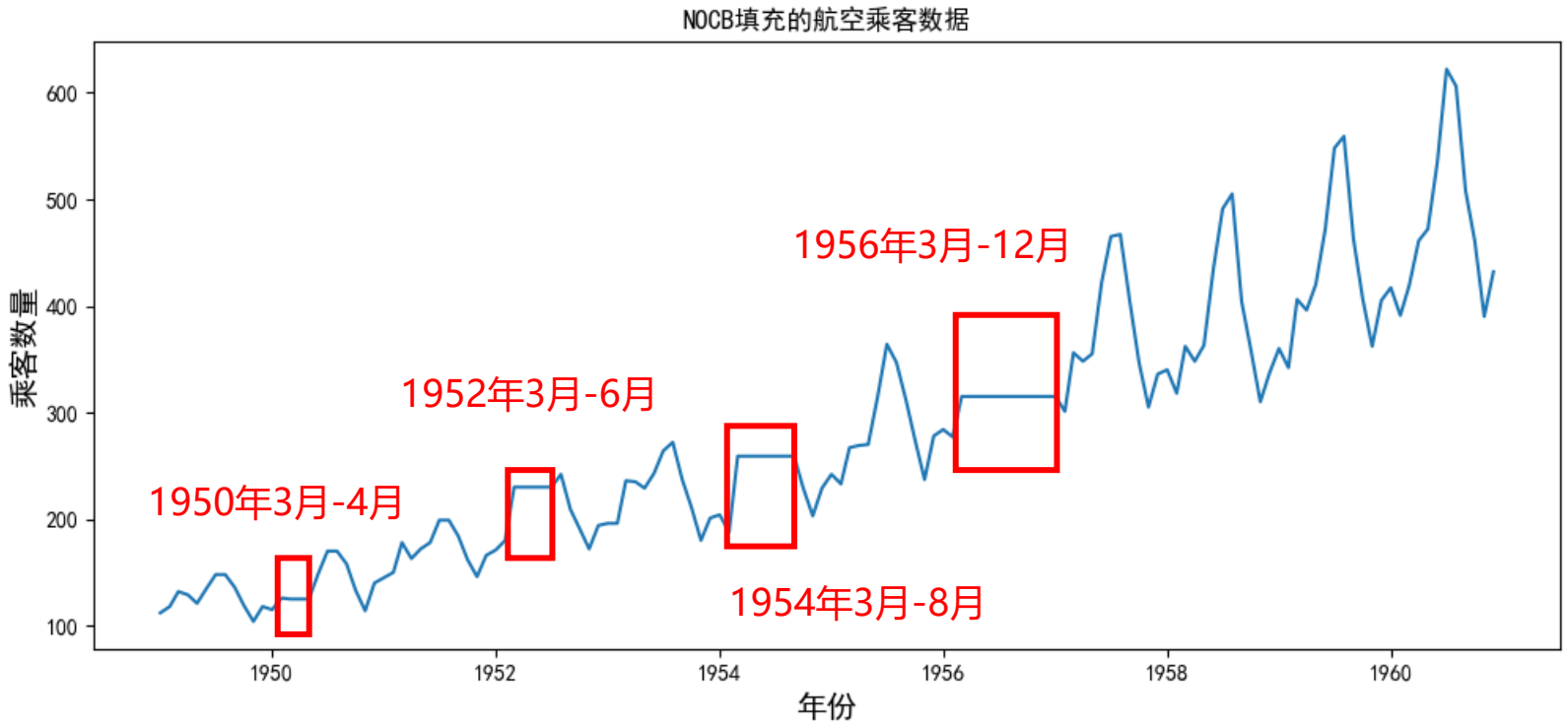


时序数据的缺失值处理

- **后推法**（Next Observation Carried Backward，NOCB）
  - 如果在时间序列中某个时间点有缺失值，那么就将缺失值替换为缺失**之后**的**第一个非缺失**观测值。
- 例：
  - 原始时序数据：[1, 2, None, None, 3, 3, 3, 4, None, None, 4, 4, 5, 5, 6, None, 7, 8, 9]
  - NOCB填充后的时序数据：[1, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 7, 7, 8, 9]

数据清理

- 时序数据的缺失值处理
  - 使用后推法填充缺失值



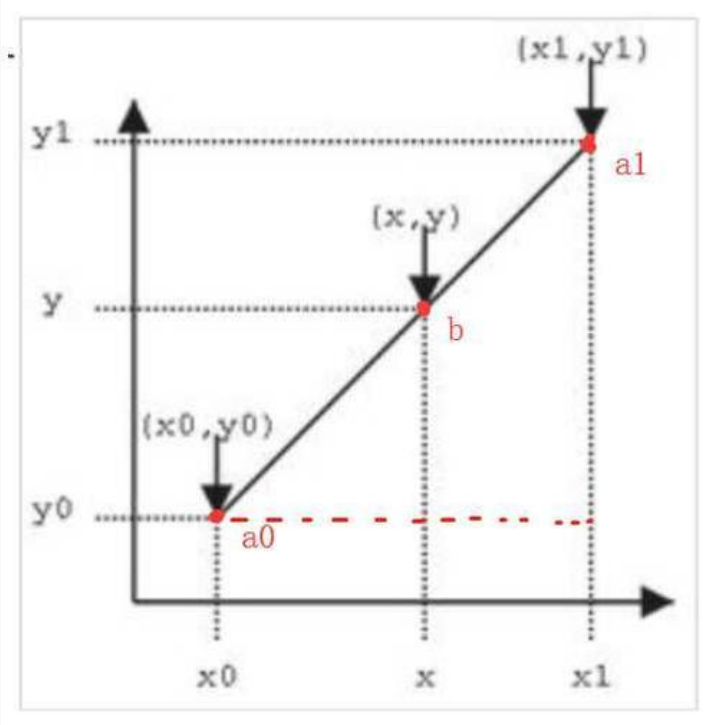
- 时序数据的缺失值处理
  - 前推法和后推法的优点
    - 简单，时间复杂度低；
    - 考虑到了时序数据的局部变化的特点；
    - 对于具有较稳定趋势的时序数据，可以保持趋势的连续性。
  - 前推法和后推法的缺点
    - 对于出现连续缺失值的情况，填补的值没有区别；
    - 对于变化剧烈或具有明显趋势的时序数据，可能导致不准确的填充。

