Modern technologies for Machine Learning and Mining of Massive Datasets

course syllabus

Borisyak M.A., Ustyuzhanin A.E., Ignatov D.I.
National Research University Higher School of Economics (HSE)
Laboratory of Methods for Big Data Analysis

November 9, 2015

Course syllabus

In real-world Machine Learning practical experience might be somewhat independent from theory: people with good theoretical background but without enough practice are usually unable to solve real-world problems, while a lot of people spent tremendous amount of time in trial-error learning process successfully win prizes in Data Analysis competitions (like ones on Kaggle) making their solutions based on intuition rather than on scientific considerations.

The main goal of the course is to provide enough experience to make students familiar with Machine Learning in practice. Given theoretical background after this course students are expected to be able to solve typical Machine Learning and Big Data problems.

The course is expected to be 28 academic hours long. After each topic estimated number of hours is indicated in parentheses.

Introduction

- course summary: what is this course about and what it is not
- organization moments

Basics before Machine Learning (2)

1. Introduction to Jupyter notebook:
    - basics

    - plotting (`matplotlib`)

2. Crush `numpy` course
3. `Scipy`
    - linear algebra: basics, pseudo-inverse, SVD
    - optimization
    - statistics:
        - distributions: building, sampling, fitting
        - hypotheses checking: χ^2-, KS-, t- tests
        - kernel density estimation

Sklearn etc (8)

- ideology: 'define model, fit, predict, validate, tune'
- supervised learning (4):
    - Naive Bayes
    - kNN (e.g. kNN with Mahalanobis, custom distances)
    - LSR, Ridge regression, Lasso, elastic net
    - LDA, QDA, dimensionality reduction via DA
    - logistic regression
    - SVM, RKHS
    - Decision Trees, ensembling: Random Forest, Gradient Boosting over Trees (+XGBoost)
    - SGD: custom learning, LSR, SVM etc via SGD (also see Big Data section)
- unsupervised learning (3):
    - clustering: k-Means, k-Medoids, DBSCAN
    - EM, Gaussian Mixtures
    - Matrix Factorization: SVD, PCA, NNMF
    - Novelty/Outlier detection: one-class SVM, PCA
- model selection, evaluation, feature selection (1):
    - cross-validation, LOO
    - grid search
    - model evaluation
- Artificial Neural Networks (?)

Mining of Massive Datasets (18)

- introduction (1):
  - Big Data as Big rubbish heap
  - history: MPI, Apache MapReduce etc
  - Alternatives: multi-threading, GPU, `IPython.parallel`, MPI
  - GRID vs. Map-Reduce vs. Dataflow/DAG
- introduction to Functional Programming and Scala/Python (2)
- Apache Spark (3)
  - state-of-the-art platform for distributed computations
  - basics: RDD, creation, transformations, actions, saving
  - execution, debugging, tuning
  - caching
  - first approximation of the internal mechanisms
- Apache Spark MLLib (8):
  - Basic statistics
  - naive Bayes, SVM, logistic regression, Random Forest, Gradient Boosting
  - k-Means, Gaussian Mixtures
  - PCA
  - recommendation system via ALS
  - NLP: tf-idf, word2vec
  - feature selection
  - evaluation
  - advanced: distributed SGD
- Apache Spark Streaming (2):
  - basics: DStream, transformations, sources
  - naive Bayes on Twitter
  - general ways of making streaming algorithms
- *Optional* Apache Spark GraphX (2)
- *Optional* Distributed data bases (1):
  - Cassandra, Elastic search

Homework and examples

Students will have 3 home assignments.

Since there is not so many students, each should prepare 2 20-30 minutes reports about either:

1. a state-of-the-art Machine Learning method:
   - advantages/disadvantages, use cases, successfully solved problems
   - modifications of the method
   - distributed implementations, applications for Big Data
   - related algorithms and their combinations (e.g. PCA Trees)
2. a domain:
   - domain peculiarity, methods for preprocessing, feature extraction
   - overview of state-of-the-art algorithms for this domain (especially distributed algorithms)
3. a machine learning library:
   - implemented algorithms
   - quick guide to the library
   - examples

Some domains are enumerated below:

- physics
- computer vision
- music, films
- natural languages
- document search

Suggested libraries:

- XGBoost
- uBoost
- Pybrain
- TMVA
- Theanets
- Neurolab

- Caffe
- NLTK

Students should base their reports on recent studies and the reports should be as close as possible to the scientific survey.

The last assignment is an individual project in which students are offered to choose a database and build a predictor for this database applying technologies studied in this course.

Example datasets

- MNIST (almost endless potential for exploration)
- Million Songs Dataset (for collaborative filtering)
- COMET track recognition
- Large Data Sets Repository

Additional dataset sourses

- UC Irvine ML repository
- Kaggle
- Datasets from [6]
- StatSci's list of data sources
- PhysioBank Archive Index

Recommended literature

[1] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge University Press, 2014.

[2] Scikit learn documentation. `http://scikit-learn.org/stable/documentation.html`, 2015.

[3] Martin Odersky, Lex Spoon, and Bill Venners. *Programming in scala.* Artima Inc, 2008.

[4] Spark documentation. `https://spark.apache.org/docs/latest/`, 2015.

[5] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[6] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction.* 2009.