Machine Learning as a Service

Learning the art of building data-driven products

Workshop @ The Fifth Elephant 2017

Amit Kapoor amitkaps.com
Anand Chitpothu anandology.com
Bargava Subramanian bargava.com

Getting Started

- Download the Repo: https://github.com/amitkaps/full-stack-data-science
- Finish installation
- Run jupyter notebook in the console

data scientist: *the people who are building products from data*

What is required to know?

- *Data Management*
- *Modelling & Prototyping*
- *Product Design*
- *Data Engineering*

"Jack of all trades, master of none, though oft times better than master of one."

The Unicorn Skillset

- *Data Management*: data ingestion & wrangling
- *Modelling & Prototyping*: statistics, visualisation, machine learning
- *Product Design*: data narrative, dashboards, applications
- *Data Engineering*: data pipelines, cloud infrastructure

Motivation for the Workshop

- Solve a business problem.
- Understand the end-to-end MLaaS approach
- Build a data-driven ML application

Approach

— Simple and intuitive

— Go wide vs. go deep

— Practical and scalable

Outline - Day 1

Session 1: **Introduction and Concepts**

- Approach for building ML products
- Problem definition and dataset
- Build your first ML Model (Part 1)

Session 2: **Build a Simple ML Service**

- Build your first ML Model (Part 2)
- Concept of ML Service
- Deploy your first ML Service - localhost API

Outline - Day 1 (contd.)

Session 3: **Build & Evaluate ML Models**

- Feature Engineering
- Build your second ML model
- ML model evaluation (metrics, validation)

Session 4: **Practice Session**

- Practice problem overview and data
- Build your ML Model
- Build your API

Outline - Day 2

Session 5: **Build a Simple Dashboard**

- Concept of Dashboard design
- Create your first dashboard
- Integrate ML model API with dashboard

Session 6: **Deploy to cloud**

- Get started with cloud server setup
- Deploy your ML service as cloud API
- Deploy your dashboard as cloud service

Outline - Day 2 (contd.)

Session 7: **Repeatable ML as a Service**

- Build data pipelines
- Update model, API and dashboard
- Schedule ML as as Service process

Session 8: **Practice Session & Wrap-up**

- Deploy on cloud - dashboard and API
- Best practices and challenges in building ML service
- Where to go from here

Schedule

08:45 to 09:30 : *Check-in & Breakfast*

09:30 to 11:00 : **Session 1**

11:00 to 11:20 : *Coffee break*

11:20 to 13:00 : **Session 2**

13:00 to 14:00 : *Lunch break*

14:00 to 15:40 : **Session 3**

15:40 to 16:00 : *Coffee break*

16:00 to 17:10 : **Session 4**

Data-Driven Learning

Two cases / dataset in the Workshop

- Loan Default
- People Attrition

Metaphor

- A start-up providing loans to the consumer
- Running for the last few years
- Now planning to adopt a data-driven lens

What are the **type of questions** you can ask?

Type of Questions

- What is the trend of loan defaults?
- Do older customers have more loan defaults?
- Which customer is likely to have a loan default?
- Why do customers default on their loan?

Type of Questions

- Descriptive
- Inquisitive
- Predictive
- Causal

Data-driven Analytics

- **Descriptive**: Understand Pattern, Trends, Outlier
- **Inquisitive**: Conduct Hypothesis Testing
- **Predictive**: Make a prediction
- **Causal**: Establish a causal link

Prediction Challenge

It's tough to make predictions, especially about the future.

— Yogi Berra

How to make a Prediction?

- **Human Learning**: Make a *Judgement*
- **Machine Programmed**: Create explicit *Rules*
- **Machine Learning**: Learn from *Data*

Machine Learning (ML)

[Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed.
— *Arthur Samuel*

Machine learning is the study of computer algorithm that improve automatically through experience
— *Tom Mitchell*

Machine Learning: Essense

- A pattern exists
- It cannot be pinned down mathematically
- Have data on it to learn from

"Use a set of observations (data) to uncover an underlying process"

ML as a Service (MLaaS) Approach

```
FRAME  ⟶  ACQUIRE  ⟶  REFINE ⟶
                              \
                          TRANSFORM ⟵
                              ↑          ↘
                              |          EXPLORE
                              ↓          ↗
                            MODEL    ⟵
                             /
INTERACT ⟵  BUILD ⟵  DEPLOY ⟵
```

MLaaS Approach

- *Frame*: Problem definition
- *Acquire*: Data ingestion
- *Refine*: Data wrangling
- *Transform*: Feature creation
- *Explore*: Feature selection
- *Model*: Model creation & selection
- *Deploy*: Model deployment
- *Build*: Application building
- *Interact*: User interaction

ML Theory: Data Types

- What are the types of data on which we are learning?
- Can you give example of say measuring temperature?