数据清理

时序数据的缺失值处理

- **线性插值法**（linear interpolation）

- 如果在时间序列中某个时间点有缺失值，那么找到缺失值前后最近的两个已知观测值，然后计算它们之间的线性插值以填充该缺失值。



$\because$ a0，b，a1处于一条直线

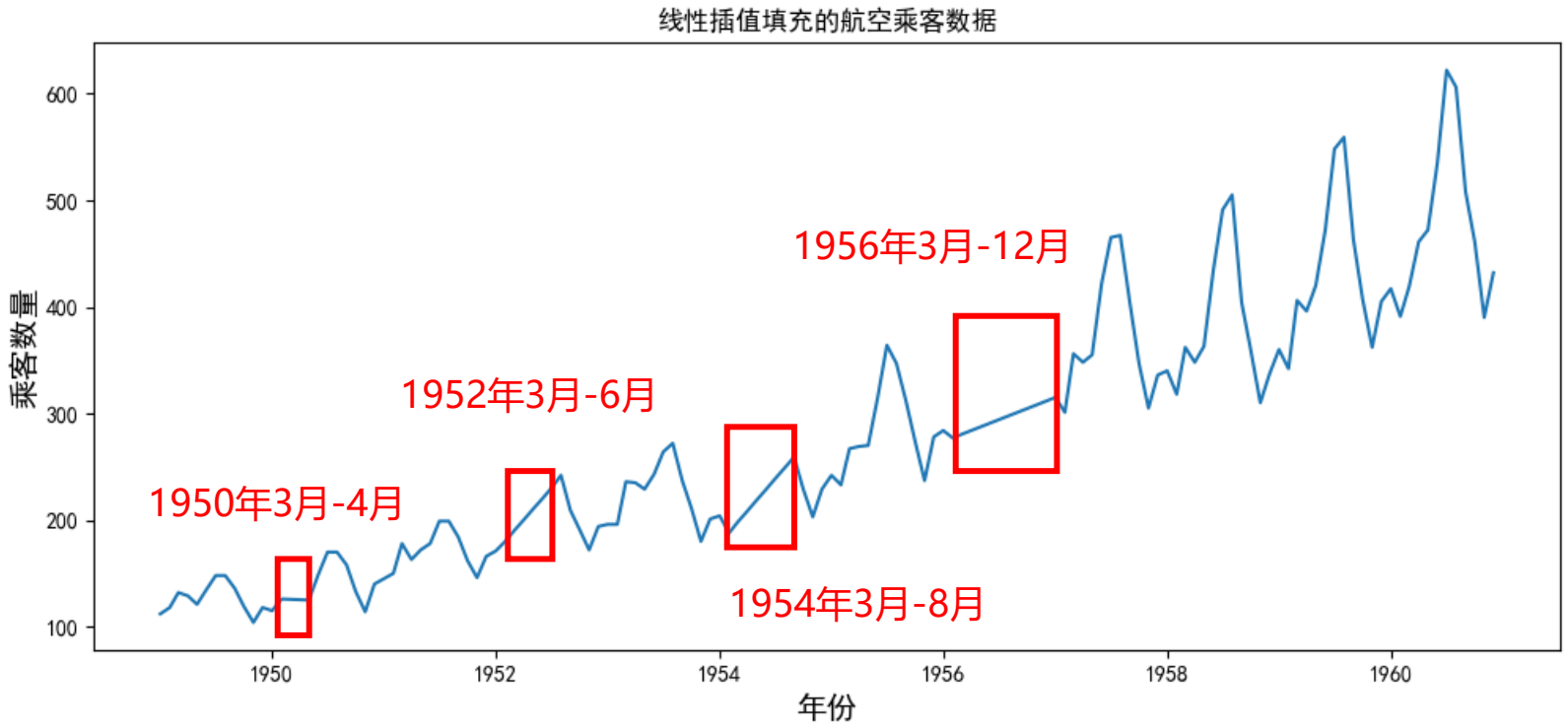
$$\therefore \frac{y-y_0}{x-x_0}=\frac{y_1-y_0}{x_1-x_0}$$

$$\therefore y=\frac{x_1-x}{x_1-x_0}y_0+\frac{x-x_0}{x_1-x_0}y_1$$

$$\therefore y=\alpha y_0+(1-\alpha)y_1$$

数据清理

- 时序数据的缺失值处理
  - 使用线性插值法填充缺失值



数据清理

时序数据的缺失值处理

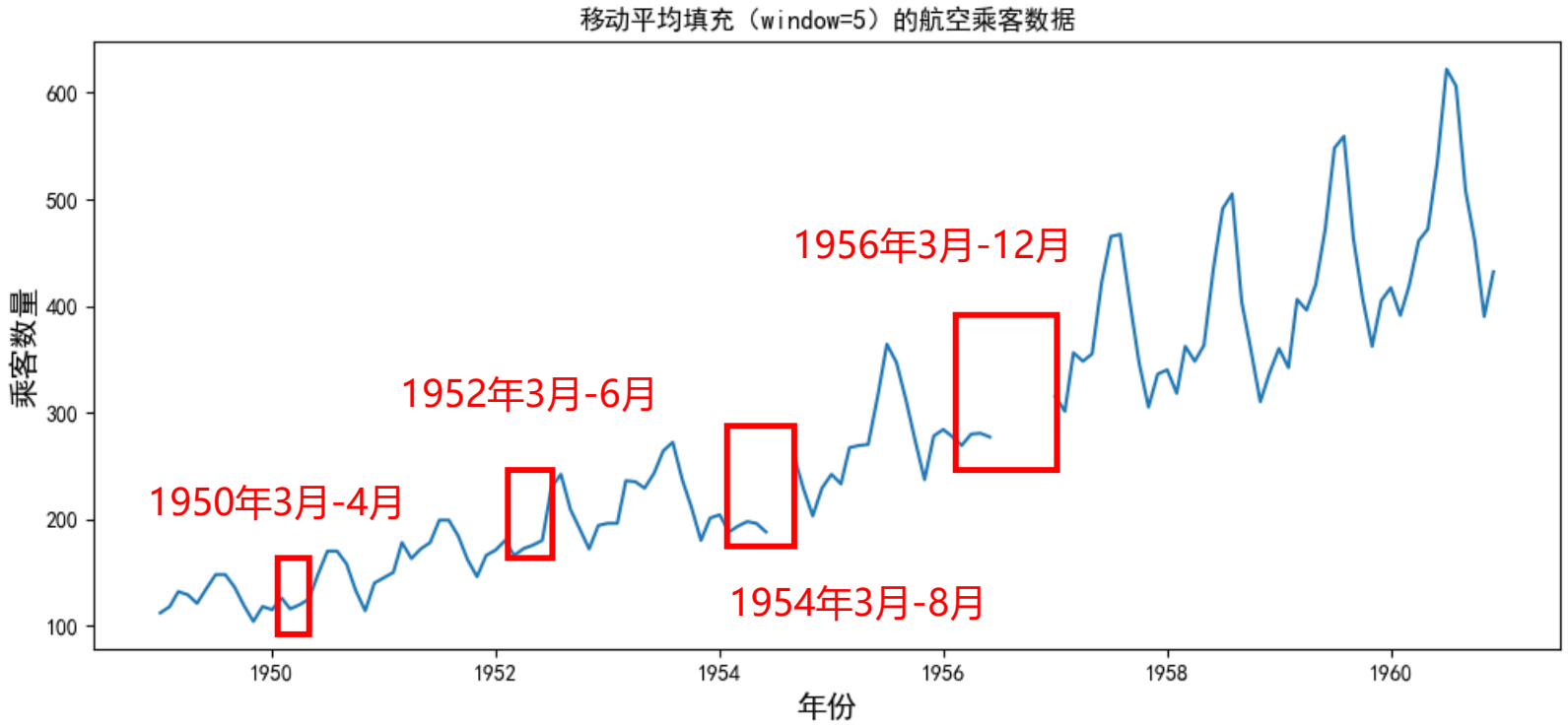
- 线性插值法的优点
  - 简单，时间复杂度低；
  - 考虑到了时序数据的局部变化的特点；
  - 可以保持稳定、上升或下降趋势的连续性。
- 线性插值法的缺点
  - 假设观测值之间的关系是线性的，这在某些情况下可能不合理；
  - 当数据的变化趋势不稳定或有较大波动时可能会产生不合理的结果，因为无法捕捉数据的非线性性质。

- 时序数据的缺失值处理
- **移动平均法**（Moving Average）
  - 移动平均法通过计算一定时间窗口内的观测值的平均值来填充缺失值。根据具体的需求，可以选择不同大小的窗口，或者对窗口内观测值采用不同的加权方式。

