第03课 2019-11-17

马克飞象

特征工程和结果可视化

□□特征工程 和 结果可视化

> 内容涵盖:主流Python数据预处理库，原始数据特征构建。特征选择，构建新特征，缺 失值填充等等特征工程方法学习，数据预处理为机器学习做准备和铺垫。

- Python 机器学习库 (★★)
- Python AI主流库介绍和典型使用 (★★★)
- 特征工程 (★★★)
- 模型加载，模型存储 (★★)
- 特征选择 (★★★)
- 降维 (★)
- 分析结果可视化 (★★)
- Python 参数搜索 (★★)
- 实战案例 (★★)

Python机器学习库

工具管理

- Anaconda 包管理 环境管理 解决各种第三方依赖问题
- conda env list
- conda avtive/ deactivate
- conda install
- conda install tensflow==1.14
- Pip Python□□□□□□
- $ pip -V #□□pip□□ 版本
- pip list

机器学习库

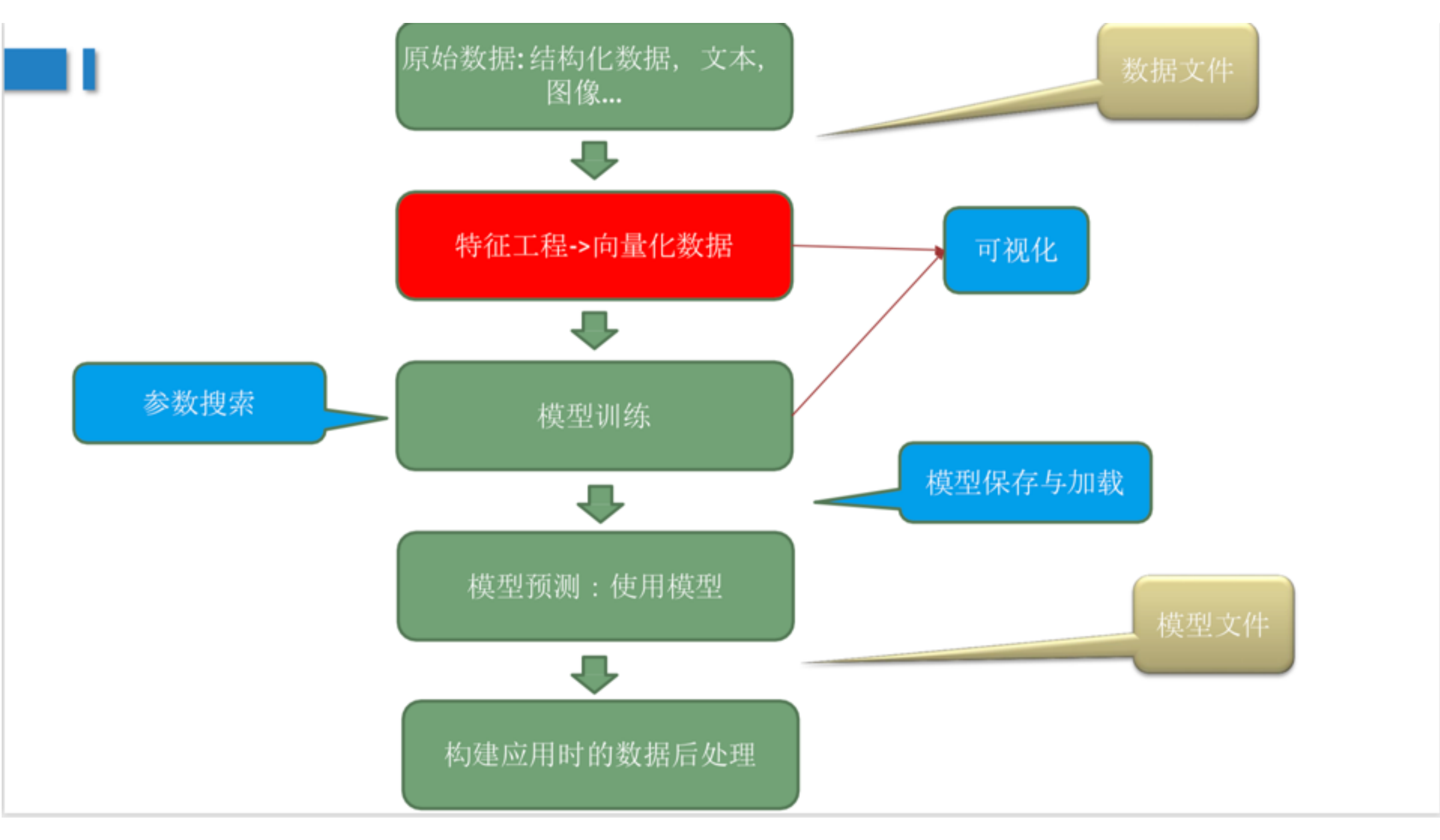
- 1) 机器学习
- Scikit-learn
- 基 于 NumPy, SciPy, Matplotlib
- 开源，涵盖分类，回归和聚类算法
- 代码和文档完备

