UNIVERSITAT DE BARCELONA

Master in Fundamental Principles of Data Science

Dr Rohit Kumar

UNIVERSITAT DE BARCELONA

ML Models in Production

Typical flow

Data Science Life Cycle

Data wrangling

Training

Prediction

src: https://www.kdnuggets.com/2019/06/approaches-deploying-machine-learning-production.html by Julien Kervizic

UNIVERSITAT DE BARCELONA

Training

1. One Off training

Where do you save your notebook/code?
Where and how do you save your models?

Models don't necessarily need to be continuously trained in order to be pushed to production. Quite often a model can be just trained ad-hoc by a data-scientist, and pushed to production until its performance deteriorates enough that they are called upon to refresh it.

Example: Image recognition, self driving cars

Training

2. Batch training (continuous learning system)

Not simply retraining the model weights !!

But also feature processing, feature selection, model selections and parameter optimization.

UNIVERSITAT DE BARCELONA

Training

3. Real time training

Incremental Training using “Online Machine Learning Algorithms”

UNIVERSITAT DE BARCELONA

AutoML

AutoWEKA is an approach for the simultaneous selection of a machine learning algorithm and its hyperparameters; combined with the WEKA package it automatically yields good models for a wide variety of data sets.

TPOT is a data-science assistant which optimizes machine learning pipelines using genetic programming. (Good for Python users)

H2O AutoML provides automated model selection and ensembling for the H2O machine learning and data analytics platform. (Nice for Java users)

TransmogrifAI is an AutoML library running on top of Spark.

UNIVERSITAT DE BARCELONA

Airflow

To manage different data workflows

TPOT

To automate Machine Learning

UNIVERSITAT DE BARCELONA

UNIVERSITAT DE BARCELONA

Offcourse there are easy to use fancy platforms !!
(if you have money and do not want to get your hands too dirty !!)

Google AutoML

Microsoft Azure

Azure ML

IBM WATSON

DataRobot

Amazon SageMaker
aws

UNIVERSITAT DE BARCELONA

Saving ML Models

In Python world:
1) Pickle:
2) Joblib: significantly faster on large numpy arrays

In R world:
1) saveRDS/readRDS – can be assigned to a new model variable
2) save/load – returns the same model variable

In Java (H20 world):
1) Pojo : Plan old java object
2) Mojo: Model ObJect, Optimized

UNIVERSITAT DE BARCELONA

Saving ML Models

"**Change**, which is the only constant in life, can be **secured and facilitated** by standardization."

UNIVERSITAT DE BARCELONA

Saving ML Models

ONNX (Open Neural Network Exchange) format: developed by Facebook and Microsoft in collaboration it is an open-source standard for representing deep learning models that will let you transfer them between CNTK, Caffe2 and PyTorch.

Link: https://github.com/onnx/tutorials

UNIVERSITAT DE BARCELONA

Saving ML Models

Framework / Tool	Installation	Tutorial
Caffe	apple/coremltools and onnx/onnxmltools	Example
Caffe2	part of caffe2 package	Example
Chainer	chainer/onnx-chainer	Example
Cognitive Toolkit (CNTK)	built-in	Example
CoreML (Apple)	onnx/onnxmltools	Example
Keras	onnx/keras-onnx	Example
LibSVM	onnx/onnxmltools	Example
LightGBM	onnx/onnxmltools	Example
MATLAB	Deep Learning Toolbox	Example
ML.NET	built-in	Example
MXNet (Apache)	part of mxnet package docs github	Example
PyTorch	part of pytorch package	Example1, Example2, export for Windows ML, Extending support
SciKit-Learn	onnx/sklearn-onnx	Example
SINGA (Apache) - Github (experimental)	built-in	Example
TensorFlow	onnx/tensorflow-onnx	Examples

UNIVERSITAT DE BARCELONA

UNIVERSITAT DE BARCELONA

Saving ML Models

PMML (Predictive model markup language): Developed by Data Mining Group (DMG) a consortium of commercial and open-source data mining companies and supported by IBM, AWS and Google.

PFA (Portable Format for Analytics): it is more flexible (using JSON instead of XML) and also handling data preparation along with the models themselves.

UNIVERSITAT DE BARCELONA

Saving ML Models

MLWriter/MLReader: Inherente to Spark

Supports not only saving a Model but a ML Pipeline.

The result of **save** for pipeline model is a JSON file for metadata while Parquet for model data, e.g. coefficients.

UNIVERSITAT DE BARCELONA

Is a model or Pipeline saved using Apache Spark ML persistence in Spark version X loadable by Spark version Y?

- Major versions: No guarantees, but best-effort.
- Minor and patch versions: Yes; (backwards compatible)
- Note about the format: There are no guarantees for a stable persistence format, but model loading itself is designed to be backwards compatible.

Does a model or Pipeline in Spark version X behave identically in Spark version Y?

- Major versions: No guarantees, but best-effort.
- Minor and patch versions: Identical behavior, except for bug fixes.

