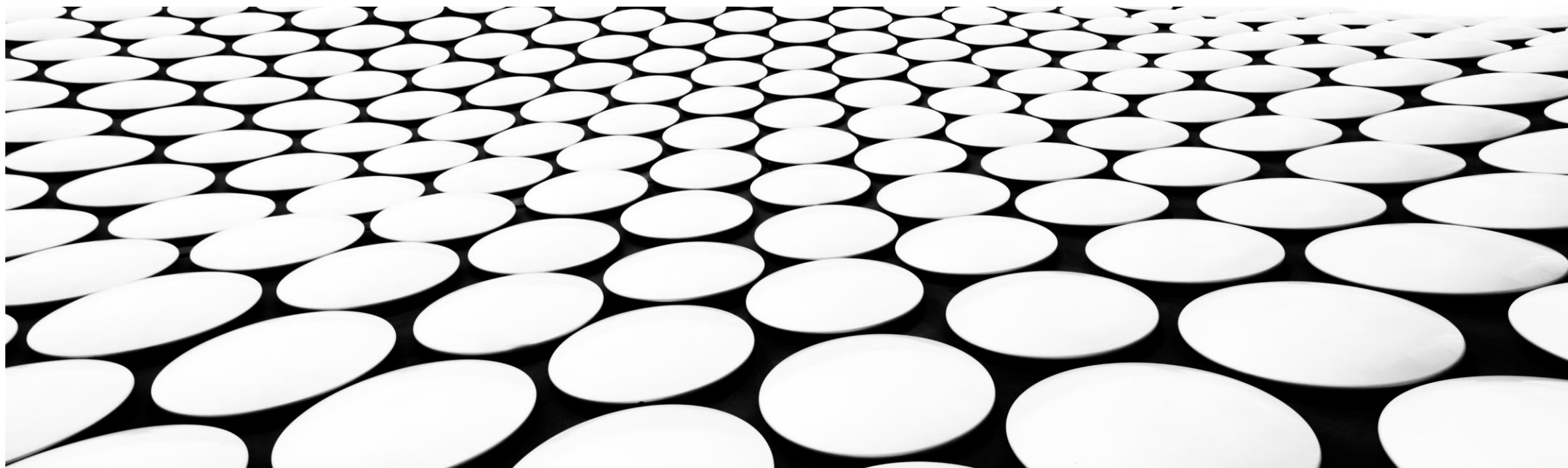


分布式计算

邱怡轩



今天的主题

- Spark 分布式机器学习库



Spark MLlib

MLlib

- Spark 还提供了一套基于分布式框架的机器学习库，称为 MLlib
- <https://spark.apache.org/mllib/>

MLlib

- 基于 Spark 的基础框架和数据结构
- 涵盖了经典的机器学习模型和任务
- 封装了较完整的数据处理流程

MLlib

- 分类
- 回归
- 聚类
- 推荐系统
- 主题模型
-

分类模型

- Logistic 回归
- 决策树