128	156		201	202	200		184	165

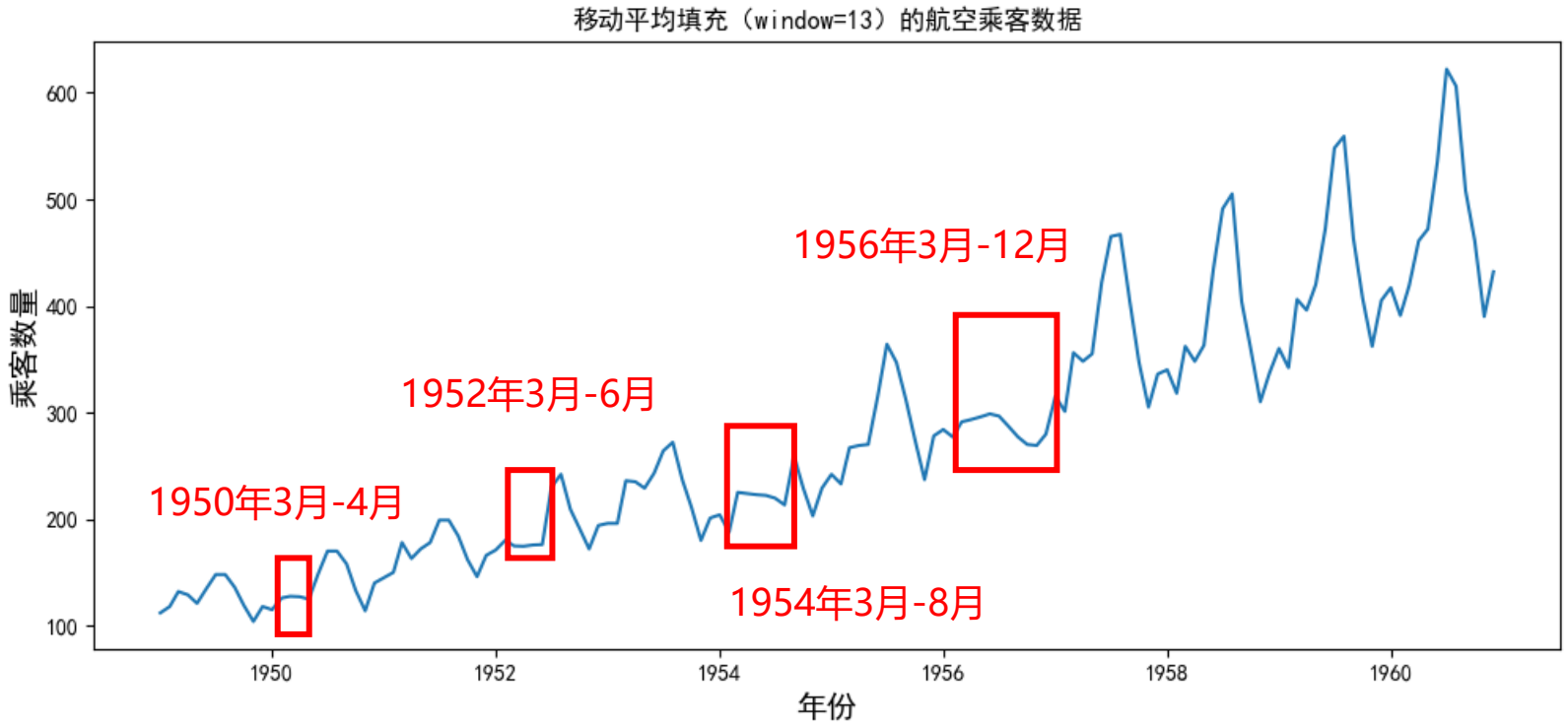
数据清理

- 时序数据的缺失值处理
  - 使用移动平均法填充缺失值



数据清理

- 时序数据的缺失值处理
  - 使用移动平均法填充缺失值



数据清理

- 时序数据的缺失值处理
  - 移动平均法的优点
    - 简单，时间复杂度低；
    - 滑动窗口的方式非常适用于时序数据。
  - 移动平均法的缺点
    - 难以选择合适的窗口尺寸；
    - 仅适用于数值属性。

噪声的定义

- 数据挖掘中，**噪声**（Noise）通常指的是数据中的随机或不相关的干扰、错误或错误的现象，它们不符合数据的真实模式或规律，可能导致数据挖掘结果的不准确性和误导性。

数据清理

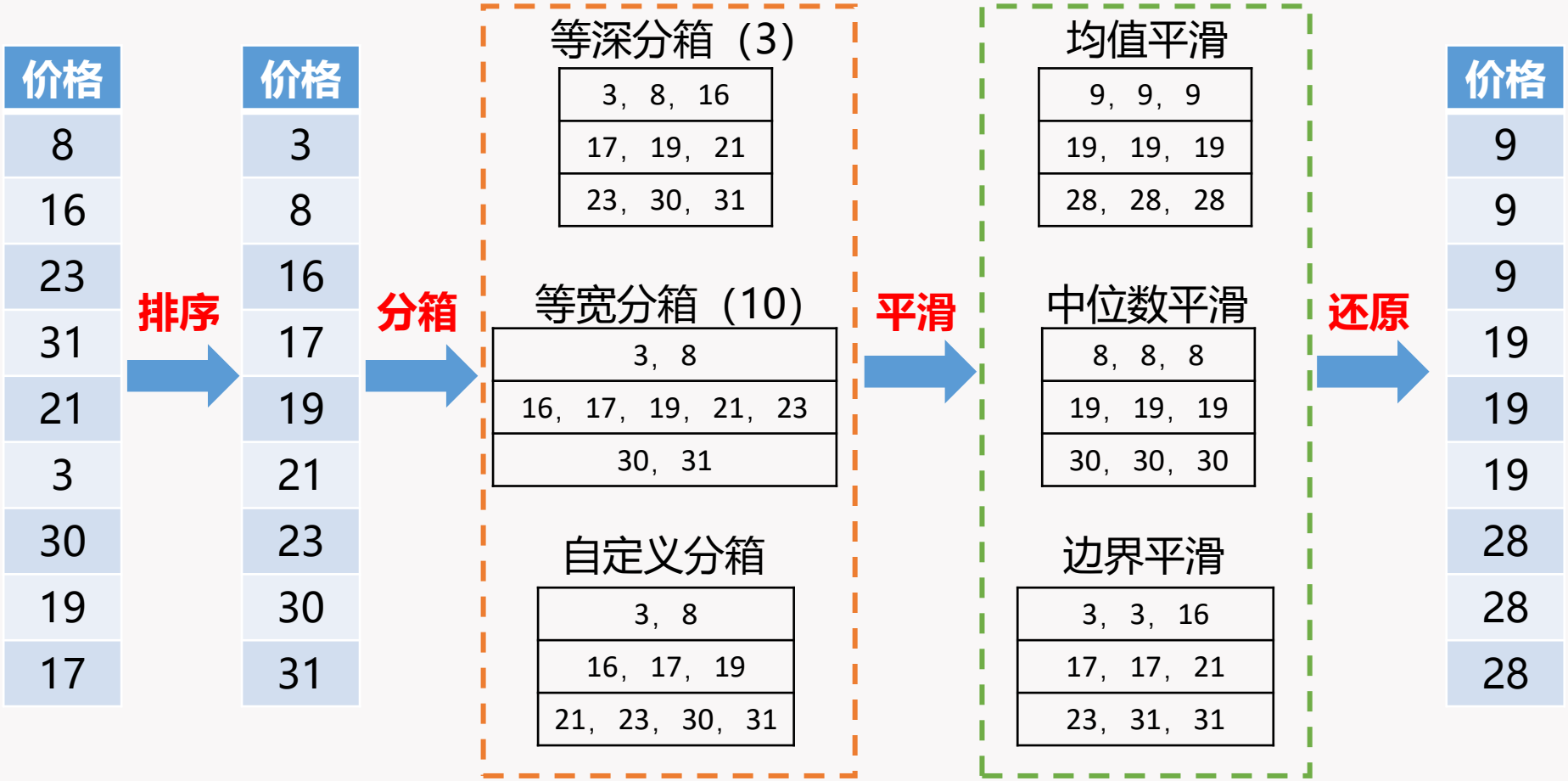
噪声的处理

- 回归分析：用一个函数拟合数据来光滑数据，描述两个属性之间的关系时使用线性回归，描述多个属性之间的关系时使用多元线性回归。
- 离群点分析：通过聚类、盒图等方法检测并删除离群数据。