- 2)数据处理
- Pandas 基于NumPy和Matplotlib 数据分析和数据可视化
- 数据结构DataFrame

- 3)科学计算
- Numpy
- 数值编程工具 矩阵数据类型、矢量处理

- 4)可视化 Matplotlib
- 绘图库
- 和matlab相似的命令API

特征工程流程



> **特征工程 -> 向量化数据**

Python AI主流库介绍和典型使用

Numpy

- 数组操作数据提升数据的处理效率
- 类似于R的向量化操作
- 可以进行数组和矢量计算

Pandas

- 带有标签的列和索引
- csv 类型的文件中导入数据
- 对数据进行复杂的转换和过滤等操作
- series 和 dataframe

> series 是一种一维的数据类型，其中的每个 元素都有各自的标签。
> dataframe 是一个二维的、表格型的数据结 构。可以把它当作一个 series 的字典。

```
import numpy as np
import pandas as pd
from pandas import Series, DataFrame
data = DataFrame(np.arange(16).reshape(4,4),index=list('abcd'),columns=list('wxyz'))
data['w']   #□□□□□□'w'□□□□□□□□□□,□□□□Series□□
data.w      #□□□□□□'w'□□□□□□□,□□□□Series□□
data[['w']] #□□□□□□'w'□□□□□□DataFrame□□
data[['w','z']] #□□□□□□□'w'□'z'□
```

框架类型

- 通用机器学习:Scikit-learn, Spark MLlib
- 特定系列算法:XGBoost
- 深度学习:Tensorflow，**Keras**

scikit-learn

- 机器学习的Python库
- 分类、聚类以及回归分析方法
- 强大的功能、优异的拓展性以及易用性
- 著名的一个开源项目之一

Matplotlib

- 2D绘图库
- 各种格式
- 跨平台的交互式环境
- 生成出版质量级别的图形

特征工程

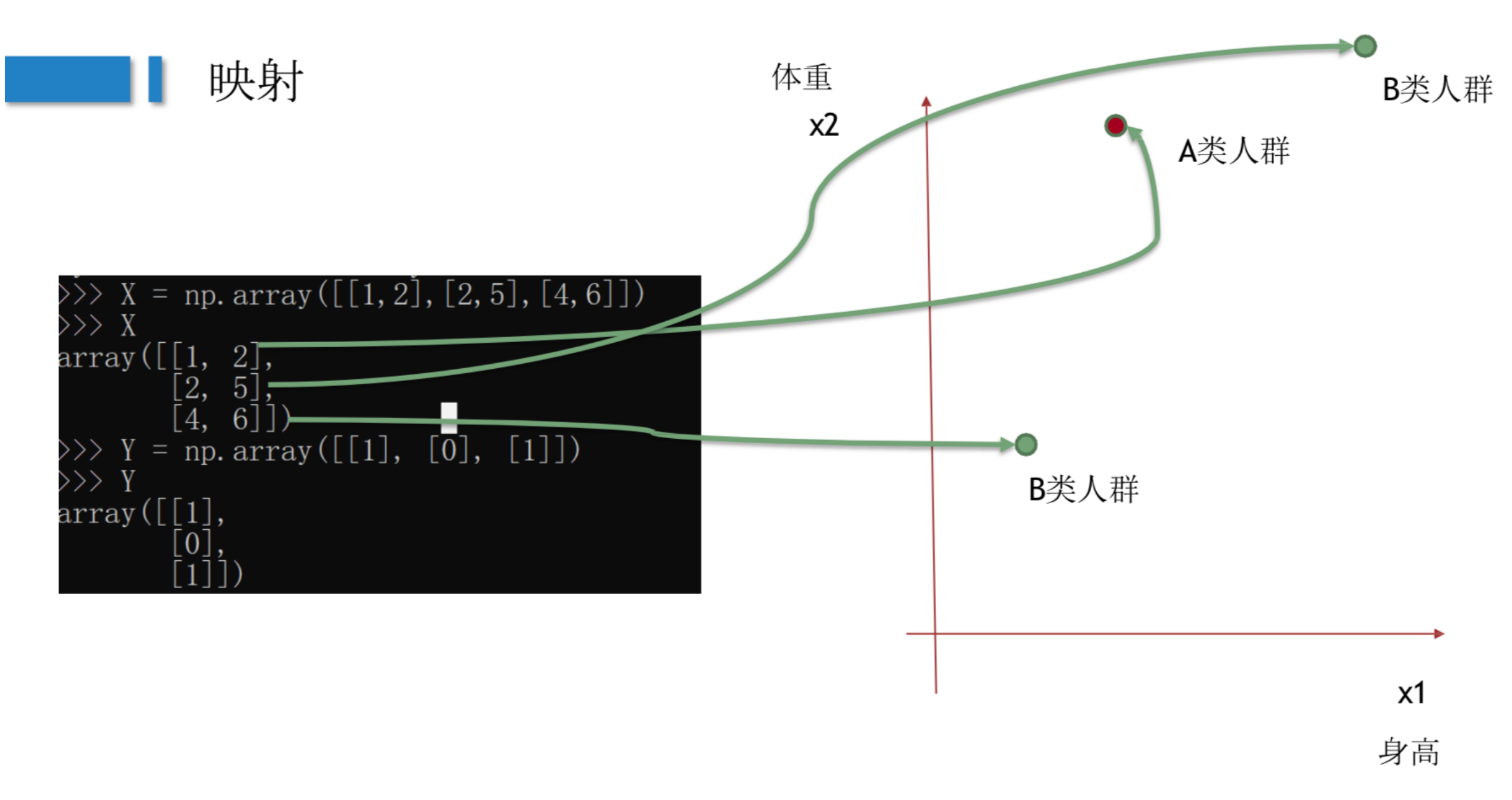
> **其本质是一项工程活动，目的是最大限度地 从原始数据中提取特征以供算法和模型使用**

