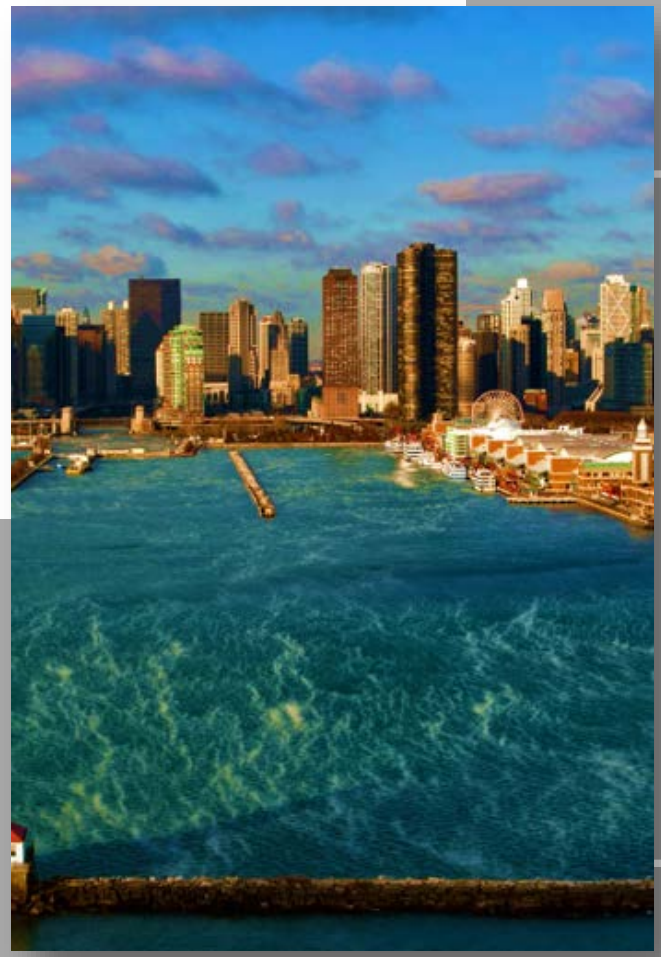
工作总结汇报

王益国

2019 / 10 / 23

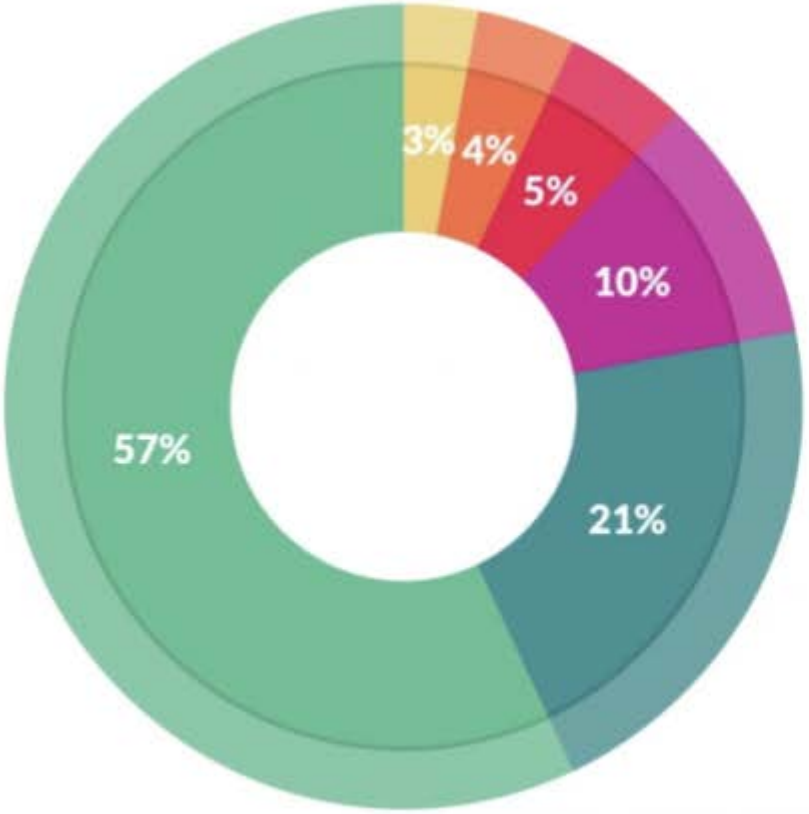


CONTENT

01 工程

如何做好项目前期的数据处理工作?

重要性



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

主要内容

数据清洗：解决数据质量问题

数据探索：了解数据的基本情况

什么是数据质量？

如何做好数据清理？

Data quality

- Validity
- Accuracy
- Completeness
- Consistency & Uniformity

Data quality

- **Validity**
- Accuracy
- Completeness
- Consistency & Uniformity

- ***Data-Type Constraints***
- ***Range Constraints***
- ***Mandatory Constraints***
- ***Unique Constraints***
- ***Set-Membership constraints***
- ***Foreign-key constraints***
- **Regular expression patterns**
- **Cross-field validation**

Data quality

- Validity
- **Accuracy** → ***Outliers***
- Completeness
- Consistency & Uniformity

Data quality

- Validity
- Accuracy
- **Completeness** → ***Missing values***
- Consistency & Uniformity

Data quality

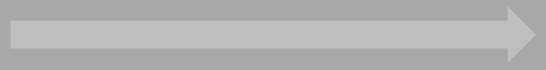
- Validity
- Accuracy
- Completeness
- **Consistency & Uniformity**

The workflow

1. **Inspection:** Detect unexpected, incorrect, and inconsistent data.
2. **Cleaning:** Fix or remove the anomalies discovered.
3. **Verifying:** After cleaning, the results are inspected to verify correctness.
4. **Reporting:** A report about the changes made and the quality of the currently stored data is recorded.

