Overview

- Open repository for machine learning data sets, learning tasks and solutions on those tasks.
- Started working on mldata/mlbench in October 2009.
- Current web site at mldata.org/mlbench.org.
Three types of objects:
1. **Data** raw data
2. **Task** learning task defined on raw data
3. **Solution** Method to solve a task

All data is editable with history ("Wiki")

Community features: Comments, rating, discussions, etc.
Current Status

- Implemented Data and Tasks, Solutions still work in progress.
- Community feature: Discussion forums, commenting, rating, wiki-like editing.
- Internally store data in HDF5 and convert to and from CSV, ARFF, svmlight, etc.
- Added existing repositories like UCI, libsvm as a starting point.
Motivation

Machine Learning (ML) is a scientific field comprised of both theoretical and empirical results. For methodological advances, one key aspect of reproducible research is the ability to compare a proposed approach with the current state of the art. Such a comparison can be theoretical in nature, but often a detailed theoretical analysis is not possible or may not tell the whole story. In such cases, an empirical comparison is necessary. To produce reproducible machine learning research, there are three main required components that need to be easily available:

- The paper describing the method clearly and comprehensively.
- The data on which the results are computed.
- Software (possibly source code) that implements the method and produces the figures and tables of results in the paper.

Acknowledgements

This project is supported by PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning)
This site is under construction! The database will be reset at irregular intervals. Thanks for your comments and criticisms.
Next steps

- Just try it at http://mldata.org
- Add a data set
- Add a task
- Comment, interact, etc
- **And most importantly:** Help us squash the bugs!