MLbase: A Distributed Machine-learning System

Tim Kraska[1],
Rean Griffith[2],
Ameet Talwalkar[3],
John Duchi[3],
Michael J. Franklin[3],
Michael Jordan[3]

[1] Brown University [2] VMware [3] AMPLab, UC Berkeley

CIDR 2013. 6th Biennial Conference on Innovative Data Systems Research
Presented by Ching-Yuan, Tsai

Outline

1 Introduction

2 Architecture

3 Optimization

4 Conclusion

MLbase

- A distributed machine learning system.
- Focus on user experience.
  - End-user
  - ML researchers

Motivation

- Explosion of abundant data.
- Data in no longer confined to academic researchers.
- Extracting value from such big data is a growing concern.
- The complexity of existing algorithms is overwhelming.
- Layman users may not understand the trade-offs between different learning techniques and parameterization.

Outline

1. Introduction
2. Architecture
3. Optimization
4. Conclusion

Workflow

- User specify simple machine learning task.
- Parse into a logical learning plan (LLP).
  - LLP is the most general workflow to perform the ML task.
- An optimizer translate LLP into physical learning plan (PLP).
  - PLP specifies exactly the parameters and the data sets to be used.
- Distribute PLP onto the worker nodes.
- Return a learned model that can be immediately used.

Architecture

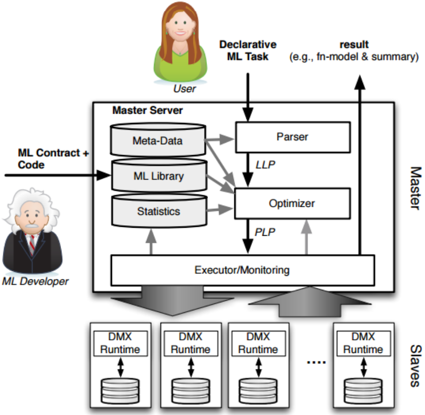


Figure 1: MLbase Architecture
MLbase

Outline

1 Introduction

2 Architecture

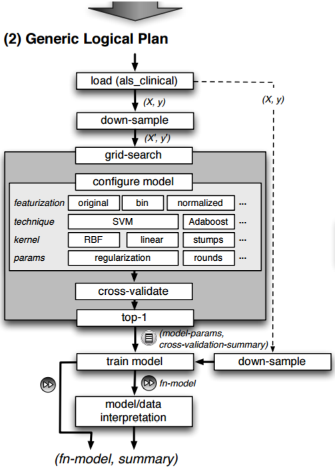
3 Optimization

4 Conclusion

LLP

- Many operations are mapped 1-to-1 to PLP which can translate earlier.
- Insert down-sample stage.

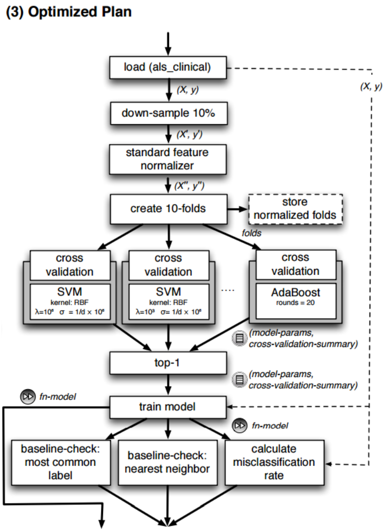
LLP


(2) Generic Logical Plan
load (als_clinical)
(X, y)
down-sample
(X', y')
grid-search
configure model
featurization
original
bin
normalized
...
technique
SVM
Adaboost
...
kernel
RBF
linear
stumps
...
params
regularization
rounds
...
cross-validate
top-1
(model-params, cross-validation-summary)
(X, y)
train model
down-sample
fn-model
model/data interpretation
(fn-model, summary)

PLP

- Transform LLP to concrete parameters and learning algorithm.
- Estimate execution time based on statistical models.
  - Normalized data sets for SVM and AdaBoost.
- Use 10-folds cross validation.

PLP

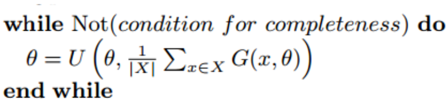

(3) Optimized Plan
load (als_clinical)
(X, y)
(X, y)
down-sample 10%
(X', y')
standard feature normalizer
(X'', y'')
create 10-folds
store normalized folds
folds
cross validation
SVM
kernel: RBF
$\lambda=10^6$ $\sigma = 1/d \times 10^6$
cross validation
SVM
kernel: RBF
$\lambda=10^3$ $\sigma= 1/d \times 10^6$
....
cross validation
AdaBoost
rounds = 20
(model-params, cross-validation-summary)
top-1
(model-params, cross-validation-summary)
fn-model
train model
fn-model
baseline-check: most common label
baseline-check: nearest neighbor
calculate misclassification rate

Return first result, find best in the background

- Become more interactive.
- Reduce the risk of stopping too early.

Good extensibility

- A set of high-level operators to enable ML researchers to implement a wide range ML algorithms without deep systems knowledge.



```
while Not($condition\ for\ completeness$) do
    $\theta = U\left(\theta, \frac{1}{|X|}\sum_{x \in X} G(x, \theta)\right)$
end while
```

Outline

1. Introduction
2. Architecture
3. Optimization
4. Conclusion

Conclusion

- This system aiming to make ML more accessible to non-experts.
- The core of MLbase is its optimizer,which transforms the hight-level ML tasks into the executable ML codes.
- MLbase quickly returns a first quality answer to the user.
- MLbase can improve the result in the background.
- ML developers can constantly add new ML techniques to the system.