特征工程- 解决的一些问题

- 加工数据为模型可处理的格式——向量化和规范化
- 某些方法能加□一些优化算法的收敛性
- 能让数据变得使模型更容易拟合:简化数据/简化机器学习问题

□□□□□□□□□□□□□□□□□□模型输入正确

映射



```
>>> X = np.array([[1,2],[2,5],[4,6]])
>>> X
array([[1, 2],
       [2, 5],
       [4, 6]])
>>> Y = np.array([[1], [0], [1]])
>>> Y
array([[1],
       [0],
       [1]])
```

特征映射

- Input space -mapping- Feature space
- 特征工程 收敛更快

Feature Scaling

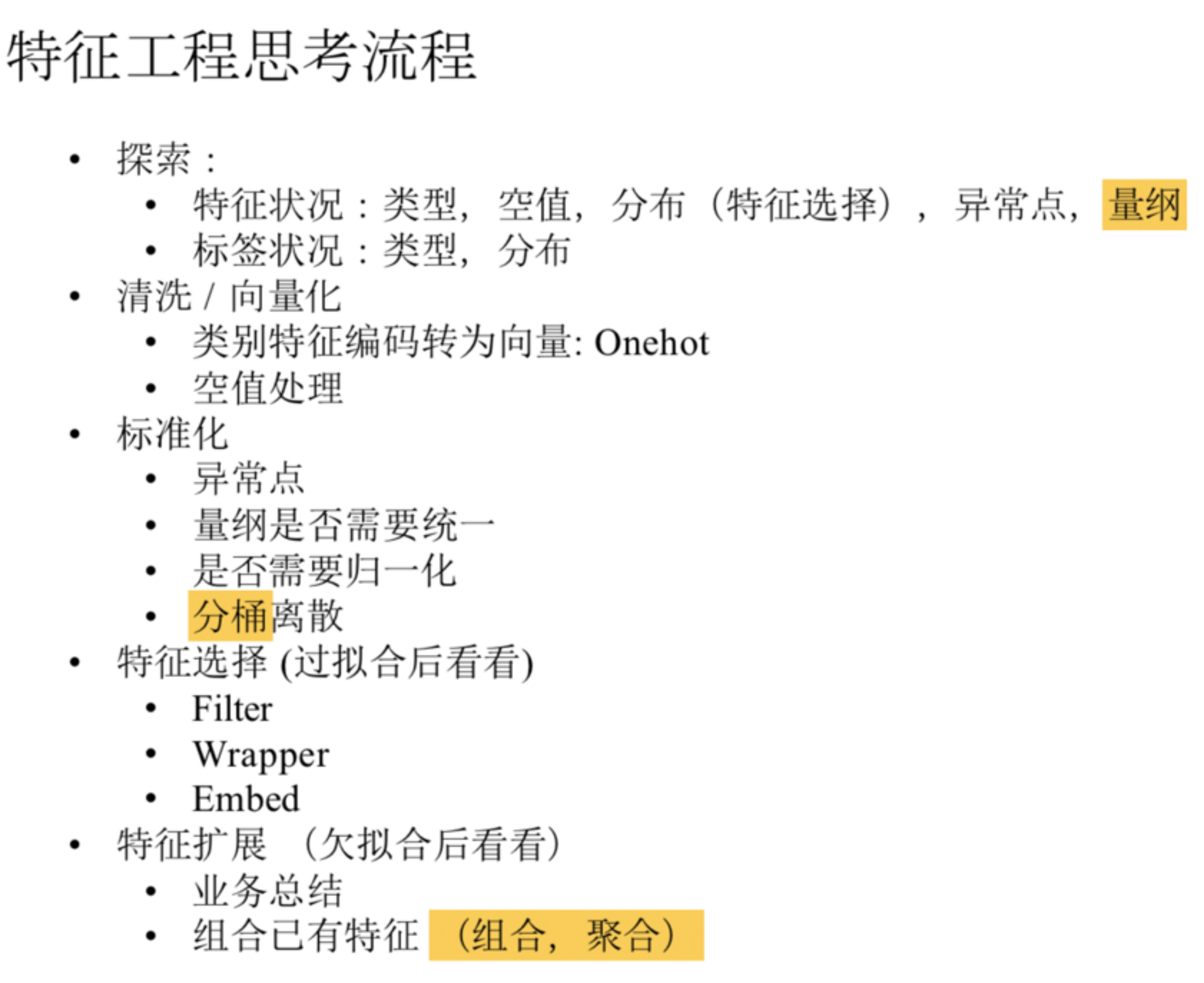
向量相似度

- 余玄相似度
- 向量的归一化

□□□□特征工程

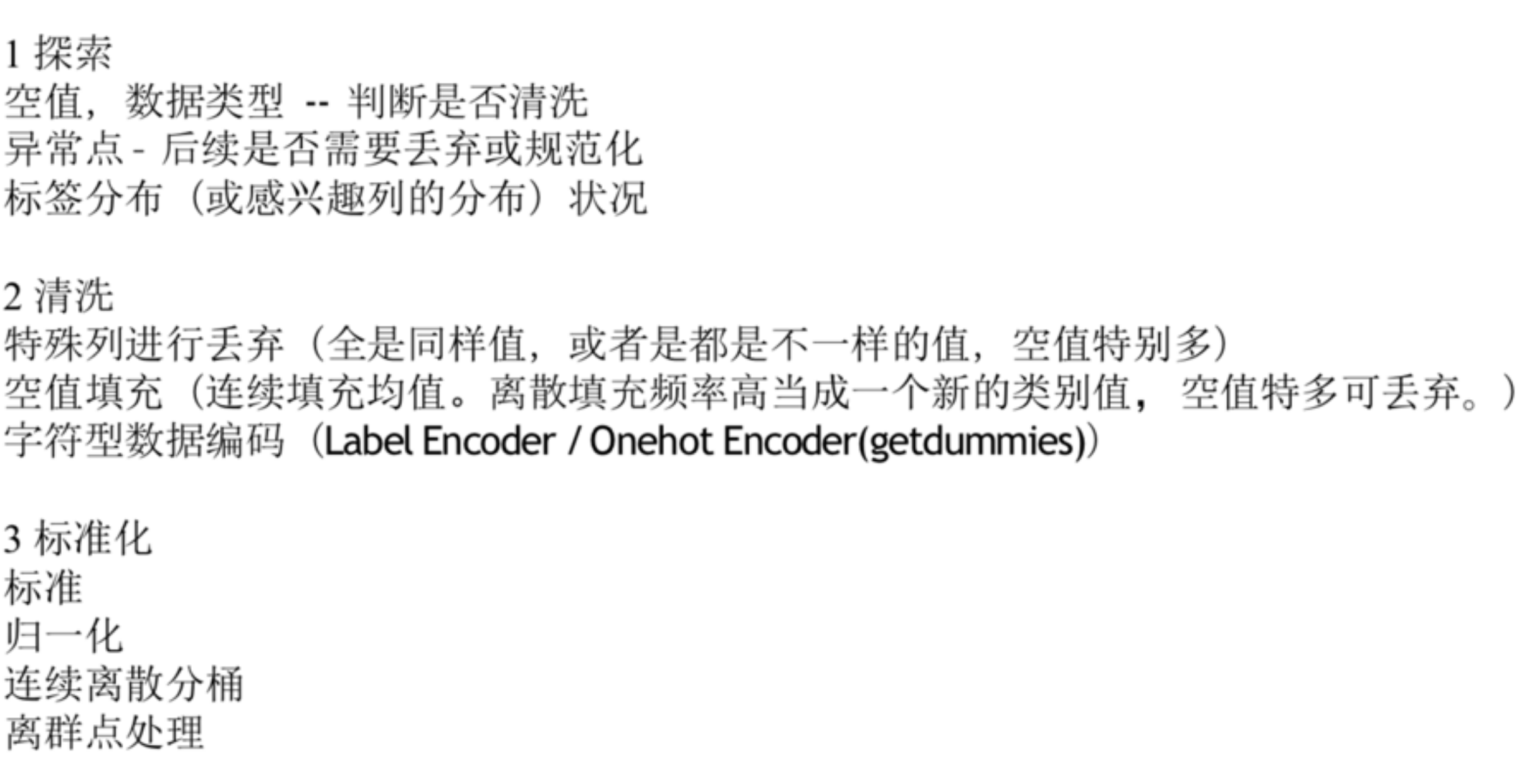
- 特征如何使用? - 业务理解，可用性评估
- 特征如何获取? - 获取与存储
- 特征如何处理?
- 探索，清洗，标准化，特征选择，特征扩展 更新特征?