数据清理

噪声的处理

- 分箱法（Binning）通过考察近邻的观测值实现局部平滑与去噪。

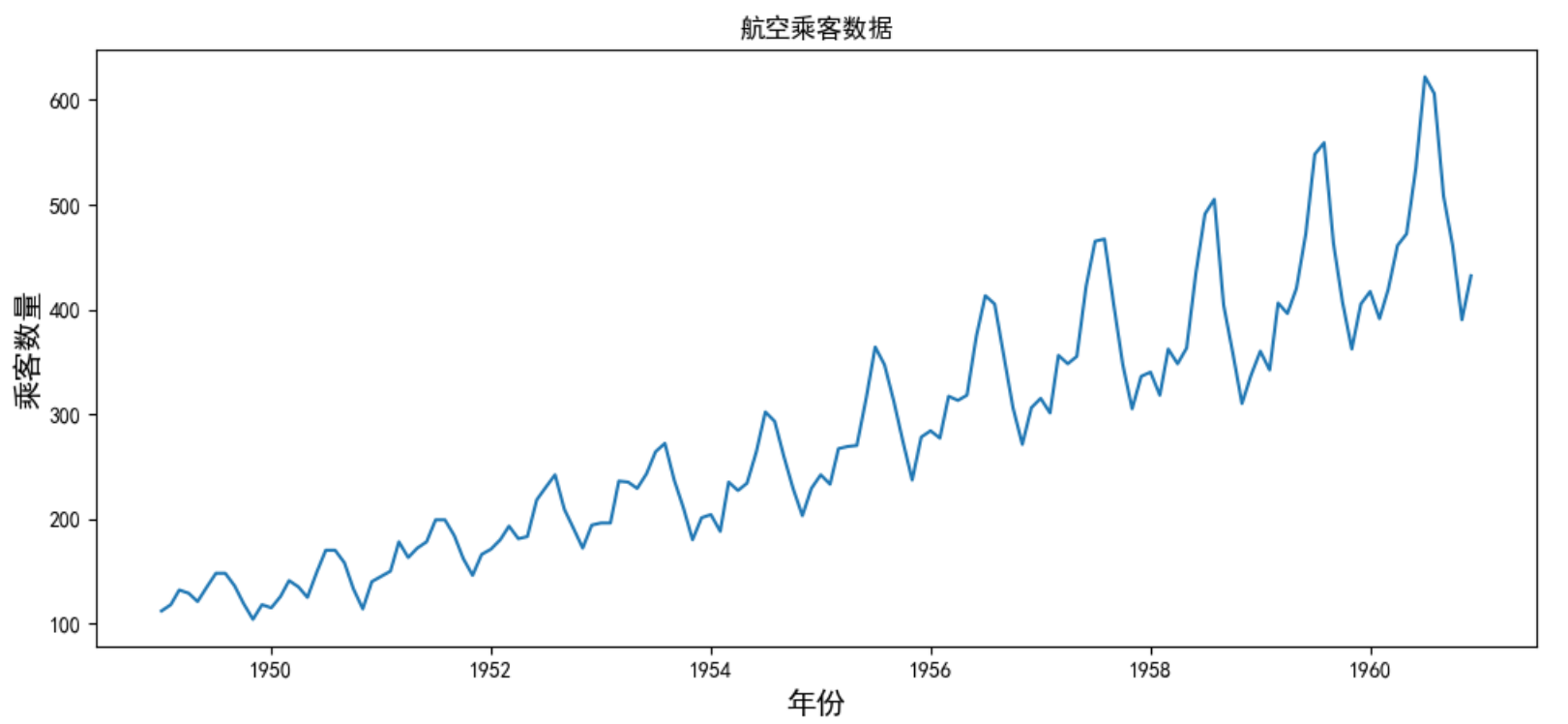


数据清理

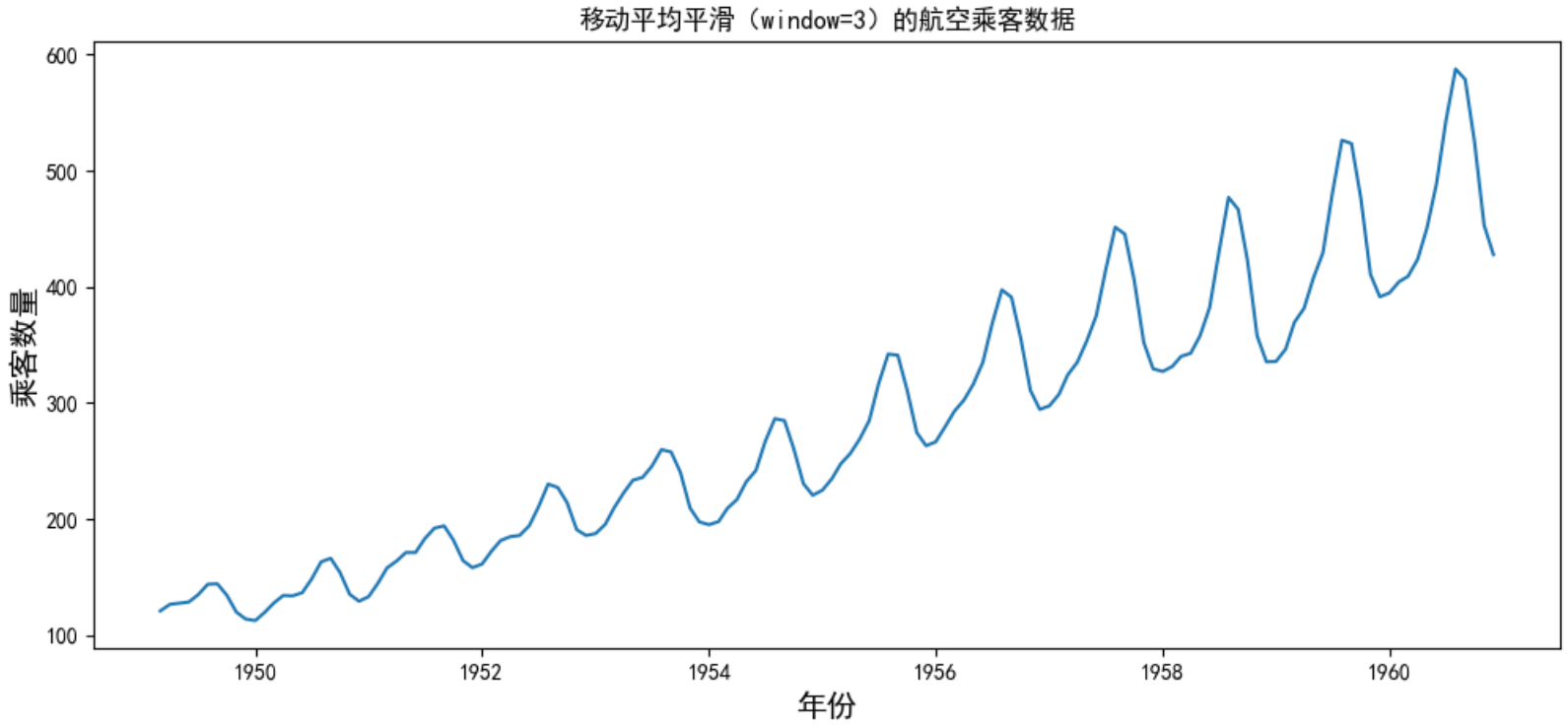
时序数据的噪声处理

- 移动平均法
  - 移动平均法利用一段时间内的平均值作为平滑值，因此可以一定程度上去除噪声的影响；
  - 移动平均法可以去除时序数据的短期波动，有助于识别趋势和周期性模式。

