**BerkeleyX:** CS190.1x Scalable Machine Learning

## COURSE OVERVIEW

Machine learning aims to extract knowledge from data and enables a wide range of applications. With datasets rapidly growing in size and complexity, learning techniques are fast becoming a core component of large-scale data processing pipelines. This course introduces the underlying statistical and algorithmic principles required to develop scalable real-world machine learning pipelines. We present an integrated view of data processing by highlighting the various components of these pipelines, including feature extraction, supervised learning, model evaluation, and exploratory data analysis. Students will gain hands-on experience applying these principles by using Apache Spark to implement several scalable learning pipelines.

## PREREQUISITES

Programming background; comfort with mathematical and algorithmic reasoning; familiarity with basic machine learning concepts; exposure to algorithms, probability, linear algebra and calculus; experience with Python (or the ability to learn it quickly). All exercises will use PySpark, but previous experience with Spark or distributed computing is NOT required. You should take this Python quiz before the course and take this Python mini-course if you need to learn Python or refresh your Python knowledge. This self-assessment document provides online resources that review additional relevant background material.

## PIAZZA DISCUSSION GROUP

We are using Piazza for all course questions and discussions. Please sign up at the discussion forum here: https://piazza.com/edx_berkeley/summer2015/cs1901x (access key: **cs1901x**)

## COURSE CONTENT

**WEEK 0:** Setup Course Software Environment - *Launches June 22 at 16:00 UTC*

- **Topics:** Step-by-step instructions for installing / using the course software environment, and submitting assignments to the course autograder.
- **Setup:** Download and install the course software environment, run your first Apache Spark notebook, and submit it for grading. (Due July 4th, 2015 at 07:00 UTC)
- **Note to CS100.1x students**: The course software environment is identical to that of BerkeleyX CS100.1X. If you've already setup the CS100.1X environment you do NOT need to go through this process again, but you MUST submit your notebook to receive credit for this course.

**WEEK 1:** Course Overview and Introduction to Machine Learning - *Launches June 29 at 16:00 UTC*

- **Topics:** Course goals, Apache Spark overview, basic machine learning concepts, steps of typical supervised learning pipelines, linear algebra review, computational complexity / big O notation review.
- **Lab 1:** NumPy, Linear Algebra, and Lambda Function Review. Gain hands on experience using Python's scientific computing library to manipulate matrices and vectors, and learn about lambda functions which will be used throughout the course. (Due July 11, 2015 at 07:00 UTC)

**WEEK 2:** Introduction to Apache Spark - *Launches June 29 at 16:00 UTC*

- **Topics:** Big data and hardware trends, history of Apache Spark, Spark's Resilient Distributed Datasets (RDDs), transformations, and actions.
- **Lab 2:** Learning Apache Spark. Perform your first course lab where you will learn about the Spark data model, transformations, and actions, and write a word counting program to count the words in all of Shakespeare's plays. (Due July 11, 2015 at 07:00 UTC)
- **Note to CS100.1x students**: This material is identical to Week 2 of BerkeleyX CS100.1x, and if you've already completed Lab 1 of CS100.1x you can simply submit your completed notebook to receive credit.

**WEEK 3:** Linear Regression and Distributed Machine Learning Principles - *Launches July 6 at 16:00 UTC*

- **Topics:** Linear regression formulation and closed-form solution, distributed machine learning principles (related to computation, storage, and communication), gradient descent, quadratic features, grid search.
- **Lab 3:** Millionsong Regression Pipeline. Develop an end-to-end linear regression pipeline to predict the release year of a song given a set of audio features. You will implement a gradient descent solver for linear regression, use Spark's machine Learning library ( mllib) to train additional models, tune models via grid search, improve accuracy using quadratic features, and visualize various intermediate results to build intuition. (Due July 18, 2015 at 07:00 UTC)

**WEEK 4:** Logistic Regression and Click-through Rate Prediction - *Launches July 13 at 16:00 UTC*

- **Topics:** Online advertising, linear classification, logistic regression, working with probabilistic predictions, categorical data and one-hot-encoding, feature hashing for dimensionality reduction.
- **Lab 4:** Click-through Rate Prediction Pipeline. Construct a logistic regression pipeline to predict click-through rate using data from a recent Kaggle competition. You will

extract numerical features from the raw categorical data using one-hot-encoding, reduce the dimensionality of these features via hashing, train logistic regression models using mllib, tune hyperparameter via grid search, and interpret probabilistic predictions via a ROC plot. (Due July 25, 2015 at 07:00 UTC)

**WEEK 5:** Principal Component Analysis and Neuroimaging - *Launches July 20 at 16:00 UTC*

- **Topics:** Introduction to neuroscience and neuroimaging data, exploratory data analysis, principal component analysis (PCA) formulations and solution, distributed PCA.
- **Lab 5:** Neuroimaging Analysis via PCA - Identify patterns of brain activity in larval zebrafish. You will work with time-varying images (generated using a technique called light-sheet microscopy) that capture a zebrafish's neural activity as it is presented with a moving visual pattern. After implementing distributed PCA from scratch and gaining intuition by working with synthetic data, you will use PCA to identify distinct patterns across the zebrafish brain that are induced by different types of stimuli. (Due August 1, 2015 at 07:00 UTC)

## COURSE GRADING POLICY

The components of the course grade are:

- 5% Course software setup
- 10% Quizzes (associated with lectures and labs)
- 85% Five Spark coding labs

The course is graded on the following scale:

- 85-100: A
- 70-84: B
- 60-69: C
- < 60: non-passing

We encourage you to start software setup and the Spark coding labs as early as possible. *There is an automatic 3-day grace period for submission deadlines. After the grace period, there is a 20% penalty for late submissions.*

## CREDITS

This course is sponsored by Databricks.

About Blog News FAQs Contact Jobs Donate Sitemap

Terms of Service & Honor Code Privacy Policy Accessibility Policy

Powered by Open edX