特征工程思考流程



特征工程思考流程

- 探索：
  - 特征状况：类型，空值，分布（特征选择），异常点，量纲
  - 标签状况：类型，分布
- 清洗 / 向量化
  - 类别特征编码转为向量: Onehot
  - 空值处理
- 标准化
  - 异常点
  - 量纲是否需要统一
  - 是否需要归一化
  - 分桶离散
- 特征选择 (过拟合后看看)
  - Filter
  - Wrapper
  - Embed
- 特征扩展 （欠拟合后看看）
  - 业务总结
  - 组合已有特征 （组合，聚合）



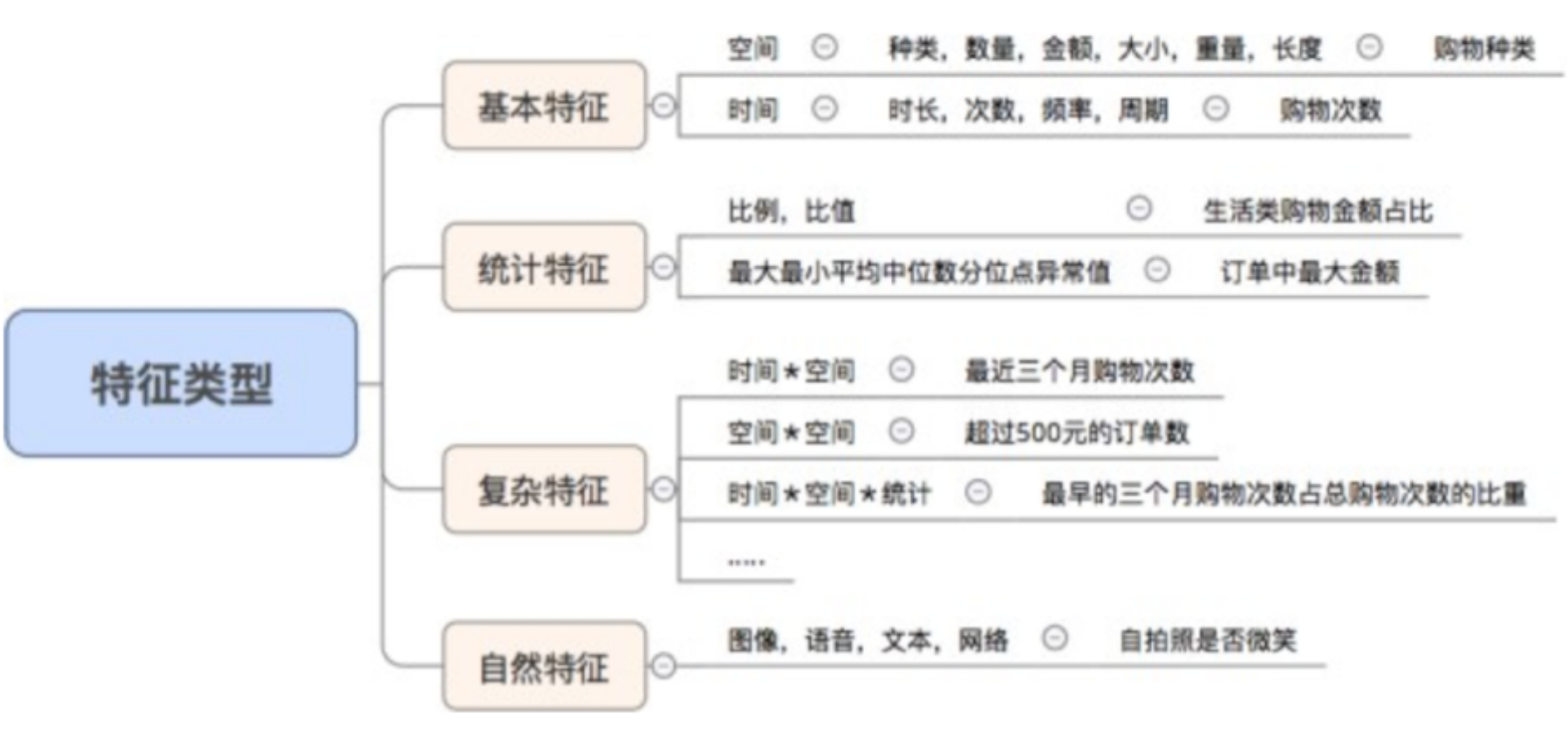
1 探索
空值，数据类型 -- 判断是否清洗
异常点 - 后续是否需要丢弃或规范化
标签分布（或感兴趣列的分布）状况