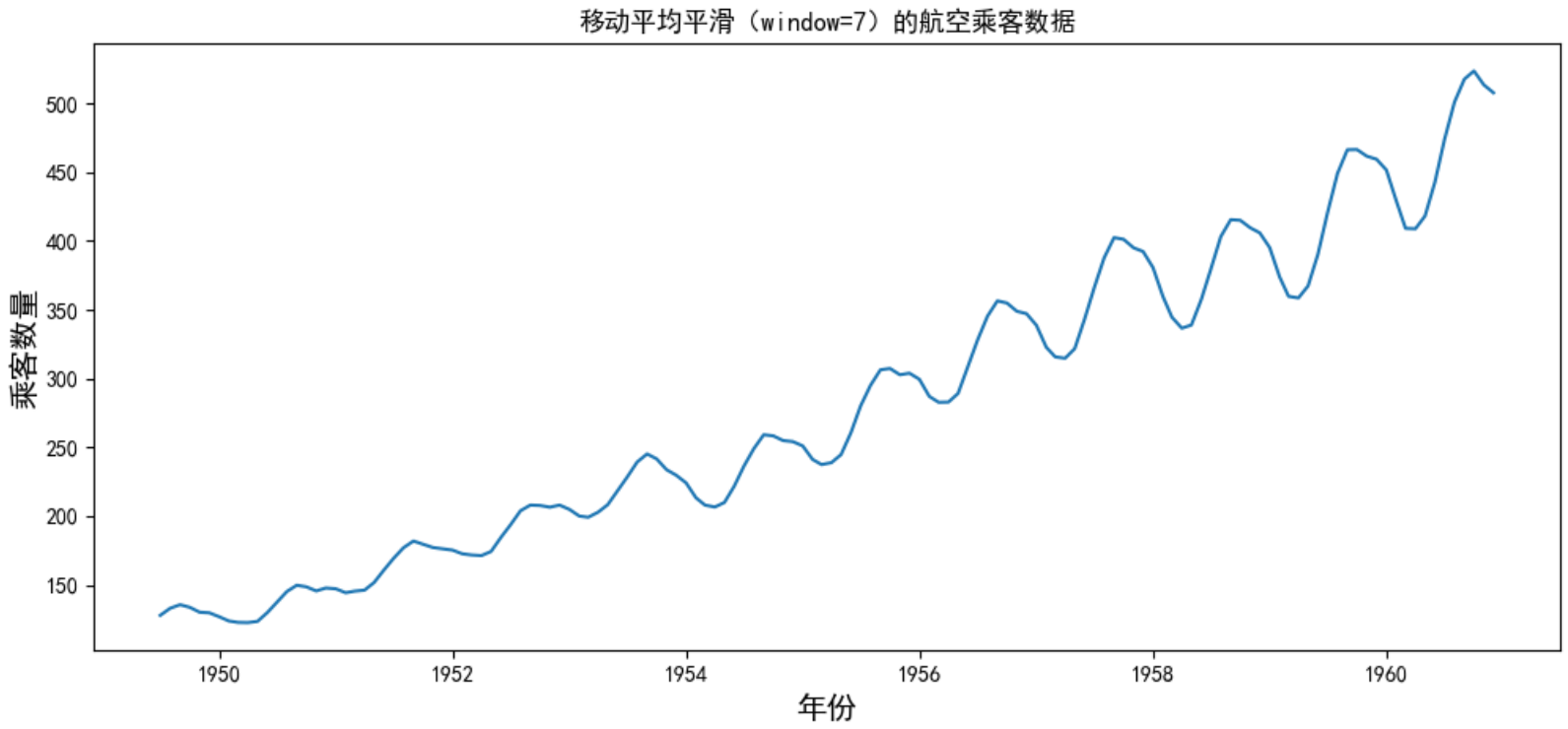
- 时序数据的噪声处理
  - 未平滑的Air Passengers数据集



- 时序数据的噪声处理
  - 移动平均平滑（window=3）的Air Passengers数据集



- 时序数据的噪声处理
  - 移动平均平滑（window=7）的Air Passengers数据集



数据清理

时序数据的噪声处理

- 指数滑动平均法
  - 移动平均法认为近期的观测值对当前平滑值具有相同的重要性，而非近期的观测值则不影响当前平滑值。
  - 时序顺序下观测值的影响应该随着时间间隔的增长而递减，因此指数平滑法利用时序顺序对观测值进行加权平均。