UNIVERSITAT DE BARCELONA

Batch vs. Real-time Prediction

Load implications

Infrastructure Implications

Cost Implications

Evaluation Implication

UNIVERSITAT DE BARCELONA

Batch Prediction

Use workflow managers like Airflow, or as simple as crontab etc.

UNIVERSITAT DE BARCELONA

Real Time Prediction

1. Database Trigger based prediction:

When?

If the overall size of your database is fairly small (< 1M user profile) and the update frequency is occasional it can make sense to integrate some of the real-time update process directly within the database.

How?

Postgres possess an integration that allows to run Python code as functions or stored procedure called PL/Python. This implementation has access to all the libraries that are part of the PYTHONPATH, and as such are able to use libraries such as Pandas and SKlearn to run some operations.

Real Time Prediction

Example Case

UNIVERSITAT DE BARCELONA

Real Time Prediction

1. Database Trigger based prediction

- PostGres supprts native Python integration and can run Python script on trigger.
- Ms SQL Server has **Machine Learning Services (In-Database)**
- Teradata can run R/Python script through an external script command.
- Oracle supports PMML model through its data mining extension.

UNIVERSITAT DE BARCELONA

Real Time Prediction

2. Message queue based prediction service

- A combination of Kafka for message queue and Spark Streaming for processing.
- Azure eventhub with Azure functions
- Google Pub/Sub with Apache Beam.

UNIVERSITAT DE BARCELONA

Real Time Prediction

Example Case

UNIVERSITAT DE BARCELONA

Real Time Prediction

3. Web services based deployment

- Get unique ids and let web service pull related data from backend system and make prediction.
- Get the complete payload with data and make the prediction.
- A combination of both.

The web service or rest api based deployment models are more popular as they are very easy to scale and also can be used to be integrated in different frontend applications like mobile, web or desktop.

UNIVERSITAT DE BARCELONA

Real Time Prediction

3.1 Using Functions!!

+ ML Net

UNIVERSITAT DE BARCELONA

```
def main(req: func.HttpRequest) -> func.HttpResponse:
    logging.info('Python HTTP trigger function processed a request.')
    if req.body:
        try:
            logging.info("Converting Request to DataFrame")
            req_body = req.get_json()
            df_body  = pd.DataFrame([req_body])

            logging.info("Loadding the Prediction Model")
            filename = "model.pckl"
            loaded_model = joblib.load(filename)
            # Features names need to have been added to the pickled model
            feature_names = loaded_model.feature_names
            # subselect only the feature names

            logging.info("Subselecting the dataframe")
            df_subselect = df_body[feature_names]

            logging.info("Predicting the Probability")
            result = loaded_model.predict_proba(df_subselect)
            # We are looking at the probba prediction for class 1
            prediction = result[0][1]

            return func.HttpResponse("{prediction}".format(prediction=predicti

        except ValueError:
            pass
    else:
        return func.HttpResponse(
             "Please pass a name on the query string or in the request body",
status_code=400
        )
```

Can any one tell me what is the issue with this approach?

UNIVERSITAT DE BARCELONA

Real Time Prediction

3.2 Using Container

docker

kubernetes

- Ununiform environments across models.
- Ununiform library requirements across models.
- Ununiform resource requirements across models.
- Scaling at model level

UNIVERSITAT DE BARCELONA

Real Time Prediction

3.3 Directly In-App

ML Kit

Google ML Kit or the likes of Caffe2or Apple's core ML allows to leverage models within native applications

JavaScript

Tensorflow.js or ONNXJs allow running models directly on browser

UNIVERSITAT DE BARCELONA

Real Time Prediction

3.4 Using Notebooks!!!!

data iku

databricks

UNIVERSITAT DE BARCELONA

So Far So Good!!

UNIVERSITAT DE BARCELONA

Why ML Life Cycle is more complicated than SDLC!

- Variety of tools and packages at disposal.
- Hard to track useful experiments.
- Reproducibility is hard.
- Deploying is hard.

UNIVERSITAT DE BARCELONA

ML Life Cycle Management Platforms

Azure ML services (launched in 2015)

What's possible with AML?

Using Azure Machine Learning service, you can track your models as you build, train, deploy, and manage them at cloud scale.

Run & Monitor Experiments

Submit Experiments for training and automatically track their progress and view logs.

Register Models

Save scoring logic operations into models to create Docker Images and Deployments.

Build Images

Quickly create Docker images that encapsulate models, scripts, and any associated files.

Deploy Models

Send scoring requests to web services in Azure Container Instances, Azure Kubernetes Service, or field programmable gate arrays (FPGA).

Create Pipeline

Pipelines are used to build, optimize, and manage machine learning workflows.

UNIVERSITAT DE BARCELONA

ML Life Cycle Management Platforms

Amazon SageMaker (launched in 2017)

UNIVERSITAT DE BARCELONA

ML Life Cycle Management Platforms

Google AI (launched in 2018)

UNIVERSITAT DE BARCELONA

ML Life Cycle Management Platforms

Cloudera DS Workbench (launched in 2018)

UNIVERSITAT DE BARCELONA

ML Life Cycle Management Platforms

Databricks Managed ML Flow (launched in 2019) : ML Flow in itself is a library and not a hosted platform.