2 清洗
特殊列进行丢弃（全是同样值，或者是都是不一样的值，空值特别多）
空值填充（连续填充均值。离散填充频率高当成一个新的类别值， 空值特多可丢弃。）
字符型数据编码（Label Encoder / Onehot Encoder(getdummies)）

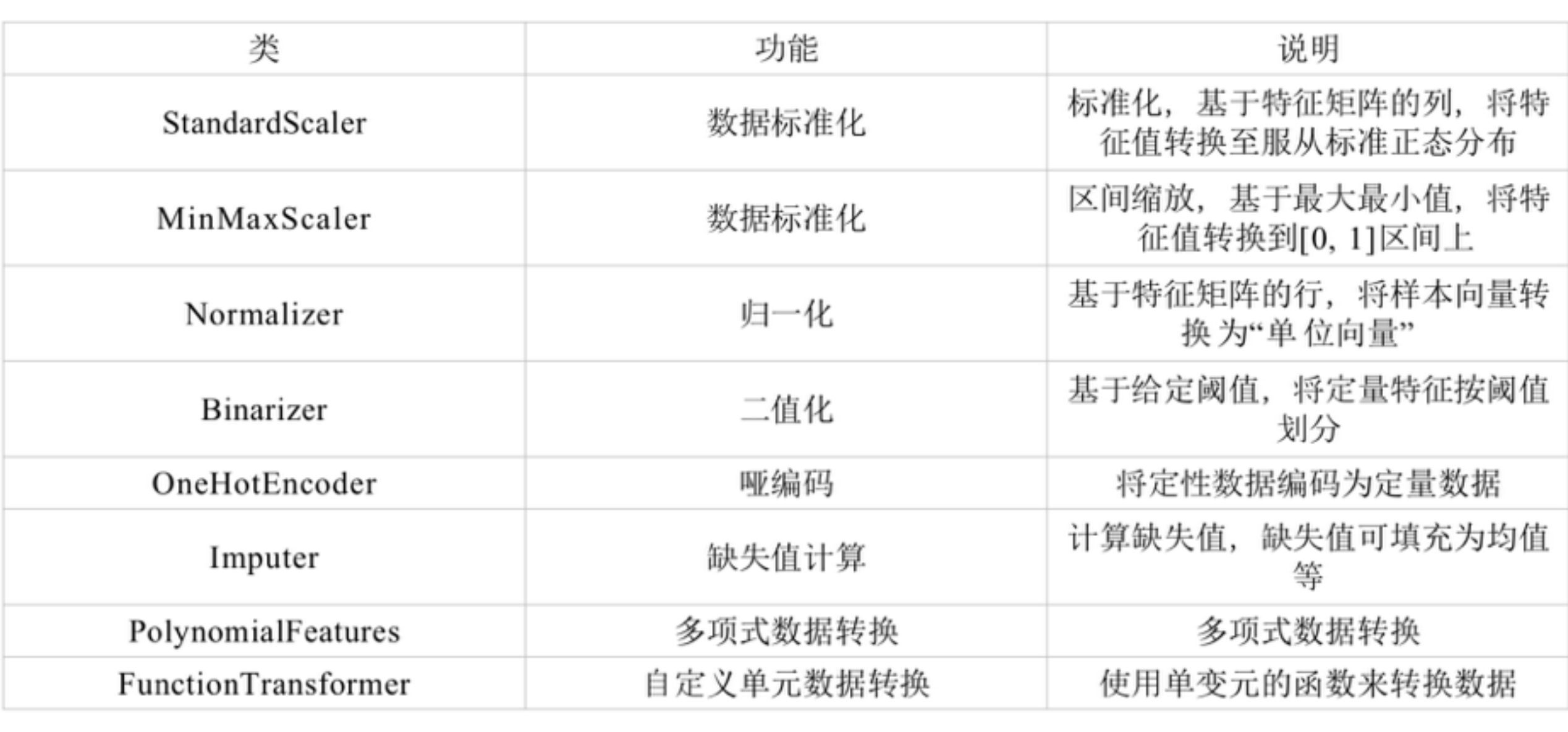
3 标准化
标准
归一化
连续离散分桶
离群点处理

- 过拟合 欠拟合

特征类型



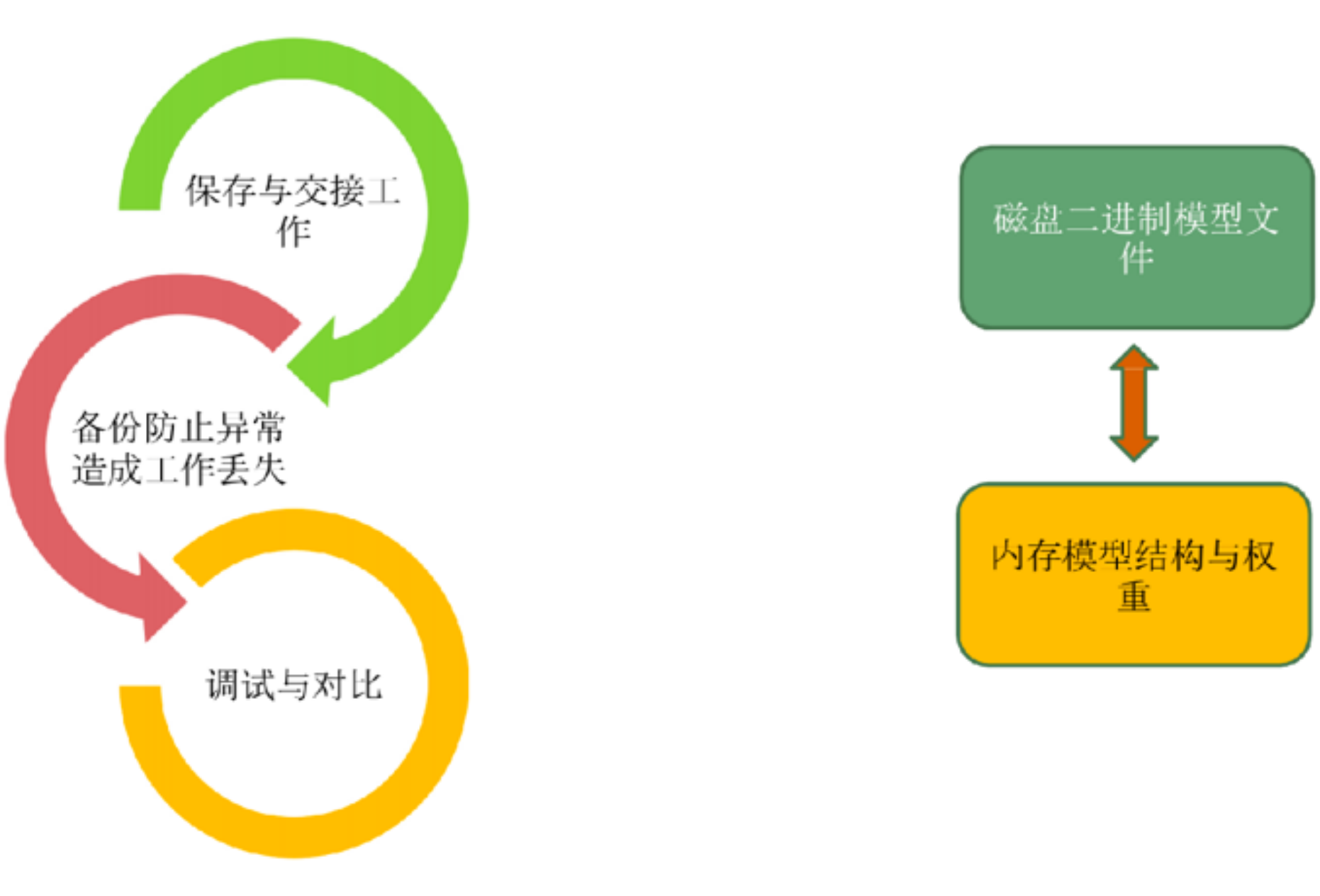
scikit-learn □□□□API



类	功能	说明
StandardScaler	数据标准化	标准化，基于特征矩阵的列，将特征值转换至服从标准正态分布
MinMaxScaler	数据标准化	区间缩放，基于最大最小值，将特征值转换到[0, 1]区间上
Normalizer	归一化	基于特征矩阵的行，将样本向量转换 为“单 位向量”
Binarizer	二值化	基于给定阈值，将定量特征按阈值划分
OneHotEncoder	哑编码	将定性数据编码为定量数据
Imputer	缺失值计算	计算缺失值，缺失值可填充为均值等
PolynomialFeatures	多项式数据转换	多项式数据转换
FunctionTransformer	自定义单元数据转换	使用单变元的函数来转换数据

