|  |  |
| --- | --- |
| **Course Title** | Machine Learning for Physics and Astronomy |
| **Proposal Date** | 1/10/2016 |
| **Proposer’s Name**  | V. Acquaviva, A. Satyanarayana |
| **Course Number** | PHYS 3600  |
| **Course Credits, Hours** | 4 class hours (1 lecture, 3 computer lab), 3 credits |
| **Course Pre / Co-Requisites** | CST 1201 or equivalent, MAT1272 or MAT1372 or MAT 2572 or permission |
| **Catalog Course Description** | The course focuses on problem solving in Physics and Astronomy through statistical inference, machine learning algorithms and data mining techniques. Students will be presented with data sets and research problems in different areas of physics and will solve them using tools such as Bayesian statistics, Monte Carlo sampling, regression and classification algorithms, dimensionality reduction, and data cleaning. The programming assignments will be carried out in current, flexible languages, such as Python.  |
| **Brief Rationale**Provide a concise summary of why this course is important to the department, school or college. | This course was developed specifically for the Applied Computational Physics major and aims to provide students with a modern, flexible toolset that can be used to answer open-ended research questions based on data. This is a fundamental skill required for both the academic or industry route. Students will become familiar with data sets and practical applications from different branches of physics, ranging from geophysics to particle physics, condensed matter physics and astrophysics. For their final project, student will select their own data set, formulate a research question, and select their own tools to answer it. Through this process, they will begin to think as researchers and learn what to do when “the answer is not in the back of the book”.  |
| **Intent to Submit as Common Core**If this course is intended to fulfill one of the requirements in the common core, then indicate which area. | This course will not be submitted to fulfill a common course requirement. |
| **Intent to Submit as An Interdisciplinary Course** | This course will be submitted as an interdisciplinary course together with the CST department. |
| **Intent to Submit as a Writing Intensive Course** | This course will not be submitted as a writing intensive course. |

**Learning Outcomes and Assessments**

*Discipline specific*

|  |  |
| --- | --- |
| **Learning outcome** | **Assessment** |
| Learn how to formulate and test hypotheses on data | In-class quizzes, homework assignments |
| Learn how to recognize classification/regression problems and diagnose bias and variance issues | In-class quizzes, homework assignments |
| Learn how to use supervised learning algorithms and compare their performance | In-class quizzes, homework assignments |
| Learn how to extract information from unstructured data, recognize patterns and mine information | In-class quizzes, homework assignments |
| Learn how to use the numpy, scipy, matplotlib, scikit in Python in order to analyze, visualize and mine data  | In-class quizzes, homework assignments, final project |

*General Education*

|  |  |
| --- | --- |
| **Learning outcomes** | **Assessment** |
| Understand and employ both quantitative and qualitative analysis to describe and solve problems, both independently and cooperatively. | In-class quizzes, homework assignments, final project |
| Employ scientific reasoning and logical thinking. | In-class quizzes, homework assignments  |
| Communicate in diverse settings and groups, using written (both reading and writing), oral (both speaking and listening), and visual means. | Final project presentation |

**Example Weekly Course Outline with tools, sources, and department assignment**

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Topic or Challenge | Data Set | Tools introduced |
| 1 | Introduction; course overview(Physics/CST) |  | Basic Python commands and tutorials |
| 2 | Basic statistics and terminology recap(Physics) |  | Statistical Analysis 1 - Basic statistical concepts. Mean, standard deviation. Rank statistics and percentiles. Package: numpy |
| 3 | How much does the Universe weigh?(Physics) | WMAP/Planck<http://pla.esac.esa.int/pla/>+ tutorial | Statistical Analysis 2 - Covariance, correlation, analysis of variance; Hypothesis testing. Package: Scipy |
| 4 | How many variables do you need to fit the Universe? And what are their values?(Physics) | WMAP/Planck | Inference; Bayes theorem; maximum likelihood/Monte Carlo methods; goodness of fit and chi2 test |
| 5 | Telling a story: how to present data in an organic manner(Physics) | WMAP/Planck | Data visualization; Types of charts, graphs and tables. Composite charts. Visualization of multidimensional data. Package: Matplotlib. |
| 6 | Introduction to machine learning terminology and concepts(CST) |  | Intro to machine learning; linear regression, gradient descent, cost function |
| 7 | Evaluating performance; Diagnostics;Troubleshooting(CST) | A simplified version of the Higgs boson data set used in the 2014 Kaggle challenge:https://higgsml.lal.in2p3.fr/ | Logistic regression, regularization, underfitting/overfitting (bias and variance).  |
| 8 | Application Of Support Vector Machines To Global Prediction Of Nuclear Propertieshttp://arxiv.org/abs/nucl-th/0603037(CST) | Brookhaven National Nuclear Data Center (NNDC) http://www.nndc.bnl.gov/ | Classification problems and algorithms I (SVMs, using regression for classification) Package: Scikit-learn |
| 9 | Using Random Forests to Classify W+W- and ttbar Events <http://arxiv.org/abs/1410.8058>(Physics) | LHC data | Classification problems and algorithms II (Decision Trees; Random Forests) |
| 10 | Decadal climate predictions using sequential learning algorithms<http://arxiv.org/abs/1509.05285>(Physics) | NCEP reanalysis datahttp://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html | Online learning  |
| 11 | Probing modifications of General Relativity using current cosmological observations<http://arxiv.org/abs/1003.0001>(Physics) | Six combined public data sets described in the source paper. | Unsupervised learning algorithms; Clustering; Dimensionality reduction (PCA) |
| 12 | Exploring the relationship between the Engineering and Physical Sciences and the Health and Life Sciences by advanced bibliometric methods<http://arxiv.org/abs/1407.0199>(Physics) | Thomson Reuters’ Web of Science (WoS) database | Text processing |
| 13 | Feature importance for machine learning redshifts applied to SDSS galaxies<http://arxiv.org/abs/1410.4696>(Physics) | Sloan Digital Sky Survey data | Data cleaning/ real world data: how to deal with formatting, outliers, variable selection |
| 14 | Big Data challenges(CST) | Google Speller | Scaling data and solutions |
| 15 | Final project(Physics/CST) |  | Final project presentations |

**Grading Policy and Procedure**

The course will be graded according to: in-class quizzes and homework (20%), homework (3 sets, 15% each), and final project (35%).

**Required and Recommended Instructional Materials**

<http://scikit-learn.org/stable/user_guide.html> (open source) is required and can be downloaded as PDF or viewed online.

Students will receive a list of data sets to be downloaded during the first week of classes.

Additional reading material (papers and tutorials) will be distributed on a week-by-week basis.

Recommended:

Machine Learning in Python: Essential Techniques for Predictive Analysis, Michael Bowles, Wiley, 2015, ISBN 9781118961759.