Machine Learning Open Source Software and Benchmark Repository

Mikio L. Braun
TU Berlin
mikiobraun.de

October 7, 2010
Validation in Statistics and Machine Learning
WIAS Berlin

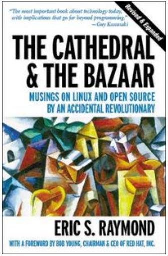
Validation in Machine Learning and Statistics

Machine Learning	Statistics
Solve hard computational tasks	Obtain scientific insight from data

→ Data and validation important both for ML and statistics, but for different reasons:

- **Machine Learning:** to share learning problems and compare existing methods.
- **Statistics:** to ensure that scientific insights are correct.

“Open Source”


"The most important book about technology today, with implications that go far beyond programming." —Guy Kawasaki
Revised & Expanded
THE CATHEDRAL & THE BAZAAR
MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY
ERIC S. RAYMOND
WITH A FOREWORD BY BOB YOUNG, CHAIRMAN & CEO OF RED HAT, INC.

- Actually, open source is about a collaborative process to develop software (not unlike science!)
- Infrastructure: Source code revision system, bug trackers, mailing lists, discussion forums... .
- Once you release your code, you should enter this process.
- Opportunity for much faster interaction with “users”!

Legal Implications of Reproducible Research

Victoria Stodden, *"Reproducible Research in Computational Science: Problems and Solutions For Data and Code Sharing"* http://videolectures.net/icml2010_stodden_rric/

- Releasing Source Code is not the same as Open Source!
- By default, "original expressions of ideas" is copyrighted (protects reproduction and derivative works, limited lifetime)
- Data? "Raw facts" not copyrightable. "Original selection and arrangement" is. Best option is to release to the public domain.

Open Source Licenses

- Main purpose: Allow derived work.
- Differences: Derived work must also be released as open source, commercial use allowed, patents allowed, etc.
- Applying such a license: As easy as downloading the code and adding it as a file called LICENSE, adding links to that code.

More on Licenses

- “Classical” source code license:
  - GNU Public License
  - BSD license,
  - Apache 2.0
  - “Lesser” GPL,
  - Affero GPL,
  - see www.opensource.org/licenses/alphabetical
- Creative Commons (creativecommons.org):
  - CC BY (attribution)
  - CC NC (no commercial use)
  - CC ND (no derived works)
  - CC SA (derived work must use same license)
- Public Domain (CC0): Waive all rights

The Reproducible Research Standard

Victoria Stodden, *"Enabling Reproducible Research: Open Licensing For Scientific Innovation"* International Journal of Communications Law and Policy, Issue 13, 2009.

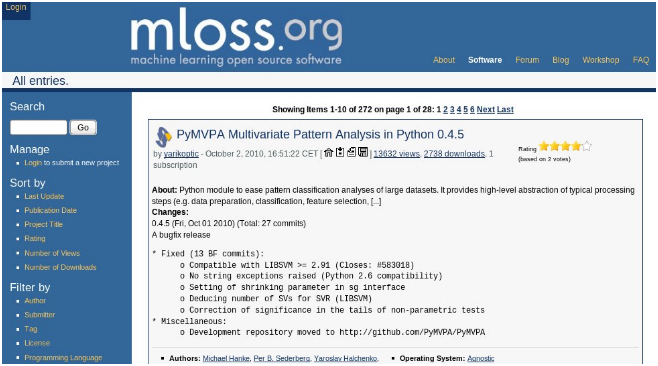
- Remove copyright's barrier to reproducible research,
- Realign the IP framework with longstanding scientific norms. A suite of license recommendations for computational science:

  1. Release media components (text, figures) under CC BY,
  2. Release code components under Modified BSD or similar,
  3. Release data to public domain (CC0) or attach an attribution license.

Machine Learning Open Source Software

- MLOSS: NIPS Workshop 2006
- Position paper: S. Sonnenburg, M. L. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K.-R. Mller, F. Pereira, C. E. Rasmussen, G. Rtsch, B. Schlkopf, A. Smola, P. Vincent, J. Weston, R. Williamson, *The Need for Open Source Software in Machine Learning*, Journal of Machine Learning Research, 8(Oct):2443-2466, 2007
- mloss.org: Machine Learning Open Source Software
- JMLR MLOSS Track
- mldata.org: Machine Learning Data Set Repository

mloss.org



All entries.