时序数据的噪声处理

- 指数滑动平均法
- 假设时序数据表示为 $x_1, x_2, \ldots$，α $(0 < \alpha < 1)$ 为加权系数，指数平滑公式为

$$S_t = \alpha x_t + (1-\alpha) S_{t-1}$$

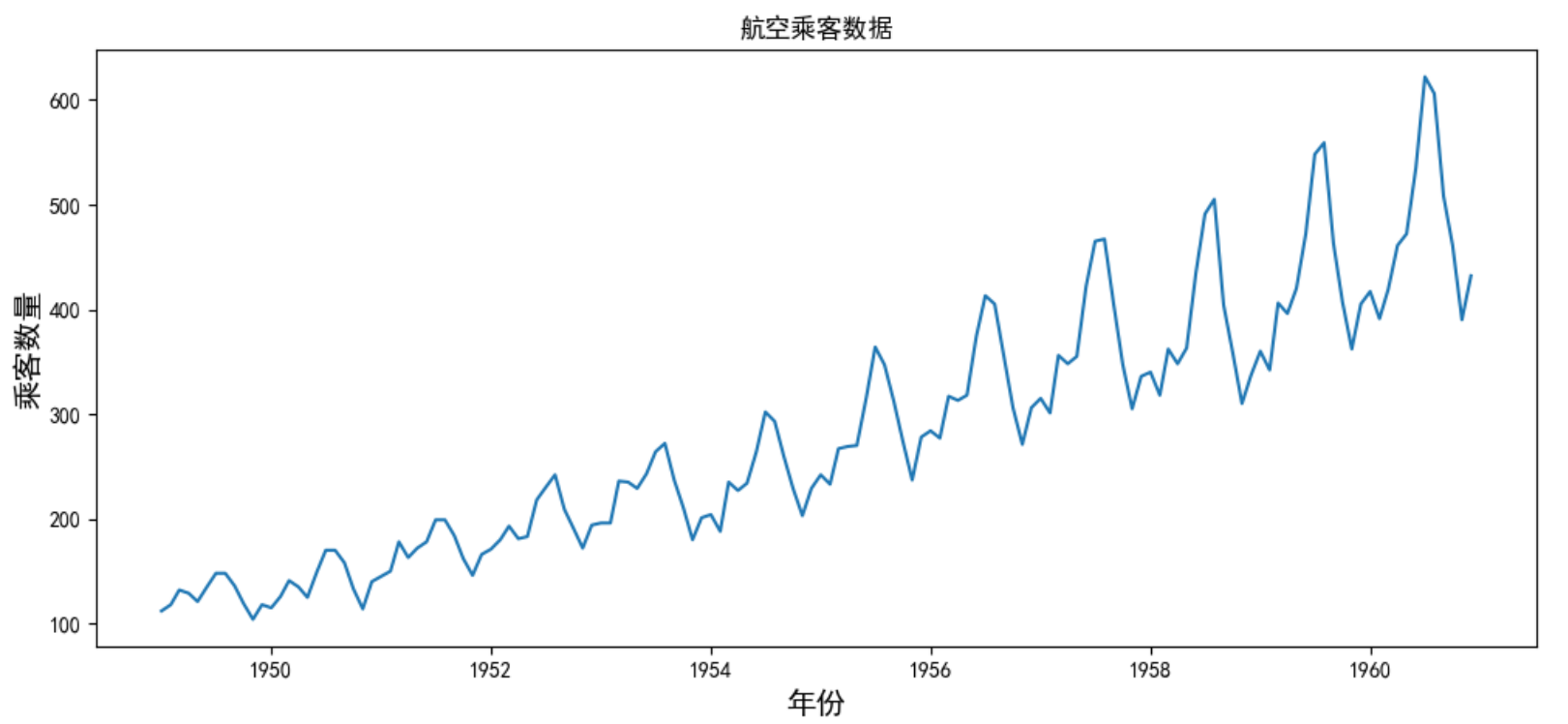
- 展开指数平滑公式，得到

$$S_t = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j x_{t-j}$$

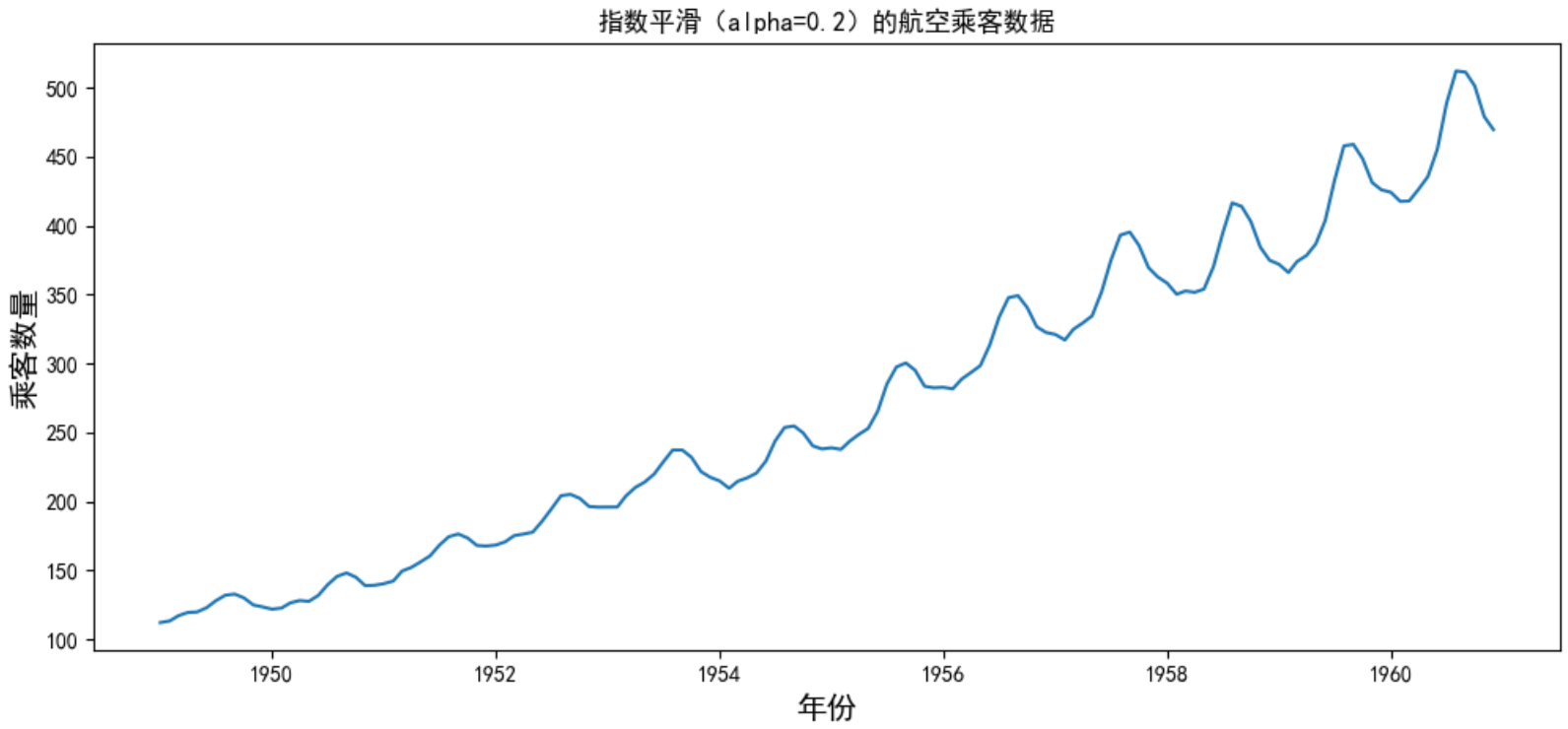
- S_t 是全部历史数据的加权平均，加权系数分别为 α，$\alpha(1-\alpha)$，…，根据等比求和公式，有