The workflow

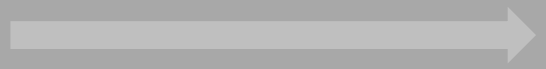
1. Inspection
2. Cleaning
3. Verifying
4. Reporting



- Data profiling
- Visualizations
- Statistics

The workflow

1.Inspection
2.Cleaning
3.Verifying
4.Reporting



- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- Normalization
- Missing values
- Outliers

The workflow

- 相等，删除一个
- 不相等，使用差值法估计点的数值，取相近的一个。

- Irrelevant data
- **Duplicates**
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- Normalization
- Missing values
- Outliers

The workflow

- Numbers
- Strings
- Date（timestamp）
- NA

- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- Normalization
- Missing values
- Outliers

The workflow

For strings, make sure all values are either in lower or upper case.

For numerical values, make sure all values have a certain measurement unit.

- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- **Standardize**
- Scaling / Transformation
- Normalization
- Missing values
- Outliers

The workflow

In most cases, we normalize the data if we're going to be using statistical methods that rely on normally distributed data.

- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- **Normalization**
- Missing values
- Outliers

The workflow

1. Drop
2. **Impute**
3. Flag

根据其他结果计算缺失值

- **使用统计值，**如均值（不偏斜），中位数（偏斜）。
- 使用**线性回归。（对异常值敏感）**
- **Hot-deck**：从其他类似记录中复制值。
- 一定情况下，随机值填充。
- 其他机器学习的方法

- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- Normalization
- **Missing values**
- Outliers

The workflow

离群点检测

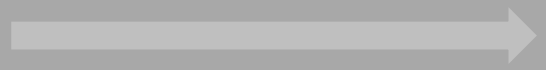
- 基于统计的方法
- 基于距离的方法
- 基于密度的方法
- 基于偏离的方法

- Irrelevant data
- Duplicates
- Type conversion
- Syntax errors
- Standardize
- Scaling / Transformation
- Normalization
- Missing values
- **Outliers**

