

Broader and Earlier Access to Machine Learning

Thomas Way, Lillian Cassel, Paula Matuszek,
Mary-Angela Papalaskari, Divya Bonagiri and Aravinda Gaddam
Villanova University
800 Lancaster Avenue
Villanova, PA 19803, USA
+1 610-519-7307
thomas.way@villanova.edu

ABSTRACT

This poster presents the results of an ongoing project that has developed learning modules to provide access to the tools and techniques associated with machine learning to a broad category of undergraduate students. The changing nature of science and its reliance on massive data sets has led to the integral use of machine learning approaches in just about every discipline. Recognizing this shift, flexible teaching materials have been developed to provide educators and students in a wide variety of academic fields with relevant machine learning topics and discipline-ready activities. Goals, module design, example material, dissemination plans and evidence of student learning are presented.

CCS Concepts

• **Computing methodologies** → **Machine learning** • **Social and professional topics** → **Professional topics** → **Education** → **Digital libraries and archives**.

Keywords

Machine learning modules; Big data; Computer science education.

1. INTRODUCTION & GOALS

In this age of Big Data, the nature of science is changing [3] and machine learning is becoming an integral part of how science (and just about everything else) is done. The typical undergraduate is unlikely to see any introduction to machine learning though all can benefit [1]. This project addresses the need for broadly accessible machine learning content across the spectrum of disciplines.

The goals of the project are threefold: (1) identify machine learning topics and relevant presentation techniques for undergraduates and faculty across disciplines, (2) produce stand-alone modules that can be adapted for learning in a variety of fields, and (3) disseminate the knowledge to a wide audience.

2. RESULTS

To support broader and easier access to machine learning, we devised a structure for educational modules consisting of: a one page topic summary and lesson plan for an instructor, slides and accompanying notes, student learning exercises targeted to the topic, appropriate software tools for the topic, other resources as needed such as references, videos, visualizations and URLs.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).
ITiCSE'16, July 09-13, 2016, Arequipa, Peru
ACM 978-1-4503-4231-5/16/07.
<http://dx.doi.org/10.1145/2899415.2925485>

For example, the starter module “Introduction to Classification and Clustering” can be done in half of a class meeting and consists of a summary, lesson plan, hands-on activity, slides, and a quiz. The more involved “Support Vector Machines” module assumes prior module experience and also includes data sets applicable to multiple disciplines and a tutorial for the popular and free WEKA software (www.cs.waikato.ac.nz/~ml/weka) to explore the topic using the provided data.

Modules are available, or in progress, for ten topics. Some are introductory and require no previous experience while others expect one or more of the other modules as background. Modules include: **Introduction to Machine Learning** (fundamental definitions, experimenting with the “Animal Game”, prerequisite module for all others), **Classification and Clustering** (definitions, unplugged and paper-based activities, minimal slides), **Using WEKA** (illustrated step-by-step instructions on installation and basic use), **Evaluating Classifiers** (WEKA activities), **Decision Trees** (discussion, WEKA), **Naive Bayes** (paper-based and Python program-based activities), **Neural Networks** (WEKA or Simbrain activities), **Support Vector Machines (SVMs)** (slides, WEKA, data sets), **Dimensionality Reduction** (slides, hands-on activities), and **K Means Clustering** (WEKA or Python-based activities).

Module content is being disseminated via ComputingPortal.org, a digital archive created as part of the Ensemble Project [2]. The content is organized hierarchically with elements available for individual download or as a zipped package. Discipline-specific data sets are being gathered to cover a wide variety of relevancy.

Results of pre- and post-test evaluations of the Introduction to Classification and Clustering module were given two successive semesters and show that student understanding improved from 20% on the pre-test to 90% on the post-test. Follow-up discussions during class made it clear that students understood the concepts as well as the distinctions between the techniques.

3. ACKNOWLEDGMENTS

This project is funded in part by NSF DUE award 1141033. Thanks to Carol Weiss for assistance with learning measure design.

4. REFERENCES

- [1] Domingos, P. “The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.” Basic Books, New York, NY, 2015.
- [2] Hislop, Gregory W., et al. “Ensemble: creating a national digital library for computing education.” Proceedings of the 10th ACM conference on SIG-information technology education. ACM, 2009.
- [3] Tansley, S, and Tolle, K. M., eds. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Redmond, WA: Microsoft Research, 2009.