Data Types e.g. Temperature

- **Categorical**
  - *Nominal*: Burned, Not Burned
  - *Ordinal*: Hot, Warm, Cold
- **Continuous**
  - *Interval*: 30 °C, 40 °C, 80 °C
  - *Ratio*: 30 K, 40 K, 50 K

Data Types - Operations

- **Categorical**
  - *Nominal*: = , !=
  - *Ordinal*: =, !=, >, <
- **Continuous**
  - *Interval*: =, !=, >, <, -, % of diff
  - *Ratio*: =, !=, >, <, -, +, %

Case: Loan Default Prediction

Application Attributes

- **age**: age of the applicant
- **income**: annual income of the applicant
- **year**: no. of years of employment
- **ownership**: type of house owned
- **amount** : amount of loan requested by the applicant

Behavioural Attributes:

- **grade**: credit grade of the applicant

Question - whether the applicant will **default** or not?

Historical Data

default	amount	grade	years	ownership	income	age
0	1,000	B	2.00	RENT	19,200	24
1	6,500	A	2.00	MORTGAGE	66,000	28
0	2,400	A	2.00	RENT	60,000	36
0	10,000	C	3.00	RENT	62,000	24
1	4,000	C	2.00	RENT	20,000	28

Data Types

- **Categorical**
  - *Nominal*: home owner [rent, own, mortgage]
  - *Ordinal*: credit grade [A > B > C > D > E]
- **Continuous**
  - *Interval*: approval date [20/04/16, 19/11/15]
  - *Ratio*: loan amount [3000, 10000]

ML Terminology

Features: $\mathbf{x}$

- age, income, years, ownership, grade, amount

Target: y

- default

Training Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_n, y_n)$

- historical records

ML Paradigm: Supervised

Given a set of **feature** $\mathbf{x}$, to predict the value of **target** y

Learning Paradigm: **Supervised**

- If y is *continuous* - **Regression**
- If y is *categorical* - **Classification**

Simple MLaaS Example (1/4)

```
#Load the libraries and configuration
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

from sklearn import tree
from sklearn.externals import joblib
from firefly.client import Client
```

Simple MLaaS Example (2/4)

```python
#Frame - predict loan default probability

#Acquire - load historical data
df = pd.read_csv("../data/historical_loan.csv")

#Refine - drop NaN values
df.dropna(axis=0, inplace=True)

#Transform - log scale
df['log_age'] = np.log(df.age)
df['log_income'] = np.log(df.income)
```

Simple MLaaS Example (3/4)

```
#Model - build a tree classifier
X = df.loc[:,('age', 'income')]
y = df.loc[:,'default']
clf = tree.DecisionTreeClassifier(max_depth=10).fit(X,y)
joblib.dump(clf, "clf.pkl")

#Build - the model API
%%file simple.py
import numpy as np
from sklearn.externals import joblib
model = joblib.load("clf.pkl")
```

Simple MLaaS Example (4/4)

```
def predict(age, amount):
    features = [age, amount]
    prob0, prob1 = model.predict_proba([features])[0]
    return prob1

#Deploy - the ML API
! firefly simple.predict

#Interact - get predictions using API
simple = Client("http://127.0.0.1:8000")
simple.predict(age=28, amount=10000)
```

Frame

Variables

- age, income, years, ownership, grade, amount, default and interest

— What are the **Features**: $\mathbf{x}$?

— What are the **Target**: y

Frame

Features: $\mathbf{x}$

- age
- income
- years
- ownership
- grade
- amount

Target: y

- default

Acquire

- Simple! Just read the data from csv file

Refine - Missing Value

- **REMOVE** - NAN rows
- **IMPUTATION** - Replace them with something?
  - Mean
  - Median
  - Fixed Number - Domain Relevant
  - High Number (999) - Issue with modelling
- **BINNING** - Categorical variable and "Missing becomes a category*
- **DOMAIN SPECIFIC** - Entry error, pipeline, etc.

Refine - Outlier Treatment

- What is an outlier?
- Descriptive Plots
  - Histogram
  - Box-Plot
- Measuring
  - Z-score
  - Modified Z-score > 3.5

    where modified Z-score = 0.6745 * (x - x_median) / MAD

Explore

- Single Variable Exploration
- Dual Variable Exploration
- Multi Variable Exploration

Transform

Encodings e.g.

- One Hot Encoding
- Label Encoding

Feature Transformation e.g.

- Log Transform
- Sqrt Transform

Model Creation

Types of ML Model

- Linear
- Tree-Based
- Neural Network

Choosing a Model

1. Interpretability
2. Run-time
3. Model complexity
4. Scalability

Tree Based Models

— Easy to interpret

— Little data preparation

— Scales well with data

— White-box model

— Instability - changing variables, altering sequence

— Overfitting

Ensemble Models

Bagging

- Also called bootstrap aggregation, reduces variance
- Uses decision trees and uses a model averaging approach

Random Forest

- Combines bagging idea and random selection of features.
- Similar to decision trees are constructed – but at each split, a random subset of features is used.

Model Selection

How to choose between competing model?

— Error Metric (Business Decision)

— Hyper-Parameter Tuning

— Cross-Validation

If you torture the data enough, it will confess.

— Ronald Case

Challenges

- Data Snooping
- Selection Bias
- Survivor Bias
- Omitted Variable Bias
- Black-box model Vs White-Box model
- Adherence to regulations

Machine Learning as a Service

Learning the art of building data-driven products

Workshop @ The Fifth Elephant 2017

Amit Kapoor amitkaps.com
Anand Chitpothu anandology.com
Bargava Subramanian bargava.com