总的方法是通过一定的方法预测该点的值，与原始值做对比判断是否是异常值。

The workflow

1.Inspection
2.Cleaning
3.Verifying
4.Reporting



Data quality

The workflow

1.Inspection
2.Cleaning
3.Verifying
4.Reporting → 记录过程；溯源

02 研究？

时间序列离群点检测

现有方法

1. Moving Average

移动平均是一种分析时间序列的常用工具，它可过滤高频噪声和检测异常点。根据计算方法的不同，常用的移动平均算法包括

- 简单移动平均，SMA
- 加权移动平均，WMA
- 指数移动平均，EMA

通过以上方法计算之前数据的平均值检测现有值是否为异常点。

现有方法

1. Moving Average
2. **同比环比**

适合数据呈周期性规律的场景中。如果同比或环比超过一定阈值，可认定为离群点。

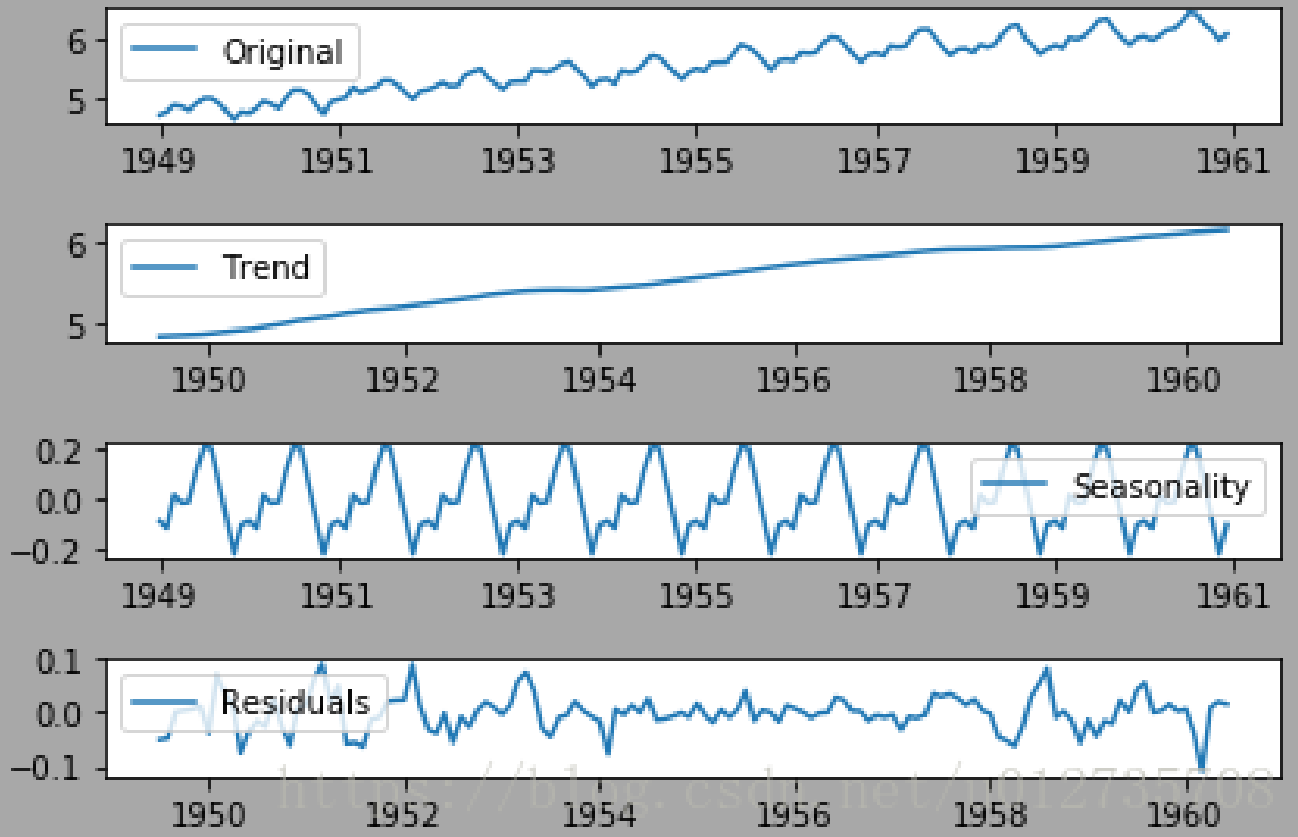
现有方法

1. Moving Average
2. 同比环比
3. **ARIMA模型**

- Autoregressive Integrated Moving Average Model，自回归差分移动平均模型
- 缺点：要求时序数据是稳定的（stationary），或者是通过差分化后是稳定的。只能捕捉线性关系。
- 如何判断稳定：稳定的数据是没有趋势(trend)，没有周期性(seasonality)的; 即它的均值，在时间轴上拥有常量的振幅，并且它的方差，在时间轴上是趋于同一个稳定的值的。

现有方法

1. Moving Average
2. 同比环比
3. **ARIMA模型**



将时序数据分离成不同的成分：长期趋势、季节趋势、随机成分。

现有方法

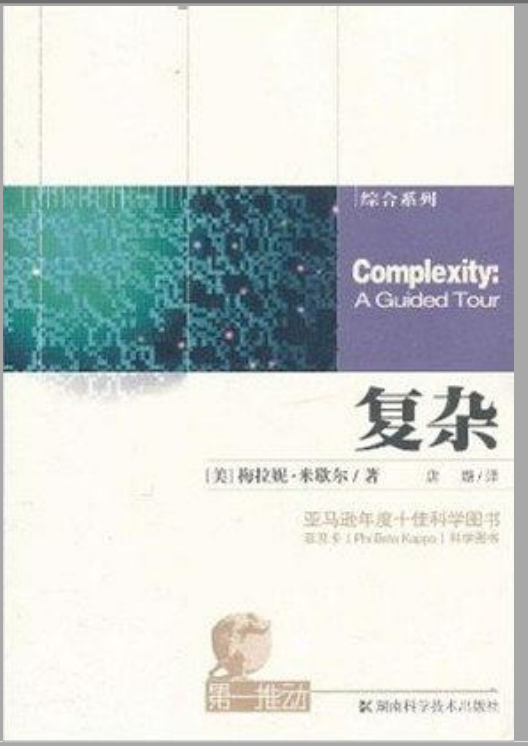
1. Moving Average
2. 同比环比
3. ARIMA模型
4. **贝叶斯方法**
5. **遗传算法**
6. **神经网络（LSTM）**
7. **小波检测**

03 感想

为什么要读在职博士？

宏观方面

- 牛顿世界观是一种“钟表宇宙”，可预测，可规划。
- 混沌系统则是一种“非线性系统”，不可预测。



→ 放弃确定性的直觉期待，用概率论的思维看待世界。

坚决服从概率,坚定不移地去实践高概率的方向。

(吴军）叠加式进步

高概率的方向

- 叠加式进步
- 每天进步一点点
- 终身学习
- 1.01的365次方是37.8
- 1.02的365次方是1377.4
- 积跬步以至千里

高概率的方向

微观方面

- 舒适的工作环境
- 固定的工作内容
- 稳定的同事朋友
- 工作领导
- 机械思维
- 横向扩展

微观方面

- 舒适的工作环境
- 固定的工作内容
- 稳定的同事朋友
- 工作领导
- 机械思维
- 横向扩展

- 竞争性环境
- 喜欢的研究方向
- 各种优秀的同学
- 科研导师
- 认知迭代
- 纵向延伸

请多指教

无限风光在险峰