- 随机森林
- Gradient boosting
- 神经网络
- 支持向量机
- 朴素贝叶斯
-

回归模型

- 线性回归（可带正则项）
- 广义线性回归
- 决策树
- 随机森林
- Gradient boosting
-

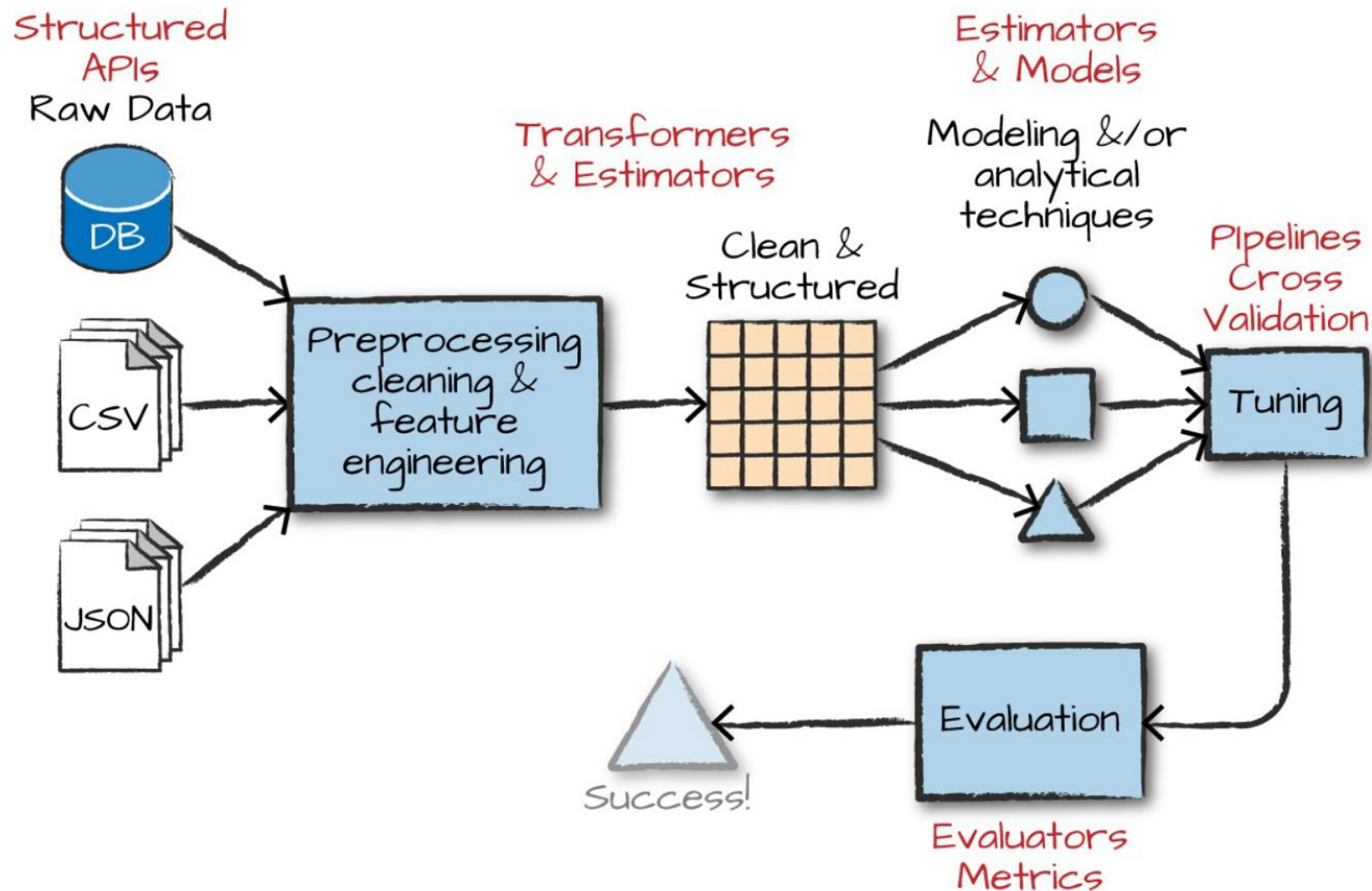
聚类模型

- Kmeans
- 高斯混合模型
-

一般流程

- 读取数据, 转换为 DataFrame
- 特征提取、变换
- 指定模型, 设定参数
- 模型训练
- 预测
- 模型评价

一般流程



读取数据

- 参见 [lec14-dataframe.ipynb](#)

