Accelerating Machine Learning Development with **mlflow**

Matei Zaharia @matei_zaharia



Talk Overview

Reproducibility matters for scientific researchers, but does it matter for practitioners? → Yes, in fact it matters more!

Designing ML development tools that users will love (and getting reproducibility in the process)



My Background



Research on infrastructure for usable machine learning



Cloud platform for large-scale data analytics and ML



ML development is harder than traditional software development



Traditional Software

Goal: meet a functional specification

Quality depends only on code

Typically pick one software stack

Machine Learning

Goal: optimize a metric (e.g., CTR)Constantly experiment to improve it

Quality depends on input data, training method, tuning params

Compare many libraries, models & algorithms for the same task



Production ML is Even Harder

ML apps must be fed new data to keep working

Design, retraining & inference done by different people

Software must work across many environments





Example

"I build 100s of models/day to lift revenue, using any library: MLlib, PyTorch, R, etc. There's no easy way to see what data went in a model from a week ago and rebuild it."

-- Chief scientist at ad tech firm



Example

"Our company has 100 teams using ML worldwide. We can't share work across them: when a new team tries to run some code, it doesn't even give the same result."

-- Large consumer electronics firm



Traditional Software Development

Became dramatically faster through dev lifecycle tools

• Version control, unit tests, logging, code review, etc

Tools useful enough that developers use them by default

• Even when working alone on a toy project!

How can we do the same for machine learning?





Open source platform to manage ML development

- Works with any ML library, algorithm, language, etc
- Key principle: "API-first" design (use with any code you already have)

Tackles three key problems:

- Experiment tracking: MLflow Tracking
- Reusable workflows: MLflow Projects
- Model packaging: MLflow Models

databricks

Model Development without MLflow

. . .

- data = load_text(file)
- ngrams = extract_ngrams(data, N=n)
- score = compute_accuracy(model)

pickle.dump(model, open("model.pkl"))

	•	👚 mate	ei — -bash — 40×13
For	n=3,	lr=0.9:	accuracy=0.79
For	n=2,	lr=0.1:	accuracy=0.71
For	n=3,	lr=0.2:	accuracy=0.79
For	n=2,	lr=0.5:	accuracy=0.83
For	n=2,	lr=0.9:	accuracy=0.79
For	n=3,	lr=0.1:	accuracy=0.71
For	n=2,	lr=0.2:	accuracy=0.79
For	n=2,	lr=0.5:	accuracy=0.83
For	n=2,	lr=0.9:	accuracy=0.79
For	n=3,	lr=0.1:	accuracy=0.83
For	n=3,	lr=0.2:	accuracy=0.82
For	n=4,	lr=0.5:	accuracy=0.75

What version of my code was this result from?



Experiment Tracking with MLflow

- data = load_text(file)
- ngrams = extract_ngrams(data, N=n)
- score = compute_accuracy(model)

mlflow.log_param("data_file", file)
mlflow.log_param("n", n)
mlflow.log_param("learning_rate", lr)
mlflow.log_metric("score", score)

mlflow.sklearn.log_model(model)

databricks

mlflow GitHub Docs Language Model Experiment ID: 0 Artifact Location: /Users/matei/mlflow/mlruns/0 Search Runs State Active -Search Clear Filter Metrics Filter Params: 10 matching runs Delete Download CSV 🕹 Parameters Metrics Date User Version input_file f1 🔻 2018-10-02 21:53:57 matei lang model.pv e55d56 data.txt 2.0 0.77 0.704 2018-10-02 21:53:55 matei lang_model.py e55d56 data.txt 2.0 4 0.835 0.609 2018-10-02 21:53:48 0.66 0.476 matei lang_model.py e55d56 data.txt 2.0 2 0.468 2018-10-02 21:53:53 matei lang model.pv e55d56 data txt 1.0 0.663 2018-10-02 21:53:49 matei □ lang_model.py e55d56 data.txt 1.0 4 0.902 0.461

Track parameters, metrics, output files & code version

MLflow UI: Inspecting Runs

ml <i>fl</i>	ow								Git	Hub Docs
ι	anguage	e Model								
E	Experiment ID:	0	Artifact Lo	ocation: /Users/matei	/mlflow/mlrun	s/0				
s	earch Runs:	metrics.rmse <	1 and para	ms.model = "tree"		Ē	State:	Act	ive -	Search
F	ilter Params:	alpha, Ir		Filte	r Metrics:	rmse, r2				Clear
1	0 matching runs	Compare	Delete	Download CSV 🛓						
						Pa	arameters		Me	etrics
1	Date 🔻		User	Source	Version	input_file	lr	n	accuracy	f1
	2018-10-0	2 21:53:57	matei	lang_model.py	e55d56	data.txt	2.0	1	0.77	0.704
	2018-10-0	2 21:53:56	matei	🖵 lang_model.py	e55d56	data.txt	1.0	2	0.254	0.222
	2018-10-0	2 21:53:55	matei	lang_model.py	e55d56	data.txt	2.0	4	0.835	0.609
ĺ	2018-10-0	2 21:53:53	matei	□ lang_model.py	e55d56	data.txt	1.0	1	0.663	0.468
	2018-10-0	2 21:53:52	matei	□ lang_model.py	e55d56	data.txt	0.2	4	0.034	0.032
	2018-10-0	2 21:53:51	matei	lang_model.py	e55d56	data.txt	0.1	4	0.177	0.16