$$\alpha \sum_{j=0}^{\infty} (1-\alpha)^j = \frac{\alpha}{1-(1-\alpha)} = 1$$

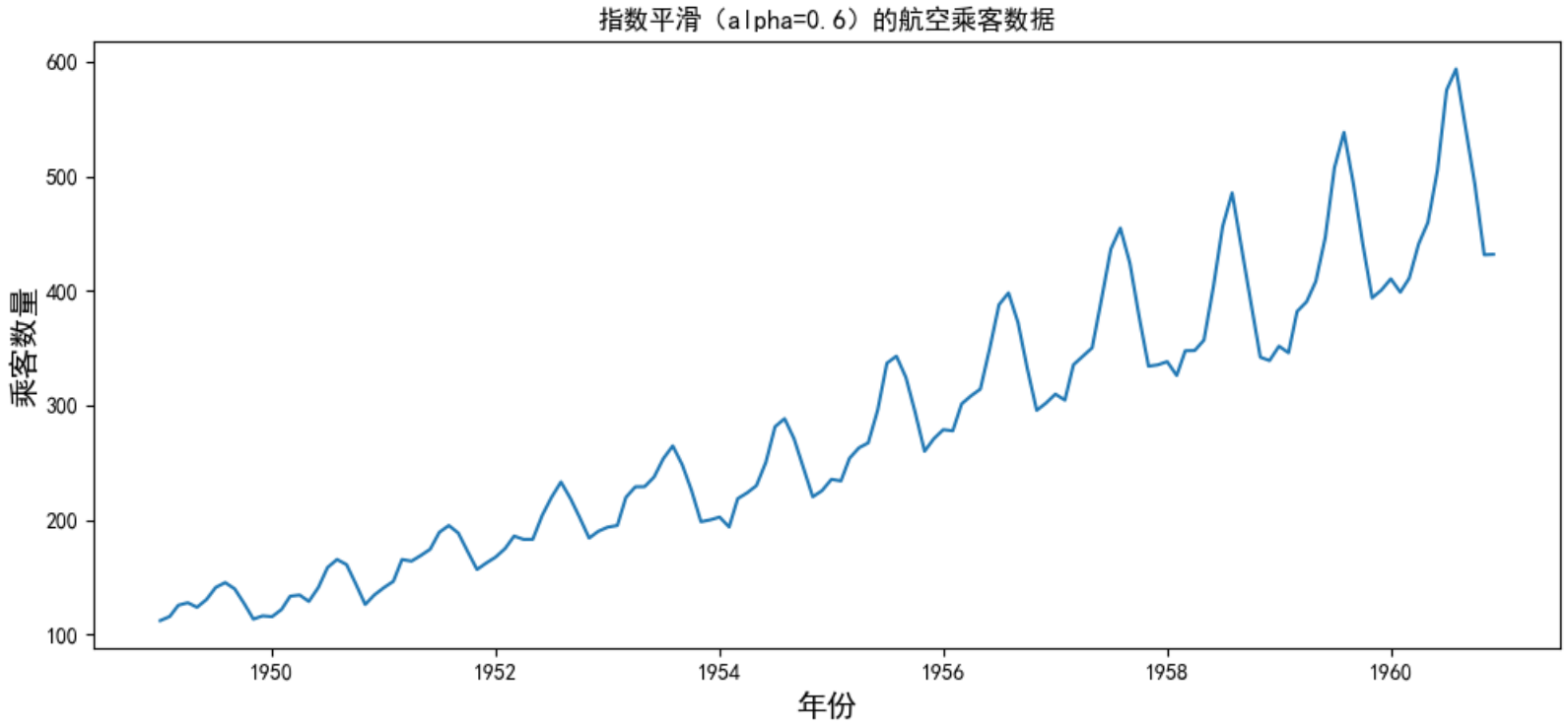
- 时序数据的噪声处理
  - 未平滑的Air Passengers数据集



- 时序数据的噪声处理
  - 指数平滑（alpha=0.2）的Air Passengers数据集



- 时序数据的噪声处理
  - 指数平滑（alpha=0.6）的Air Passengers数据集



数据清理

- 噪声的处理
  - 回归分析
  - 离群点分析
  - 分箱法
  - 移动平均法
  - 指数平滑法

Chapter 3.3

数据集成

数据集成

数据集成

- 数据集成的任务是合理地集成不同来源、格式、特点和性质的数据，形成完整的数据集后将其存放在统一的数据存储器。
- 相关性分析
  - 标称属性：采用 χ^2 检验进行相关性分析
  - 数值属性：采用相关系数和协方差进行相关性分析

Chapter 3.4

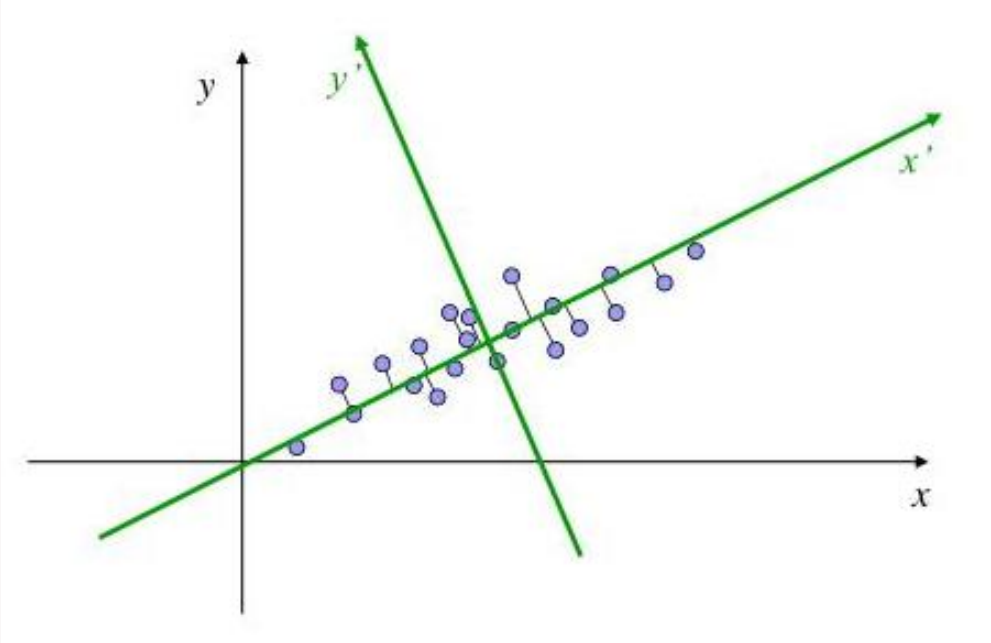
数据归约

数据归约

- 数据归约
  - **数据归约**（Data reduction）也称作数据消减或数据约简，指的是在保证规约前后的数据挖掘结果相似的条件下，缩小数据集规模的技术。
    - 属性冗余的处理
    - 数据对象冗余的处理

属性冗余的处理

- **主成分分析**（Principal components analysis）是一种常用的线性特征提取方法，其主要思想是寻找一个正交投影矩阵，使得投影后的数据具有最大的**可分性（方差）**。正交矩阵可以看作一组正交基，将数据投影到这个新的子空间后可以保证新的特征之间是线性无关的，从而获得一种**简洁**（不冗余）的低维表示。



属性冗余的处理

主成分分析

- 假设有数据集$X = [x_1, x_2, \dots, x_n] \in R^{d \times n}$，并且已经**零均值化处理**（$\sum x_i = 0$）。我们希望找到一个正交矩阵$W \in R^{d \times k}$（$W^T W = I, WW^T = I$），使得投影后的数据$W^T X$具有最大的可分性。
- 我们希望找到一个单位向量 w_1（$\|w_1\|_2^2 = 1$），使得 X 在 w_1 上具有最大的可分性。所以 X 在 w_1 上的方差表示为

$$Var[w_1^T X] = \sum_{i=1}^{n} (w_1^T x_i - w_1^T \bar{X})^2 = w_1^T X X^T w_1$$

- 属性冗余的处理
- 主成分分析

要优化的目标函数为

$$\max_{w_1} w_1^T X X^T w_1$$
$$s.t. w_1^T w_1 = 1$$

借助拉格朗日乘子法，优化的目标变为

$$\max_{\alpha} \alpha$$
$$s.t. w_1^T w_1 = 1$$

因为$XX^T w_1 = \alpha w_1$，所以 α 是 XX^T 的特征值，而 w_1 是 XX^T 的特征向量。

属性冗余的处理

主成分分析

假设得到 w_1 后，继续计算下一个单位向量 w_2（$\|w_2\|_2^2 = 1, w_2^T w_1 = 0$）。对于任意的 x_i，将其分解为平行于 w_1 的分量与残差分量，则有

$$x_i = w_1 w_1^T x_i + (x_i - w_1 w_1^T x_i)$$

残差分量就是垂直于 w_1 的超平面，所以只要在这个超平面中找到一个单位向量 w_2，即可满足$w_2^T w_1 = 0$的约束。

属性冗余的处理

主成分分析

问题变为找到一个单位向量 w_2，使得 $\hat{x}\,(x - w_1 w_1^T x)$ 在 w_2 上的投影具有最大方差。求解 w_2 的目标函数为

$$\max_{w_2} w_2^T X X^T w_2$$
$$s.t. w_2^T w_2 = 1$$

所以 w_2 是 XX^T 的第二大特征值对应的特征向量。

属性冗余的处理

主成分分析

经过 m 次计算后，得到了正交矩阵$W = [w_1, w_2, \ldots, w_m]$。对于任意的 w_i，都是通过最大化$w_i^T XX^T w_i$得到的。因此，只需要对 XX^T 进行一次特征值分解，然后按特征值大小对特征向量排序，就可以直接得到 W。