特征工程 方法作用

- 让算法能正常运行的，输入向量化与规范null和 onehot encoder 等
- 加速算法收敛的 标准化相关的
- 让数据更加可分（解决过拟合和欠拟合现象），问题更加简化特征扩展和特征选择



- Everything to vector
- as_matrix()

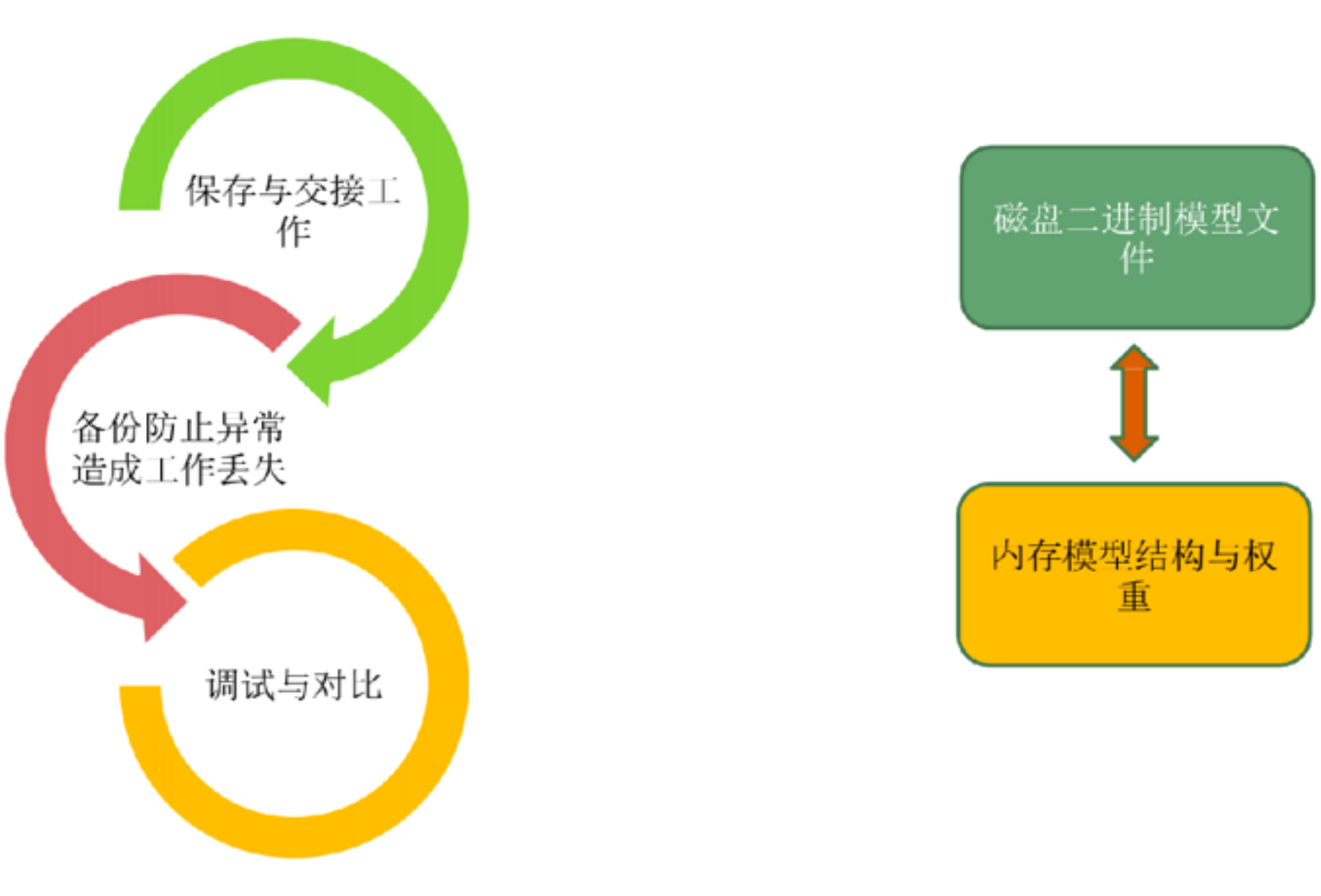
特征选择□□□□（实例）

降维

> 降维与PCA
>
> - 缓解维度灾难
> - 压缩数据时让信息损失最小化
> - 高维数据可视化

模型加载 和 模型存储

> 模型存储和加载



- RFE 递归特征消除（RFE）+ 交叉验证
- Pearson Correlation Coefficient

> **模型备份 加载 参数配置保存 json txt**

分析结果可视化

Python 参数搜索

- 参数搜索 Grid Search OR Random Search
- AutoML
- NAS