Machine Learning with MLlib and scikit-learn

Christopher Homa
scikit-learn
Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification
Identifying to which set of categories a new observation belong to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, ...
— Examples

Regression
Predicting a continuous value for a new example.

Applications: Drug response, Stock prices
Algorithms: SVR, ridge regression, Lasso, ...
— Examples

Clustering
Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes.
Algorithms: k-Means, spectral clustering, mean-shift, ...
— Examples

Dimensionality reduction
Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency
Algorithms: PCA, Isomap, non-negative matrix factorization.
— Examples

Model selection
Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: Grid search, cross validation, metrics.
— Examples

Preprocessing
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms
Modules: preprocessing, feature extraction
— Examples
Goal
Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size.

1. Generate datasets
2. Train classifiers
3. Record performance
4. Analyze results
Goal

Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size

Generate datasets

Train classifiers

Record performance

Analyze results
Generate datasets

**Type**
- Binary Classification
- Multiclass Regression
- Regression

**Size**
- Instances
- Features
Generate datasets

Type
- Binary Classification
- Multiclass Regression
- Regression

Size
- Instances [1000, 10]
- Features [2000000, 100]
**Goal**

Compare performance of scikit-learn and MLlib machine learning libraries on datasets of varying size

1. Generate datasets
2. Train classifiers
3. Record performance
4. Analyze results
Choose classifiers

- Stochastic Gradient Descent
- Gradient Boosted Decision Trees
- Random Forests

Match parameters

- Iterations
- Depth
- Most defaults match

Iteratively train classifiers on all datasets and record training times
Goal
Compare performance of sk-learn and MLlib machine learning libraries on datasets of varying size

- Generate datasets
- Train classifiers
- Record performance
- Analyze results
Analyze results
Future Considerations

- Fewer, (much) larger datasets
- Utilize EC2 instances to run sklearn scripts
- Improve data storage