数据归约

- 数据对象冗余的处理
  - 抽样
    - **抽样**（Sampling）技术的目的是用随机抽取的数据子集代替原始的数据集。
  - 抽样方法
    - 不放回简单随机抽样
    - 有放回简单随机抽样
    - 聚类抽样
    - 分层抽样

数据归约

数据对象冗余的处理

- 无放回简单随机取样
  - 假定数据集 X 包含 N 个数据对象，从中随机（每一数据对象被选中的概率为$\frac{1}{N}$）抽取出 n 个数据对象以构成数据子集。
- 有放回简单随机取样
  - 与无放回简单随机抽样方法类似，但是每次随机抽取一个数据对象后，会将抽中的数据对象放回数据集 X。

数据归约

数据对象冗余的处理

- 聚类抽样
  - 将数据集 X 中的数据对象放入 M 个簇中，从 M 个簇中选择 c 个簇，在每个簇中进行有放回/无放回的简单随机抽样。
- 分层抽样
  - 根据属性将数据集 X 划分为互不相交的部分，称作“层”，在每个层中继续有放回/无放回的简单随机抽样。

Chapter 3.5

数据变换与离散化

数据变换与离散化

数据变换

- 数据变换指的是将数据变换或统一成适合于挖掘的形式，常有以下策略：
  - **规范化**：把属性数据按比例缩放，使之落入一个特定的小区间；
  - **离散化**：数值属性的原始值用区间标签或概念标签替换，这些标签可以递归地组织成更高概念；

数据变换与离散化

规范化

- 将数据按比例进行缩放，使之落入一个特定的区域，以消除数值型属性因大小不一而造成的挖掘结果的偏差。规范化的目的是赋予所有属性相等的权重，以消除量纲对挖掘结果的影响。
- 常用方法：
  - 最小-最大规范化
  - z-score规范化
  - 小数定标规范化

规范化

最小-最大规范化

- 假定A_{min}和A_{max}分别为属性 A 的最小观测值和最大观测值，对于任意的观测值 v_i，最小-最大规范化后的值表示为

$$v_i' = \frac{v_i - A_{min}}{A_{max} - A_{min}}(A'_{max} - A'_{min}) + A'_{min}$$

其中 v_i' 表示规范化的属性值，A'_{min}为映射区间的最小值，A'_{max}为映射区间的最大值。

特点

易受异常值的影响，且未来的观测值可能在$[A_{min}, A_{max}]$之外。

数据变换与离散化

规范化

- 基于标准差的z-score规范化
  - 对于属性 A 的任意观测值 v_i，z-score规范化的值表示为

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

$\bar{A}$ 表示属性 A 的平均值，σ_A 表示属性 A 的标准差。

- 特点
  - 使用均值而非极值，z-score规范化不易受异常值的影响；
  - 当属性 A 的实际最小值和最大值未知时，该方法的效果更好。

数据变换与离散化

规范化

- 小数定标规范化
  - 对属性值的小数点位置进行规范化。将属性 A 的观测值除以10的 j 次幂，使其落在[-1,1]的范围内。观测值 v_i 的小数定标规范化表示为

$$v_i' = \frac{v_i}{10^j}$$

  - j 是使 $\max|v_i| < 1$ 的最小整数。
- 例子
  - 某个属性的最小观测值为-12000，最大观测值为98000，则 j 取5，使得最小观测值规范化为-0.12，最大观测值规范化为0.98。

数据变换与离散化

数据规范化的适用范围

- **梯度下降法优化的模型需要对数据进行规范化**
- 考虑一个线性回归模型：

$$y = w_1 a_1 + w_2 a_2 + b$$

- 损失函数关于 w_1 的梯度表示为

$$\frac{\partial L}{\partial w_1} = \frac{1}{N} \sum_j^N (y_j - \hat{y}_j)\, a_{j1}$$

- 如果 a_1 和 a_2 具有不同的尺度，那么梯度下降时需要设置很小的学习率，以确保不会在更新参数时中出现数值溢出或振荡的问题，但这会导致收敛速度非常慢，需要更多的迭代次数才能达到收敛。

- 数据规范化的适用范围

- **基于距离的模型需要对数据进行规范化**

  - 考虑使用KNN模型进行分类，假设数据对象包含两个属性 a_1 和 a_2，则测试数据对象 y 与训练数据对象 x 的欧式距离表示为

$$d = \sqrt{(a_1^y - a_1^x)^2 + (a_2^y - a_2^x)^2}$$

  - 如果 a_1 和 a_2 具有不同的尺度，那么 d 将主要受尺度较大的属性的影响，而忽略了尺度较小的属性。

数据变换与离散化

- 数据规范化的适用范围
- **基于概率的模型不需要对数据进行规范化**
  - 朴素贝叶斯等基于概率的模型基于属性的分布和条件概率进行分类和预测，不涉及属性的数值关系或距离度量，所以不需要对数据规范化。

数据变换与离散化

- 数据规范化的适用范围
- **基于决策树的模型不需要对数据进行规范化**
  - 决策树、随机森林等模型通过比较特征值的大小进行决策，而不是进行特征之间的比较，也与距离度量无关，因此不需要对数据进行规范化。

数据变换与离散化

离散化

- 由于部分数据挖掘算法，比如关联分析相关算法、决策树、朴素贝叶斯分类器等，只适用于离散属性，因此需要将连续属性离散化，然后将数值属性的原始观测值用区间标签或概念标签替换。

数据变换与离散化

- 无监督离散化
  - 等宽离散化
    - 将属性的值域划分为若干个区间，且每个区间的宽度相等。
  - 例子
    - 将年龄属性的取值范围分成若干年龄段，如[0,10)，[10,20)，[20,30）等。

数据变换与离散化

- 无监督离散化
  - 等频离散化
    - 根据属性的观测值出现的频数将属性的值域划分为若干个区间，且每个区间的观测值数量相同。
  - 例子
    - 将月收入属性分成三个区间，分别代表高、中、低三个收入水平，每个区间包含相等数量的观测值。

数据变换与离散化

- 无监督离散化
  - 等宽离散化与等频离散化
    - 等宽离散化计算简单，更适用于分布均匀的属性
    - 分布不均匀时，过度集中的趋势可能造成不平衡问题
  - 例子
    - 假如年龄属性的观测值为[10, 15, 20, 25, 30, 40, 60, 70, 80]，宽度设置为30，区间表示为[10,40)，[40,70)，[70,100），则离散化后的观测值表示为[青年，青年，青年，青年，青年，中年，中年，老年，老年]。

数据变换与离散化

- 无监督离散化
  - 等宽离散化与等频离散化
    - 等频离散化更适用于分布不均匀的属性。
  - 例子
    - 假如年龄属性的观测值为[10, 15, 20, 25, 30, 40, 60, 70, 80]，频数设置为3，则三个区间的观测值分别为[10,15,20]，[25,30,40]，[60,70,80]。

数据变换与离散化

- 无监督离散化
- 聚类离散化
  - 对属性的观测值进行聚类（一维），根据聚类结果将观测值替换为对应簇的标记。

数据变换与离散化

- 有监督离散化
  - 基于决策树的离散化
  - 基于支持向量机的离散化
  - 基于卡方检验的离散化

Chapter 3

本章小结

本章小结

- 数据质量用准确性、完整性、一致性、时效性、可信性和可解释性定义。
- 数据清理通过填充缺失值，光滑噪声的方式提高数据质量。
- 数据集成将多个数据集整合成一致的数据存储。
- 数据归约以损失信息最小作为条件，实现数据集的简洁表示。
- 数据变换的目的是将数据变换成适合挖掘方法的形式，主要包括规范化和数据离散化。
- 数据预处理的主要任务之间虽然侧重点不同，但是也存在技术上的重叠。

第三章 完结