9.5 机器学习流水线



9.5.1 流水线的概念

9.5.2 流水线工作过程

9.5.1 流水线的概念



在介绍流水线之前，先来了解几个重要概念：

DataFrame：

- 使用Spark SQL中的DataFrame作为数据集，它可以容纳各种数据类型。
- 较之RDD，DataFrame包含了schema 信息，更类似传统数据库中的二维表格。
- 它被ML Pipeline用来存储源数据。
- 例如，DataFrame中的列可以是存储的文本、特征向量、真实标签和预测的标签等

9.5.1 流水线的概念



Transformer：

- 翻译成转换器，是一种可以将一个DataFrame转换为另一个DataFrame的算法。
- 比如一个模型就是一个 Transformer。
- 它可以把一个不包含预测标签的测试数据集 DataFrame 打上标签，转化成另一个包含预测标签的 DataFrame。

- 技术上，Transformer实现了一个方法transform()，它通过附加一个或多个列将一个DataFrame转换为另一个DataFrame

9.5.1 流水线的概念



Estimator： 翻译成估计器或评估器，

- 它是学习算法或在训练数据上的训练方法的概念抽象。在 Pipeline 里通常是被用来操作 DataFrame 数据并生成一个 Transformer。
- 从技术上讲，Estimator实现了一个方法fit()，它接受一个DataFrame并产生一个转换器。
- 比如，一个随机森林算法就是一个 Estimator，它可以调用fit()，通过训练特征数据而得到一个随机森林模型。

9.5.1 流水线的概念



•**PipeLine：**翻译为流水线或者管道。流水线将多个工作流阶段（转换器和估计器）连接在一起，形成机器学习的工作流，并获得结果输出

•**Parameter：Parameter** 被用来设置 Transformer 或者 Estimator 的参数。现在，所有转换器和估计器可共享用于指定参数的公共API。
ParamMap是一组（参数，值）对

9.5.2 流水线工作过程



要构建一个 Pipeline流水线，首先需要定义 Pipeline 中的各个**流水线阶段**PipelineStage（包括转换器和评估器），比如指标提取和转换模型训练等。有了这些处理特定问题的转换器和评估器，就可以按照具体的处理逻辑有序地组织PipelineStages 并创建一个Pipeline

```
>>> pipeline = Pipeline(stages=[stage1,stage2,stage3])
```

然后就可以把训练数据集作为输入参数，调用 Pipeline 实例的 fit 方法来开始以流的方式来处理源训练数据。这个调用会返回一个 PipelineModel 类实例，进而被用来预测测试数据的标签