特征转换

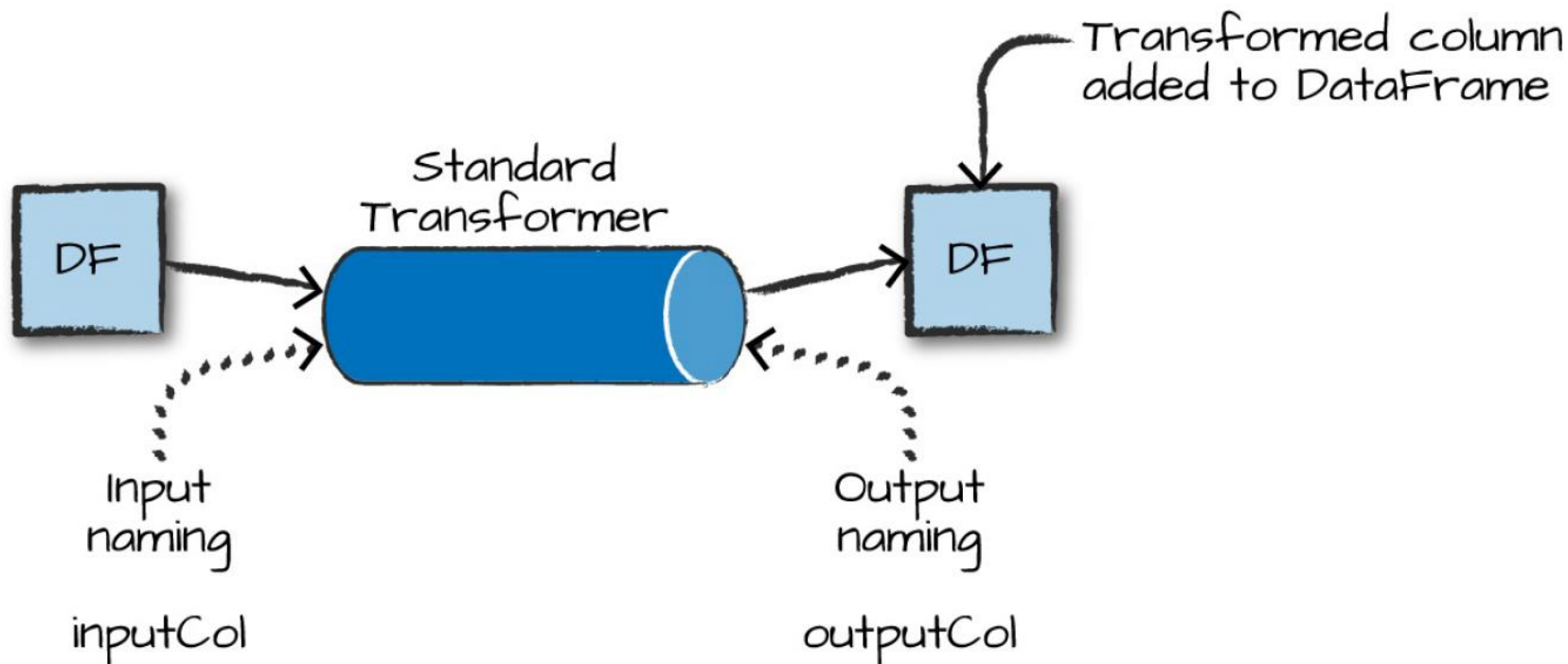
- MLlib 定义了一系列对特征进行操作的转换器，称为 Transformer，如：
- One-hot 编码 ([OneHotEncoder](#))
- 字符进行数字编码 ([StringIndexer](#))
- 标准化 ([StandardScaler](#))
- <https://spark.apache.org/docs/latest/ml-features.html>

特征转换

- MLlib 需要将所有用于预测的变量组合成一个向量
- 使用 VectorAssembler 进行合并
- MLlib 还提供了类似 R 中 formula 的接口 RFormula

特征转换

- Transformer 本质上是把一个 DataFrame 转换成另一个 DataFrame 的对象
- `newdat = trans.transform(dat)`



建立模型

- 选取合适的模型，查找其参数设定方法
- <https://spark.apache.org/docs/latest/ml-guide.html>

<https://spark.apache.org/docs/latest/ml-guide.html>

APACHE Spark 3.2.0 Overview Programming Guides ▾ API Docs ▾ Deploying ▾ More ▾

MLlib: Main Guide

- Basic statistics
- Data sources
- Pipelines
- Extracting, transforming and selecting features
- Classification and Regression
- Clustering
- Collaborative filtering
- Frequent Pattern Mining
- Model selection and tuning
- Advanced topics

Machine Learning Library (MLlib) Guide

MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable. It provides tools such as:

- ML Algorithms: common learning algorithms such as classification, regression, clustering, and
- Featurization: feature extraction, transformation, dimensionality reduction, and selection
- Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
- Persistence: saving and load algorithms, models, and Pipelines
- Utilities: linear algebra, statistics, data handling, etc.

Announcement: DataFrame-based API is primary

The MLlib RDD-based API is now in maintenance mode.

As of Spark 2.0, the [RDD](#)-based APIs in the `spark.mllib` package have entered maintenance mode.

建立模型

- 对于有监督学习模型，常见的参数设定有
- `setLabelCol()`：指定预测目标的列名
- `setFeaturesCol()`：指定特征的列名
- `setPredictionCol()`：指定输出预测的列名
- `setSeed()`：设置随机数种子
- 无监督学习模型通常没有 `setLabelCol()`

建立模型

- 具体每个模型的建立方法可查阅官方文档

模型训练

- 通常是调用模型对象的 `fit()` 方法
- `fitted = model.fit(dat)`

模型预测

- 将训练好的模型在预测集上调用 `transform()` 方法
- `pred = fitted.transform(newdat)`

模型评价

- 给定预测的结果，计算评价模型优劣的指标
- 不同的模型对应不同的评价指标
- 如分类有
MulticlassClassificationEvaluator
- 聚类有 ClusteringEvaluator
- `evaluator.evaluate(pred)`

例子

- 参见 [lec15-mllib.ipynb](#)

分布式计算2023

Ask me anything



长按图片扫码