Practical Machine Learning

The art of building data-driven Machine Learning applications

Amit Kapoor
Anand Chitipothu
Bargava Subramanian

Overview

One of the common pain points that we have come across in big organisations is the last-mile delivery of data science applications. One common delivery vehicle is to create dashboards (BI). But the one, that's very useful and neglected more often than not, is to create APIs and provide seamless integration with other applications within the company. This requires you to have a basic understanding of machine learning, server-side programming and front-end application.

In this workshop, you would learn how to build a seamless end-to-end data driven application - Data Exploration, Machine Learning Model, RESTful API and Web Application - to solve a business prediction problem.

Duration & Format

This is a hands-on workshop and the participants will build and deploy end-to-end ML application. There will be emphasis on actual building and will be roughly 70% coding and 30% theory.

The duration for this workshop is 3-days. Each day will have 4 sessions each.

Motivation for the Workshop

— Solve a business problem.

— Understand the end-to-end ML approach

— Develop a deeper understanding of ML

— Build a data-driven ML application

Approach

— Simple and intuitive

— Go wide vs. go deep

— Practical and scalable

Target Audience

A programmer but not a data science practioner:

A programmer with experience in server-side or front-end development and maybe has some familiarity with doing data analysis. You could be looking to transition in to building data driven products or a create a richer product experience with data.

A data science practioner but not a programmer:

A data science with some experience in doing data analysis, preferably in a scripting language (R/Python/Scala), but wants to get a deeper and a more applied perspective on creating data driven products.

Pre-requisites

Participants should be comfortable with **Python programming language** and have **prior experience** with using Python for Data Science.

Specifically, some experience in working with the following python libraries will help the participants to learn faster:

— *jupyter*: For doing literate programming in notebooks

— *numpy*: For scientific computation

— *pandas*: For data wrangling of tabular data (dataframes)

— *matplotlib*: for data visualisation

— *scikit-learn*: For building machine learning models

Software Requirements

We will be using Python data stack for the workshop. Please install Ananconda for Python 3.5 for the workshop. Additional requirement will be communicated to participants.

Data-Driven Learning

Four cases / datasets will be available during the workshop.

1. Peeling the Onion (Regression)

2. Loan Default (Classification)

3. Propensity to Buy (Classification)

4. Employee Attrition (Classification)

We will focus on two datasets during the workshop and the rest two will be used for practice sessions.

Outline - Day 1

Session 1: **Introduction and Concepts**

- Approach for building ML products - the process

- Problem definition and dataset

- Framing a problem (Case #1)

Session 2: **Data Wrangling**

- Concept of Tidy Data

- Acquire, Refine and Transform the Data

- Conduct Data Wrangling (Case #1)

Outline - Day 1 (contd.)

Session 3: **Explore and Visualise**

- Focus of Visual Exploration & Hypotheses Generation
- Introduction to Grammar of Graphics
- Make Single/Dual and Multi-Dim Vis (Case #1)

Session 4: **Linear ML Model**

- Practical Machine Learning
- Supervised Learning - Linear Model
- Build a Linear ML Model (Case #1)

Outline - Day 2

Session 5: **Tree-based ML Models**

- Deeper dive into Machine Learning

- Supervised Learning - Tree Based Model

- Build a Tree-Based ML Model (Case #2)

Session 6: **Evaluate & Regularise ML Models**

- ML model evaluation and metrics

- Concept of regularisation

- Evaluate a regularised ML model (Case #2)

Outline - Day 2 (contd.)

Session 7: **Model Selection, Ensemble**

- ML model validation & selection

- Concept of ensemble model

- Build an Random Forest ML model (Case #3)

Session 8: **Build a Simple Dashboard**

- Concept of dashboard design

- Create your first dashboard

- Integrate ML model with dashboard

Outline - Day 3

Session 9: **Build a Simple ML Service**

- Concept of ML Service

- Rest API and design

- Deploy your ML Service - localhost API

Session 10: **Deploy to cloud**

- Get started with cloud server setup

- Deploy your ML service as cloud API

- Deploy your dashboard as cloud service

Outline - Day 3 (contd.)

Session 11: **Repeatable ML as a Service**

- Setting up periodic tasks

- Scheduling re-training of model every day

Session 12: **Practice Session & Wrap-up**

- Best practices in building ML service

- Challenges in managing ML in production

- Where to go from here

Trainers

Amit Kapoor

Amit Kapoor teaches the craft of telling visual stories with data. He conducts workshops and trainings on Data Science in Python and R, as well as on Data Visualisation topics. His background is in strategy consulting having worked with AT Kearney in India, then with Booz & Company in Europe and more recently for startups in Bangalore. He did his B.Tech in Mechanical Engineering from IIT, Delhi and PGDM (MBA) from IIM, Ahmedabad. You can find more about him at http://amitkaps.com/ and tweet him at @amitkaps.

Anand Chitipothu

Anand has been crafting beautiful software since a decade and half. He's now building a data science platform, <u>rorodata</u>, which he recently co-founded. He regularly conducts advanced programming courses through <u>Pipal Academy</u>. He is co-author of web.py, a micro web framework in Python. He has worked at Strand Life Sciences and Internet Archive.

Bargava Subramanian

Bargava Subramanian is a practicing Data Scientist. He has 14 years of experience delivering business analytics solutions to Investment Banks, Entertainment Studios and High-Tech companies. He has given talks and conducted workshops on Data Science, Machine Learning, Deep Learning and Optimization in Python and R. He has a Masters in Statistics from University of Maryland, College Park, USA. He is an ardent NBA fan. You can tweet to him at @bargava.