9.5.2 流水线工作过程

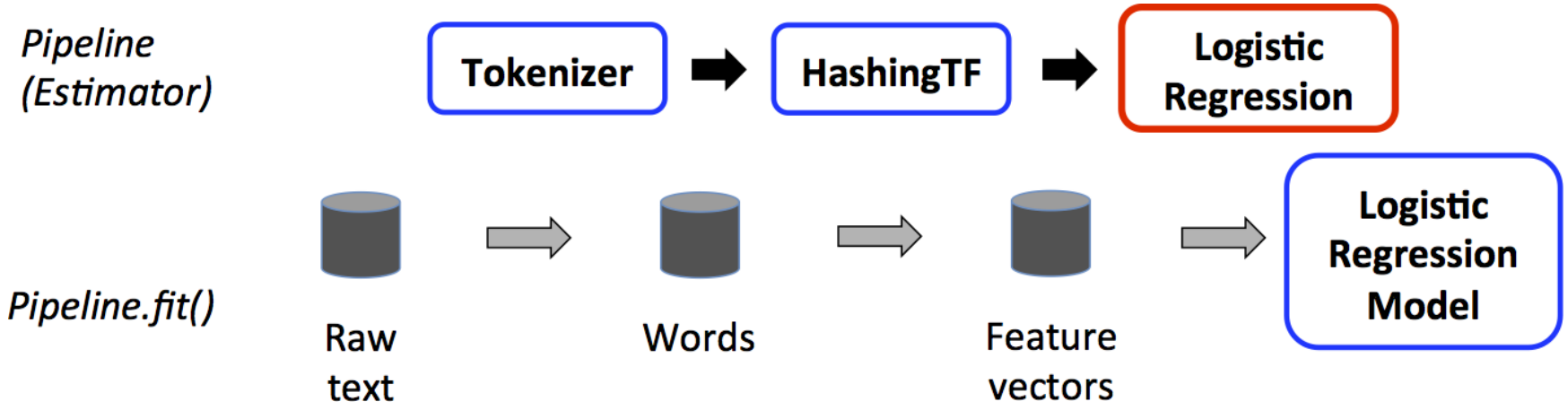


- 流水线的各个阶段按顺序运行，输入的DataFrame在它通过每个阶段时被**转换。**
- 对于转换器阶段，
- 在DataFrame上会调用transform()方法，对于评估器阶段，先调用fit()方法来生成一个转换器，
- 然后在DataFrame上调用该转换器的transform()方法。

9.5.2 流水线工作过程



一个流水线具有3个阶段，前两个阶段（Tokenizer和HashingTF）是转换器，第三个阶段（Logistic Regression）是评估器。图9-2中下面一行表示流经这个流水线的数据，其中，圆柱表示DataFrame。在原始DataFrame上调用Pipeline.fit()方法执行流水线

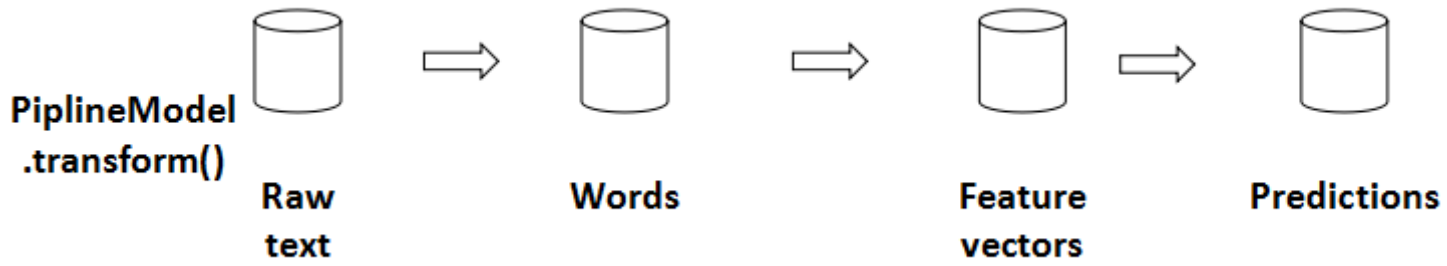


9.5.2 流水线工作过程



流水线本身就是一个评估器，因此，在流水线的fit()方法运行之后，会产生一个流水线模型（PipelineModel），这是一个转换器，可在测试数据的时候使用。





9.5.2 流水线工作过程



MLlib评估器和转换器，使用统一的API指定参数。其中，Param是一个自描述包含文档的命名参数，而ParamMap是一组（参数,值）对。将参数传递给算法主要有以下两种方法：

（1）设置实例的参数。例如，lr是的一个LogisticRegression实例，用lr.setMaxIter(10)进行参数设置以后，可以使lr.fit()至多迭代10次；

（2）传递ParamMap给fit()或transform()函数。ParamMap中的任何参数，将覆盖先前通过set方法指定的参数。

需要特别注意参数同时属于评估器和转换器的特定实例。如果同一个流水线中的两个算法实例（比如LogisticRegression实例lr1和lr2），都需要设置maxIter 参数，则可以建立一个ParamMap，即ParamMap(lr1.maxIter -> 10, lr2.maxIter -> 20)，然后传递给这个流水线。