Announcements

- 1/16: Welcome to the class! Hope you will enjoy it :)

CLASS MEETS:

Time: Tue & Thu 4:30PM - 5:50PM
Place: HBH A301

PEOPLE:

Instructor: Leman Akoglu
- Office hours: Thu 2-3 PM; also, by appointment
- Office: HBH 2118C, office ph 412-268-30 four three
- Email: invert (cs.cmu.edu @ lakoglu)

Teaching Assistants:
Runshan Fu
- Office hours: TBD
- Location: TA room HBH 3034
- Email:invert (andrew.cmu.edu @ runshanf)

Abhinav Maurya
- Office hours: TBD
- Location: TA room HBH 3034
- Email:invert (andrew.cmu.edu @ amaurya)

Graders:
Darshan Mohan
- Email:invert (andrew.cmu.edu @ darshanm)
- Office hours: by appointment

Yashu Pant
- Email:invert (andrew.cmu.edu @ ypant)
- Office hours: by appointment

COURSE DESCRIPTION:

Machine Learning (ML) is centered around automated methods that improve their own performance through learning patterns in data, and then using the uncovered patterns to predict the future and make decisions. ML is heavily used in a wide variety of domains such as business, finance, healthcare, security, etc. for problems including display advertising, fraud detection, disease diagnosis and treatment, face/speech/handwriting/object recognition, automated navigation, to name a few. See this for an extended introduction.

"If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions." -- Albert Einstein

"A problem well put is half solved." -- John Dewey

This course aims to equip students with the practical knowledge and experience of recognizing and formulating machine learning problems in the wild, as well as of applying machine learning techniques effectively in practice. The emphasis will be on learning and practicing the machine learning process, involving the cycle of feature design, modeling, and scaling.

"All models are wrong, but some models are useful." -- George Box

As there exists "no free lunch", we will cover a wide range of different models and learning algorithms, which can be applied to a variety of problems and have varying speed-accuracy-scalability-interpretability tradeoffs. In particular, the topics include generalized linear models, decision trees, Bayesian networks, feature selection, ensemble methods, semi-supervised learning, density estimation, latent factor models, network-based classification, and sequence models. See the syllabus for more.

This course is designed to give a graduate-level student a thorough grounding in the methodologies, technologies,
and best practices used in machine learning. This course does not assume any prior exposure to machine learning theory or practice. Undergraduates need instructor's permission to enroll. PhD students can either enroll or by permission audit the course.

**Learning Objectives**

By the end of this class, students will

- learn the main concepts, methodologies, and tools for machine learning
- be able to recognize machine learning tasks in real-world problems
- develop the critical thinking for comparing and contrasting models for a given task
- learn to reliably perform model selection and evaluation
- gain the experience of applying the data science process to various problems end-to-end

**BULLETIN BOARD and other info**

- For course material, assignments, announcements, and grades please see the [Canvas](http://www.andrew.cmu.edu/user/lakoglu/courses/95828/index.htm).
- For questions and discussions please use [Piazza](http://www.andrew.cmu.edu/user/lakoglu/courses/95828/index.htm).

**RECOMMENDED TEXTBOOKS:**

There is no official textbook for the course. I will post all the lecture notes and several readings on course website.

Below you can find a list of recommended reading. We will follow different parts of these various books. I recommend the top 3 books in this list as regular reading for the course, and the rest for consulting various subjects and for further reading.

- **An Introduction to Statistical Learning: with Applications in R**, FREE!  
  Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

  Ian H. Witten, Eibe Frank, Mark A. Hall

- **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**, O'Reilly  
  Foster Provost and Tom Fawcett

- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, FREE!  
  Trevor Hastie, Robert Tibshirani, Jerome Friedman

  Christopher M. Bishop

  Kevin P. Murphy

- **Advanced Data Analysis from an Elementary Point of View**, Cambridge U. Press 2015.  
  Cosma R. Shalizi

Further reading: see the list [here](http://www.andrew.cmu.edu/user/lakoglu/courses/95828/index.htm) and [Amazon's best selling ML books](http://www.andrew.cmu.edu/user/lakoglu/courses/95828/index.htm).

