Spark Machine Learning

CMPT 732, Fall 2017

Recap: Machine Learning

Consider a family of functions or models $y(x;\theta)$ that map input columns (*features* x) to an output column (*prediction* y).

Learning a model means to find a function with *parameters* θ that minimizes a certain measure of error or loss between prediction y and given target *labels* t.

For discrete outputs $y \in \{\text{apples}, \text{oranges}, \ldots\}$ this is called *classification* or *clustering* (if no labels are given). For continuous y (e.g. floating point) this is called *regression*.

Spark ML

- **ML Algorithms:** common learning algorithms
- **Featurization:** feature extraction, transformation, dimensionality reduction, and selection
- **Pipelines:** tools for constructing, evaluating, and tuning ML Pipelines
- **Persistence:** save and load algorithms, models, and Pipelines
- **Utilities:** linear algebra, statistics, data handling

Pipeline Components

- **DataFrame:** from Spark SQL with columns storing text, feature vectors, true labels, and predictions
- **Transformer:** maps one DataFrame to another, e.g., feature extractors, or model predictions
- **Estimator:** E.g., a learning algorithm that trains on a DataFrame and produces a model (Transformer)
- **Pipeline:** chains multiple Transformers and Estimators together
- **Parameter:** common API for Transformers and Estimators

Example

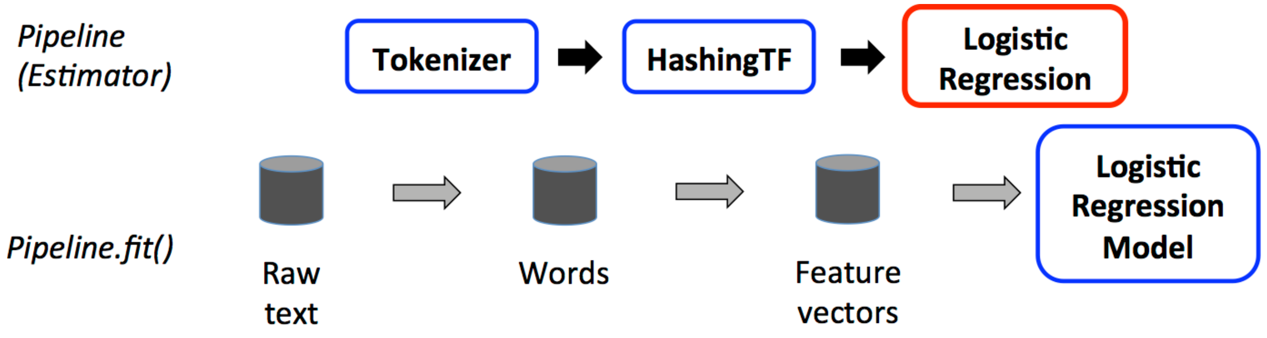
Check out the Logistic Regression example.

Feature Transformers

Let's go over the Feature Transformer list.

Pipeline as Estimator

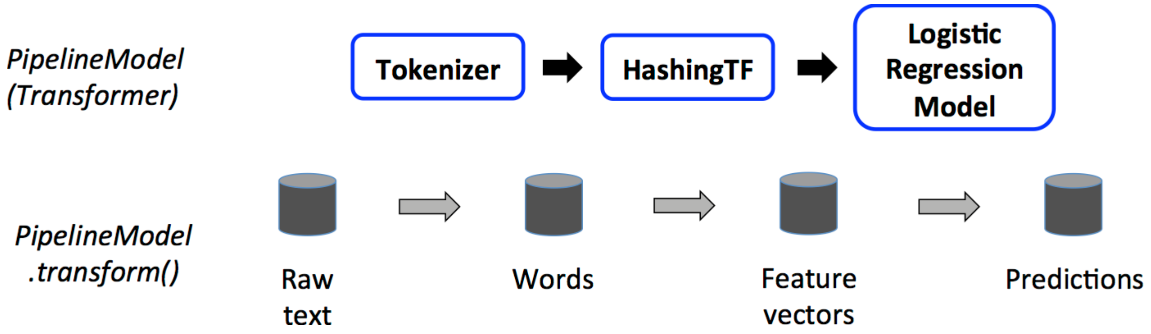
Specifes sequence of stages that are either Transformers or Estimators.



The picture shows a Pipeline at *training time*.

PipelineModel

`Pipeline.fit()` produces a model that is used during *test time*.



ML Algorithms

Lots of learning algorithms to choose from. All of them implement the `Estimator` interface.

Advanced Topics

Regularization

Hyperparameter tuning (Model selection)

Course Notes Home. CMPT 732, Fall 2017.