Features Business Explore Marketplace Pricing

This repository Search

Sign in or Sign up

jmportilla / LearnDataScience

forked from nborwankar/LearnDataScience

Watch 2 Star 1 Fork 959

Code Pull requests 0 Projects 0 Insights

Open Content for self-directed learning in data science

58 commits 2 branches 0 releases 5 contributors

Branch: master New pull request Find file Clone or download

This branch is 2 commits behind nborwankar:master. Pull request Compare

nborwankar next pylab update Latest commit `97d8d7d` on Jul 3, 2015

datasets	added ucihar	4 years ago
notebooks	next pylab update	2 years ago
styles	initial commit with notebooks etc but no datastest	4 years ago
.gitignore	Initial commit	4 years ago
LICENSE.txt	Update LICENSE.txt	4 years ago
README.md	Update README.md	3 years ago

README.md

Who

- Nitin Borwankar - primary developer
  Sponsored by Pivotal Inc. and Alpine Data Labs
  Community forming at Google Group "learnds"

What

- A collection of Data Science Learning materials in the form of IPython Notebooks.
- Associated data sets.

The initial beta release consists of four major topics

- Linear Regression
- Logistic Regression
- Random Forests
- K-Means Clustering

Each of the above has at least three IPython Notebooks covering

- Overview (an exposition of the technique for the math-wary)
- Data Exploration (the nuts and bolts of real world data wrangling)
- Analysis (using the technique to get results)

One or more of these may have supplementary material. Each of these have worksheets that contain mostly the code sections so you can iteratively explore the code.

Three openly available data sets are used.

- For the Linear and Logistic Regression we use a data set on loans and interest rates provided by Learning Club http://learningclub.com
- For Random Forests we use a data set of Android accelerometer and gyroscope readings used to predict body position and motion from the Human Activity Recognition project http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
- UN data on economic indicators of countries

Why

There's a need for open content to raise the level of awareness and training in basics, in the Data Science field (circa early 2013).

IPython Notebook provides an appropriate platform for rapid iterative exploration and learning.

When

Starting in 2013 and intended to extend for a long while.

Where

Today github, tomorrow the world. Google Group "learnds"

How

Learn Data Science is based on content developed by me (Nitin Borwankar) for the Open Data Science Training project http://opendst.org Most of the content (circa July 2013) is copyright (c) Alpine Data Labs as per the license at opendst.org, and is freely available. Extensions to the content embodied in this projects content are also released under the same license - see the LICENSE.txt file.

IPython Notebooks at Beta.

- A0. Before You Begin
- A1. Linear Regression - Overview
- A2. Linear Regression - Data Exploration - Lending Club
- A3. Linear Regression - Analysis
- B1. Logistic Regression - Overview
- B1a. Odds, LogOdds and Logit Function
- B2. Logistic Regression - Data Exploration
- B3. Logistic Regression - Analysis
- C1. Random Forests - Overview
- C2. Random Forests - Data Exploration
- C3. Random Forests - Analysis
- D1. K-Means Clustering - Overview
- D2. K-Means Clustering - Data Exploration
- D3. K-Means Clustering Analysis
- WA1. Linear Regression Overview Worksheet
- WA2. Linear Regression - Data Exploration - Lending Club Worksheet
- WA3. Linear Regression - Analysis Worksheet
- WA4. Linear Regression - Data Cleanup Worksheet
- WB3. Logistic Regression - Analysis- Worksheet

- WC3. Random Forests - Analysis - Worksheet
- WC4. Random Forests - Data Cleanup Worksheet
- WD2. K-Means Clustering - Data Exploration-Worksheet
- WD3. K-Means Clustering Analysis - Worksheet
- Z0. A quick tour of the IPython notebook
- Z1. Appendix 1 Plotting code snippets

Background

If you are unfamiliar with IPython Notebook you can start with http://ipython.org/notebook

Installation

- Prerequisites
  One of the following distributions is needed. Please note that even if you have Python installed it is important to have one of these distributions installed and the binary for this installation in your path. This is because these distributions come packaged with all the supplementary libraries needed and these have been historically difficult to install separately.
  - EPD Free Enthought Python Distribution from http://enthought.com
  - Anaconda Python from http://continuum.io
  - Development has been done on v 1.5 of Anaconda distribution but EPD Free should work just as well.
- The following steps assume you have installed one of the distributions mentioned in prerequisites.
- From a zip or tar file
  - download the zip or tar file
  - unpack the file to a directory called learnds
  - cd to the 'notebooks' subdirectory
  - start IPython Notebook 'ipython notebook --pylab=inline'
- From the git repo
  - clone the repo
  - cd to 'notebooks'
  - start IPython Notebook 'ipython notebook --pylab=inline'