**MISC - FUN:**

Fake (ML) protest

---

_Last updated by Leman Akoglu, Dec 2017_
### Tentative Syllabus

**Disclaimer:** This is an ambitious list of topics that I aim to cover in this course. I will adjust the pace depending on the progress of and the feedback from the students in class. As such, it is possible that only some subset of these topics will end up being covered. HW and exams will be adjusted accordingly.

<table>
<thead>
<tr>
<th>Date</th>
<th>Lectures and Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/16</td>
<td>Lecture 1: Intro to ML</td>
</tr>
<tr>
<td></td>
<td>• What is ML?</td>
</tr>
<tr>
<td></td>
<td>• ML applications</td>
</tr>
<tr>
<td></td>
<td>• Machine learning paradigms</td>
</tr>
<tr>
<td></td>
<td>○ Supervised learning (classification, regression, feature selection)</td>
</tr>
<tr>
<td></td>
<td>○ Unsupervised learning (density estimation, clustering, dimensionality reduction)</td>
</tr>
<tr>
<td></td>
<td>• Data mining concepts and tasks</td>
</tr>
<tr>
<td></td>
<td>○ Association rules, similarity search, cluster analysis, outlier analysis</td>
</tr>
<tr>
<td></td>
<td>• Basic data types</td>
</tr>
<tr>
<td></td>
<td>○ (Mixed) attribute data, text, time series, sequence, network data</td>
</tr>
<tr>
<td></td>
<td>• The problem solving process:</td>
</tr>
<tr>
<td></td>
<td>○ Business/project understanding, data understanding through EDA, data preparation, modeling, evaluation, deployment</td>
</tr>
<tr>
<td></td>
<td><strong>Readings:</strong></td>
</tr>
<tr>
<td></td>
<td>• Witten &amp; Frank</td>
</tr>
<tr>
<td></td>
<td>Chapter 1.1-1.3</td>
</tr>
<tr>
<td></td>
<td>• Provost &amp; Fawcett</td>
</tr>
<tr>
<td></td>
<td>Chapter 2</td>
</tr>
</tbody>
</table>

**PART I: PRELIMINARY ANALYSIS AND DATA PREPARATION**

#### Lecture 2: Exploratory Data Analysis

1/18

- Getting to know your data
- Data types
- Attribute types
- Data quality issues
- Data visualization
  - Histogram, Kernel Density Estimation
  - Charts, plots, infographics
- Correlation analysis

**Readings:**

- Aggarwal        Chapter 2
- Witten & Frank  Chapter 2
- Hastie          Chapter 6.6.1

#### Lecture 3: Data Preparation

1/25

- Feature creation
- Data cleaning
  - Missing, inaccurate, duplicate values

HW1 out
• Data transformation
  - Feature type conversion
  - Discretization
  - Normalization / Standardization
• Data reduction
  - Feature and record selection
  - Principal Component Analysis
  - Multidimensional scaling
  - Manifold learning (Isomap, LLE)

Readings:
- Witten & Frank  Chapter 7.1-7.4
- PCA tutorial
- Isomap website  Isomap paper
- LLE website  LLE paper

PART II: SUPERVISED LEARNING

Lecture 4: Learning Distributions

2/1
• Point estimation
  - Maximum Likelihood Estimation (MLE)
  - Bayesian learning
  - Maximum A Posterior (MAP) Estimation
  - MLE vs. MAP
• Gaussians
• What is ML revisited

Readings:
- Bishop: Sec 1.5, 2.2, 2.3 (up to 2.3.6)
- (Additional Resource) Andrew Moore's basic probability tutorial

Lecture 5: Linear Models

2/6
• Linear Regression
• Robust Regression
• Sparse Linear Models
  - Feature subset selection: revisited
2/8
  - Shrinkage methods: ridge regression and Lasso
  - Principal components regression, Partial least squares

Readings:
- ISLR (James, Witten, Hastie, Tibshirani)  Chapter 3.1, 3.2, 3.3, 3.4
- ISLR (James, Witten, Hastie, Tibshirani)  Chapter 6.1, 6.2.1, 6.2.2, 6.3.1, 6.3.2