databric

MLflow UI: Comparing Runs

ml <i>flow</i>								Git	Hub Docs		
Languag	e Model										
Experiment ID	Experiment ID: 0 Artifact Location: /Users/matei/mlflow/mlruns/0										
Search Runs:	metrics.rmse	s.rmse < 1 and params.model = "tree"				State:	Act	ive -	Search		
Filter Params:	alpha, Ir		Filte	r Metrics:	rmse, r2				Clear		
10 matching rur	S Compare	Delete	Download CSV								
					Pa	arameters		Me	trics		
□ Date ▼		User	Source	Version	input_file	lr	n	accuracy	f1		
2018-10-	02 21:53:57	matei	🗆 lang_model.py	e55d56	data.txt	2.0	1	0.77	0.704		
2018-10-	02 21:53:56	matei	🗆 lang_model.py	e55d56	data.txt	1.0	2	0.254	0.222		
2018-10-	02 21:53:55	matei	🗆 lang_model.py	e55d56	data.txt	2.0	4	0.835	0.609		
2018-10-	02 21:53:53	matei	□ lang_model.py	e55d56	data.txt	1.0	1	0.663	0.468		
2018-10-	02 21:53:52	matei	🖵 lang_model.py	e55d56	data.txt	0.2	4	0.034	0.032		
2018-10-	02 21:53:51	matei	lang_model.py	e55d56	data.txt	0.1	4	0.177	0.16		

databricl

MLflow Tracking: Extensibility

Using a notebook? Log its final state as HTML

Using TensorBoard? Record the logs for each run

Etc.



MLflow Projects: Reusable Workflows

"How can I split my workflow into modular steps?"

"How do I run this workflow that someone else wrote?"



MLflow Projects

Simple packaging format for code + dependencies

my_project/ ├── MLproject	conda_env: conda.yaml
	entry_points:
	main:
-	parameters:
	training_data: path
	<pre>lr: {type: float, default: 0.1}</pre>
	command: python main.py {training_data} {lr}
conda.yaml	
<pre>main.py model.py</pre>	<pre>\$ mlflow run git://<my_project< pre=""></my_project<></pre>
•••	<pre>mlflow.run("git://<my_project></my_project></pre>
databricks	

• • •

Composing Projects

r1 = mlflow.run("ProjectA", params)

if r1 > 0: r2 = mlflow.run("ProjectB", ...) else:

r2 = mlflow.run("ProjectC", ...)

r3 = mlflow.run("ProjectD", r2)

			Paramete			
	Date	User	batch_size	epochs	lr	
ŧ	2018-12-07 12:14:08	matei	512		0.05	
	2018-12-07 12:14:05	matei	512		0.05	
	2018-12-07 12:14:05	matei		20		
	2018-12-07 12:14:05	matei		40		
	2018-12-07 12:14:05	matei		40		
+	2018-12-07 12:12:23	matei	512		0.05	

Combine projects into arbitrary workflows

databricks

MLflow Models: Packaging Models

"How can I reliably pass my model to production apps?"



MLflow Models: Packaging Models



Example MLflow Model

my_model/ ├── MLmodel

run_id: 769915006efd4c4bbd662461
time_created: 2018-06-28T12:34
flavors:
 tensorflow:
 saved_model_dir: estimator
 signature_def_key: predict
 python_function:
 loader_module: mlflow.tensorflow

Usable by tools that understand TensorFlow model format Usable by any tool that can run Python (Docker, Spark, etc!)

- estimator/

--- saved_model.pb
--- variables/

\$ mlflow pyfunc serve -r <run_id>
spark_udf = pyfunc.spark_udf(<run_id>)



MLflow Project Status

Fast-growing open source community

- 70 contributors from >30 companies since we started in June 2018
- External contributions: R API, Docker support, and integrations with PyTorch, H2O, Docker, GCP & more

Try it out: pip install mlflow





"Which MLflow use cases are important to you?"

Creating a standard process to build and maintain ML applications Tracking and sharing results during experimentation Tracking the performance of models in production Enabling reproducible runs in different hardware environments Packaging models for easy deployment to production Ability to deploy same model to multiple platforms databricks



Conclusion

Better ML development tools can improve reproducibility while also improving productivity for ML engineers

Plenty of need for new tools in the industry

Learn about MLflow at mlflow.org



