MODERN TECHNOLOGIES FOR MACHINE LEARNING AND MINING OF MASSIVE DATASETS

express course

Borisyak M.A., Ustyuzhanin A.E., Ignatov D.I

November 9, 2015

National Research University Higher School of Economics (HSE)
Laboratory of Methods for Big Data Analysis

What is this course about?

This is a practical course.

Two parts:

1. Ordinary Machine Learning
2. Machine Learning for Big Data

We will learn how to use:

- Python stack: `numpy`, `scipy`, `sklearn`
- Spark stack: `Apache Spark`, `Apache Spark MLlib`, ...

What is this course about?

1. NumPy crash course
2. SciPy
3. Sklearn:
   - supervised learning
   - unsupervised learning
   - model selection, evaluation, feature selection etc
4. introduction to Scala
5. Apache Spark
6. Apache Spark MLlib:
7. Apache Spark Streaming
8. Apache Spark GraphX (?)

Course structure

Each ‘lecture’:

1. introduction into the topic
2. some examples
3. practice and exercises (optional)

3 home assignments:

1. 2 $\times$ 30 min report
2. project

work done	the final grade
2 reports + project	excellent
report + project	good
project	satisfactory

Report

Machine Learning method

Choose a state-of-the-art Machine Learning method and describe:

- advantages/disadvantages, use cases, domains
- implementations, examples of usage
- distributed implementations, applications for Big Data, examples
- modifications, related algorithms (e.g. PCA Trees)

Report

Domain

Choose a domain and describe:

- domain peculiarity, methods for preprocessing, feature extraction
- overview of state-of-the-art algorithms for this domain (especially distributed algorithms)
- implementations, libraries, examples of problems and algorithms

Possible domains: physics, computer vision, music, films, natural languages, document search

Report

Machine Learning library

Choose a modern machine learning library (preferably Python or Scala/Java):

- implemented algorithms
- quick guide to the library
- examples

Suggested libraries: XGBoost, uBoost, Pybrain, TMVA, Theanets, Neurolab, Caffe, NLTK

Note: some libraries are design for a particular domain (e.g. NLTK), watch for possible duplication in reports

Contacts

Maxim Borisyak:

email mborisyak@hse.ru

phone +7 (985) 898 53 52

Questions?