Other readings:
- Hastie  Chapter 3.1-3.4, 4.4
- Shalizi  Chapter 2, 11
- Murphy  Chapters 1.4, 7.1-7.5, 13.3-13.5
- Provost & Fawcett  Chapter 4
- Witten & Frank  Chapter 7.5
Lecture 6: Naive Bayes

- Bayes Optimal Classifier
- Conditional Independence
- Naive Bayes
- Gaussian Naive Bayes

Readings:
- Bishop 1.3, 1.5, 3.2
- Mitchell's Chapter on Naive Bayes and Logistic Regression (Sections 1 and 2)

Other readings:
- Murphy Chapter 3.4

Lecture 7: Logistic Regression and Generalized Models

- Logistic Regression
- Generalized Linear Models (GLMs)
- Generalized Additive Models (GAMs)
  - Basis expansions
  - Generalizations, shape functions

Readings:
- ISLR (James, Witten, Hastie, Tibshirani) Chapter 4.1, 4.2, 4.3
- ISLR (James, Witten, Hastie, Tibshirani) Chapter 7.1, 7.2, 7.3, 7.4, 7.6, 7.7
- Intelligible Models for Classification and Regression by Yin Lou, Rich Caruana, and Johannes Gehrke.

Other readings:
- Hastie Chapter 9.1, 9.3, 9.6
- Shalizi Chapter 12

Lecture 8: Model Selection

- What is a good model?
- Overfitting
- Decomposition of error
- Bias-Variance tradeoff
- Cross Validation
- Regularization
- Information Criteria (AIC, BIC, MDL)

Readings:
- Hastie Chapter 7.1 - 7.10
- Provost & Fawcett Chapter 5

HW1 due
HW2 out

Project proposal due

Lecture 9: Model Evaluation

- Performance measures for Machine Learning
- Creating baseline methods for comparison
- Visualizing model performance

Readings:
- Witten & Frank Chapter 5
- Provost & Fawcett Chapter 7, 8, 11
- Shalizi Chapter 3, 10

Lecture 10: Tree-based Methods

- Classification trees
From trees to rules
Missing values and pruning
Regression trees

Readings:
- Hastie Chapter 9.2
- Witten & Frank Chapter 4.3-4.4, 6.1-6.2
- Provost & Fawcett Chapter 3
- Shalizi Chapter 13
- Murphy Chapter 16.2

Midterm Exam (in class)

Spring Break; No Classes

Lecture 11: Support Vector Machines
SVM intuition, formulation, and the dual
Slack variables, Hinge loss
The Kernel trick
  - Kernel SVM
  - Kernel Logistic Regression
  - Kernel PCA

Readings:
- Witten & Frank Chapter 6.3
- ISL-with R Chapter 9
- Murphy Chapter 14.2, 14.5

Lecture 12: Instance-based Learning
Kernel Density Estimation
k-Nearest Neighbor Classifier
Kernel Regression
Locally-Weighted Linear Regression

Readings:
- Hastie Chapter 6.1-6.3, 6.6.1-6.6.2
- Murphy Chapter 1.4.1-1.4.3, 14.7
- Shalizi Chapter 7.1, 7.5

Lecture 13: Ensemble Learning
Combining multiple models
Bagging
Random Forests
Boosting

Readings:
- Witten & Frank Chapter 8
- Hastie Chapter 10.1, 15, 16
- ISL-with R Chapter 8.2

PART III: UNSUPERVISED AND SEMI-SUPERVISED LEARNING

Lecture 14: Clustering
Distance functions
Hierarchical clustering
k-means clustering
Kernel k-means clustering
k-medians clustering
Mixture models
The EM algorithm
Spectral clustering

Readings:
- Witten & Frank Chapter 6.8
- ISLR (James, Witten, Hastie, Tibshirani) Chapter 10.3
- Provost & Fawcett Chapter 6, 12 (part)
- Spectral Clustering tutorial by Ulrike von Luxburg