- Login to submit a new project

Sort by
- Last Update
- Publication Date
- Project Title
- Rating
- Number of Views
- Number of Downloads

Filter by
- Author
- Submitter
- Tag
- License
- Programming Language

Showing Items 1-10 of 272 on page 1 of 28: 1 2 3 4 5 6 Next Last

PyMVPA Multivariate Pattern Analysis in Python 0.4.5

by yarikoptic - October 2, 2010, 16:51:22 CET [ ] 13632 views, 2738 downloads, 1 subscription

Rating ★★★★☆ (based on 2 votes)

About: Python module to ease pattern classification analyses of large datasets. It provides high-level abstraction of typical processing steps (e.g. data preparation, classification, feature selection, [...]

Changes:

0.4.5 (Fri, Oct 01 2010) (Total: 27 commits)

A bugfix release

```
* Fixed (13 BF commits):
      o Compatible with LIBSVM >= 2.91 (Closes: #583018)
      o No string exceptions raised (Python 2.6 compatibility)
      o Setting of shrinking parameter in sg interface
      o Deducing number of SVs for SVR (LIBSVM)
      o Correction of significance in the tails of non-parametric tests
* Miscellaneous:
      o Development repository moved to http://github.com/PyMVPA/PyMVPA
```

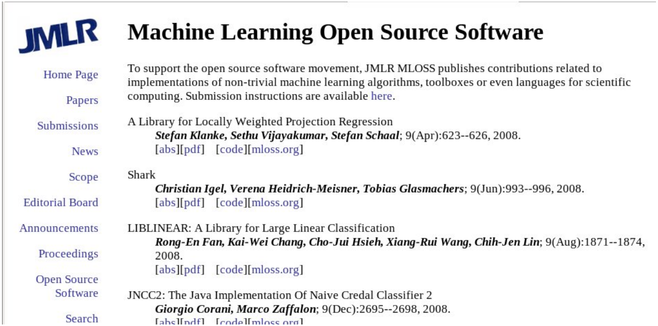
- **Authors:** Michael Hanke, Per B. Sederberg, Yaroslav Halchenko,
- **Operating System:** Agnostic

- Open directory of machine learning related open source software projects.
- 272 projects registered so far.
- About 250 visitors per day.

Soeren Sonnenburg, Cheng Soon Ong, Mikio Braun

JMLR MLOSS Special Track

http://jmlr.csail.mit.edu/mloss/



Machine Learning Open Source Software

To support the open source software movement, JMLR MLOSS publishes contributions related to implementations of non-trivial machine learning algorithms, toolboxes or even languages for scientific computing. Submission instructions are available here.

A Library for Locally Weighted Projection Regression
Stefan Klanke, Sethu Vijayakumar, Stefan Schaal; 9(Apr):623--626, 2008.
[abs][pdf] [code][mloss.org]

Shark
Christian Igel, Verena Heidrich-Meisner, Tobias Glasmachers; 9(Jun):993--996, 2008.
[abs][pdf] [code][mloss.org]

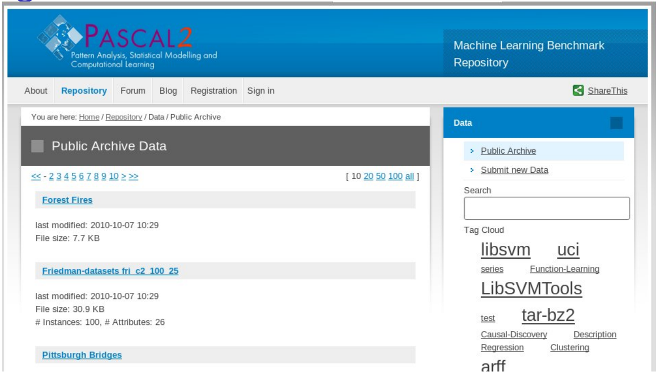
LIBLINEAR: A Library for Large Linear Classification
Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin; 9(Aug):1871--1874, 2008.
[abs][pdf] [code][mloss.org]

JNCC2: The Java Implementation Of Naive Credal Classifier 2
Giorgio Corani, Marco Zaffalon; 9(Dec):2695--2698, 2008.
[abs][pdf] [code][mloss.org]

- Submit software together with 4 page description.
- About 20 projects published so far.

Soeren Sonnenburg, Cheng Soon Ong, Mikio Braun

mldata.org



PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Machine Learning Benchmark Repository

You are here: Home / Repository / Data / Public Archive

Public Archive Data

<< - 2 3 4 5 6 7 8 9 10 > >>   [10 20 50 100 all]

Forest Fires

last modified: 2010-10-07 10:29
File size: 7.7 KB

Friedman-datasets fri_c2_100_25

last modified: 2010-10-07 10:29
File size: 30.9 KB
Instances: 100, # Attributes: 26

Pittsburgh Bridges

- Repository of data sets.
- Fully versioned, editable like a Wiki.
- Defines own data format based on HDF5. If used, additional features are available (e.g. automatic evaluation of prediction errors, download of data in other formats)

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich