CS249 – Principles of Data Mining/Data Science – Spring 2014

DRAFT SYLLABUS 4/03/14

D.S. Parker
Tuesday/Thursday 4–6
Boelter Hall 9436

1 Course Goals

The course will study two classes of topics using leading open-source implementations:

1. **Statistical Linear Modeling** (with R)
   emphasizes linear algebra and generalized linear models as a foundation for data analysis.
2. **Data Science** (with Python)
   extends this with higher-level methods for data science (notably IPython and other PyData tools)

The course emphasizes hands-on use of these environments; all lectures and assignments will rely on IPython notebooks.

- Texts: there are several texts for these topics (see below); PDF for the primary texts is freely available online.
- Software: students must use their own R and Python distributions, downloaded as described below.
- Prerequisites: Students who lack experience in either linear algebra or in programming should not take this course.

2 Course Outline

Linear Modeling — with R — texts: [ESL], [ISL]

Apr	1	Overview; Data Mining and Data Science; R and IPython
Apr	3	Statistical Learning [ESL ch.2; ISL ch.2]; Exploratory Data Analysis; Distributions, Laws
Apr	8	Matrix Algebra, Eigenstructure, SVD, Matrix Approximations
Apr	10	Dimensionality Reduction; Distance and Similarity, MDS; Covariance, PCA [ISL ch.10]
Apr	15	Linear Methods for Regression [ESL ch.3; ISL ch.3, ch.6.1-2]
Apr	17	Linear Methods for Classification [ESL ch.4; ISL ch.4]
Apr	22	Basis Expansions and Regularization [ESL ch.5; ISL ch.6]; Kernel Smoothing Methods [ESL ch.6]
Apr	24	Model Assessment and Selection [ESL ch.7; ISL ch.6]
Apr	29	MIDTERM — Linear Modeling (IN CLASS) — [ESL chs.2-7; ISL chs.2-6]

Data Science — with Python — texts: [PDA],[ESL],[ISL]

May	1	Data Science Methodology (ipython, scikit-learn) [PDA ch.3]
May	6	Feature Engineering, Data Munging, Data Wrangling (pandas) [PDA chs.5,6,7]
May	8	STUDENT PROJECT INITIAL OVERVIEWS (5-minute data/project concept presentations; attendance required)
May	13	Exploratory and confirmatory Visualization; Information graphics (matplotlib, pandas) [PDA ch.8]
May	15	OLAP, Business Analytics; Time Series (pandas) [PDA chs.9,10,11]
May	20	Arrays; Graph Analysis; Matrix and Tensor Decomposition (numpy; networkx; scikit-learn) [PDA chs.4,12]
May	22	Unsupervised Learning; Mixture models; Clustering; Outlier Detection [ESL ch.14; ISL ch.10] (scikit-learn)
May	27	Ensemble Methods [ESL chs.10,15,16; ISL ch.8] (scikit-learn)
May	29	STUDENT PROJECT PROGRESS OVERVIEWS (5-minute notebook presentations; attendance required)
June	3	Meta-Modeling and Meta-Programming [ESL chs.8-16; ISL chs.7-9] (Spark)
June	5	High-Dimensional Problems [ESL ch.18]

Conclusion — Finals Week

June	11 ?	STUDENT PROJECT FINAL OVERVIEWS (5-minute notebook presentations; attendance required)
June	13	FINAL COURSE PROJECT NOTEBOOK DUE (5:00pm Friday, end of Finals Week)

3 Course Grading

Homework	20%	
Midterm examination	30%	
In-class project reviews	10%	(helpful comments on in-class presentations of project ideas and progress)
Course project notebook	40%	(final notebook, logging experiences and lessons learned)

Midterm: an in-class, closed-book exam covering linear modeling.

Course Project: *a term paper implemented as an IPython notebook* that includes: (1) a description of a dataset you used, and how it was obtained; (2) a step-by-step log of the analyses you attempted; (3) a summary of experiences, insights, and lessons learned. Groups of up to 4 students are encouraged to work together on a challenging project.

Initial, intermediate, and semi-final versions of the notebooks, overviewing project ideas, are to be presented in class. (After these presentations other students in the course are required to submit (anonymized) helpful written comments at a class website.) Final versions of the notebooks are due at the end of Finals Week.

4 Software

4.1 Data Science (in Python)

Python (http://www.python.org) is a great language for interoperation, and many Python modules have become de facto standards for Data Science. We will focus on modules in the PyData initiative (http://pydata.org):

- get the Anaconda data science distribution, which has all data science tools needed and avoids configuration hassles. Download the (free) Anaconda distribution from: http://continuum.io/downloads
- get the book *Python for Data Analysis* ([PDA], described below), a popular introduction to these tools.

4.2 Linear Modeling (in R)

R is a large and well-established statistical environment based on an earlier system called S, developed at AT&T Bell Labs. If your system doesn't have a way to install R automatically, you can download an installation package from http://r-project.org/. There are many books about R: http://www.r-project.org/doc/bib/R-books.html. However, the documentation is excellent http://cran.r-project.org/manuals.html, and you can teach yourself R with the help system:

```
help.search( "kernel" )    #  search R documentation (grep-style)
help( parcoord )           #  help about a specific function
? parcoord                 #  equivalent

library( MASS )            # load the library/package you need
example( parcoord )        # run example of use of the function
```

UCLA ATS Statistical Consulting (http://www.ats.ucla.edu/stat) also offers classes (http://www.ats.ucla.edu/stat/seminars), online books (http://www.ats.ucla.edu/books), resources about R (http://www.ats.ucla.edu/stat/r), and free consulting.

R also has a powerful facility for installing and running packages. There are thousands of user-contributed packages at http://cran.r-project.org/web/packages/, and this has made R a leading platform for Data Mining research.

After you get R working on your system, please install the packages for the [ESL] and [ISL] texts:

```
install.packages("ElemStatLearn")   #  package for the text "Elements of Statistical Learning" by Hastie et al.
install.packages("ISLR")            #  package for the text "Introduction to Statistical Learning" by James et al.
```

5 Course Reading

5.1 Required Course Texts

[ESL] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, 2009, ISBN: 9780387848570. [google products] [google books]

> This is a classic text, covering most principles needed to get to the frontier of research in data mining. However it is written with the perspective of researchers in statistics, and some Computer Scientists find it hard to read. The [ISL] book below attempts to make it more accessible.
> home page: http://statweb.stanford.edu/~tibs/ElemStatLearn (includes data, R code, and errata)
> free book PDF download: http://statweb.stanford.edu/~tibs/ElemStatLearn/download.html (Large PDF)

[ISL] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer-Verlag, 2013, ISBN: 9781461471387. [google products] [google books]

> An introductory companion book to [ESL], with all examples implemented in R scripts.
> home page: http://www-bcf.usc.edu/~gareth/ISL/ (with links to data, R code, and errata)
> free book PDF download: http://www-bcf.usc.edu/~gareth/ISL/ (Large PDF)

[PDA] Wes McKinney, *Python for Data Analysis*, O'Reilly, 2012. [google products] [google books]

> A good introduction to how to use the data science modules IPython, pandas, NumPy, and matplotlib.
> O'Reilly page: http://www.oreilly.com (you can buy the PDF)
> github page: https://github.com/pydata/pydata-book (data and code from the book)
> McKinney's blog: http://blog.wesmckinney.com

5.2 Ways to learn R

[r-project.org] A great place to start is at the R project site: http://r-project.org

- There is an excellent *Introduction to R* among the many R manuals at: http://www.r-project.org/manuals.html
- There is even a translation into Chinese: http://www.biosino.org/R/R-doc/
- The r-project.org site lists other documentation, including books about R: http://www.r-project.org/other-docs.html
- The R project site also has a list of short R tutorial writeups (many in languages other than English) at: http://cran.r-project.org/other-docs.html

[videos] There are good tutorial videos about R online.

- A sequence of tutorial videos by developers.google.com is at: https://www.youtube.com/playlist?list=PLOU2XLYxmsIK9qQfztXeybpHvru-TrqAP
- Many other R tutorial videos: https://www.youtube.com/results?search_query=R+programming+tutorials

[cookbooks] Advanced how-to books, with code snippets and semi-real demos:

- O'Reilly has a large set of books about R: http://search.oreilly.com/?q=R
- UCLA used to have free access to Safari, O'Reilly's online access portal: http://learnit.ucla.edu/safari

5.3 Ways to learn Python

[python.org] A great place to start is at the Python site: http://python.org

- There is an excellent *Beginner's Guide to Python* https://wiki.python.org/moin/BeginnersGuide
- There is even a translation into Chinese: https://wiki.python.org/moin/BeginnersGuideChinese
- Th python.org site lists other documentation, including books about Python: https://wiki.python.org/moin/PythonBooks

[videos] There are good tutorial videos about Python online.

- The Google Python class: https://developers.google.com/edu/python/
- Associated videos: https://www.youtube.com/playlist?list=PL61E606149255B362
- Official Python video site: https://www.python.org/doc/av

[cookbooks] Advanced how-to books, with code snippets and semi-real demos:

- O'Reilly has a large set of books about Python: http://search.oreilly.com/?q=Python
- UCLA used to have free access to Safari, O'Reilly's online access portal: http://learnit.ucla.edu/safari

5.4 Ways to learn Data Science Tools

[IPython] The IPython site offers many resources:

- videos and presentation slides about IPython
- quick reference card for IPython.
- a gallery of interesting IPython notebooks.

[PyData] The PyData.org project integrates key Python modules (IPython, NumPy, SciPy, Pandas, SciKit-learn, MatPlotLib, etc.)

PyData is an open-source organization promoting Python-based tools for Data Science. Some of the tools are developed by other groups or companies, and they have tutorial materials (like videos) as well; however the videos are a painless way to get oriented.

There will be a PyData Conference on May 2-4 in Menlo Park.

[SciPy] SciPy is the user conference about quantitative analysis using Python.

The SciPy 2013 Tutorial videos are another great source of information about these tools, offering unvarnished user experiences and interesting applications.

[Strata] The Strata Conference is the showcase industry conference for the Big Data/Data Science field.

Short videos from presentations by Big Data Rock Stars at the 2014 Strata Conference.

[cookbooks] Advanced how-to books, with code snippets and semi-real demos:

- O'Reilly has a large set of books about Data Science: http://search.oreilly.com/?q=data+science
- These books include the text [PDA] above: Wes McKinney, *Python for Data Analysis*, O'Reilly, 2012.
- UCLA used to have free access to Safari, O'Reilly's online access portal: http://learnit.ucla.edu/safari