Lecture 15: Semi-supervised Learning

- Assumptions (smoothness, cluster, manifold)
- Semi-supervised learning
  - Self-training
  - Generative methods
  - Graph-based methods
  - Co-training

Readings:
- Witten & Frank Chapter 6.9
- Introduction to Semi-Supervised Learning
- Graph-based Semi-Supervised Learning Algorithms
- Combining Labeled and Unlabeled Data with Co-Training, Avrim Blum, Tom Mitchell

PART IV: LEARNING WITH COMPLEX DATA

Lecture 16: Unstructured Data: ML for Text

- Representing text
- Topic modeling, Applications
- Latent Dirichlet Allocation (LDA)
- Inference: Gibbs sampling
- Collapsed Gibbs sampling for LDA

Readings:
- Witten & Frank Chapter 9.5, 9.6
- Provost & Fawcett Chapter 10

Lecture 17: Dependent Data: ML for Networks

- Transductive learning
- Learning in networks with and without attributes
- Probabilistic relational network classifier
- Iterative classification
- Loopy belief propagation
- Applications to auction, accounting, opinion fraud

Readings:
- Collective Classification in Network Data, Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, Tina Eliassi-Rad
- A Simple Relational Classifier, Sofus A. Macskassy, Foster Provost

Project Presentations I (today's presenters return final report on 5/3)

Project Presentations II (today's presenters return final report on 5/3)
Assignments

ASSIGNMENTS ARE DUE AT THE BEGINNING OF LECTURE ON THE DUE DATE

COURSEWORK:

Coursework consist of (grading in parentheses):
- Homework (40%)
- Midterm exam (15%)
- Final exam (25%)
- Project (20%)

NOTE: All assignments (except projects) are to be done individually. Please see the Collaboration policy.

IMPORTANT DATES:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Note</th>
<th>Out</th>
<th>Due</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework 1</td>
<td>EDA, Linear Models, Naive Bayes</td>
<td>Jan 30</td>
<td>Feb 20</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 2</td>
<td>Logistic Regression, Decision Trees, Model Selection and Evaluation</td>
<td>Feb 20</td>
<td>Mar 12</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 3</td>
<td>SVM, Kernels, Instance-based learning, Ensembles</td>
<td>Mar 12</td>
<td>Apr 10</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 4</td>
<td>Clustering, Semi-supervised learning, Topic modeling, Collective classification</td>
<td>Apr 10</td>
<td>May 3</td>
<td>10%</td>
</tr>
<tr>
<td>Midterm Exam</td>
<td>(in class)</td>
<td>Mar 8</td>
<td>--</td>
<td>15%</td>
</tr>
<tr>
<td>Final Exam</td>
<td></td>
<td>TBD</td>
<td>--</td>
<td>25%</td>
</tr>
<tr>
<td>Project proposal</td>
<td>List of [datasets] [more project ideas]</td>
<td>--</td>
<td>Feb 22</td>
<td>1%</td>
</tr>
<tr>
<td>Midway report</td>
<td></td>
<td>--</td>
<td>Apr 10</td>
<td>5%</td>
</tr>
<tr>
<td>Project presentation</td>
<td>(in class)</td>
<td>--</td>
<td>May 1 &amp; 3</td>
<td>5%</td>
</tr>
<tr>
<td>Project final writeup</td>
<td></td>
<td>--</td>
<td>May 1 &amp; 3</td>
<td>9%</td>
</tr>
</tbody>
</table>

HOMEWORK:

Homework should be turned in at the beginning of the class on the day it is due. If you are taking late day(s), please send your homework as an email to the TA and also submit a hard copy next time in class. Note the number of late days you used on the top front of the first page of your homework.

We ask that you submit all your code that was used to complete the assignment electronically only (no print outs) via Blackboard.

EXAMS:

There will be a midterm and a final exam. Note: Both the midterm and the final will be open book, notes, papers, etc., but you are not allowed to use a computer. The tentative dates are posted above, the finalized dates will be announced during the semester.

PROJECTS:
Your class project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. Below, you will find some project ideas (will be posted some time during the semester). Your class project must be about new things you have done this semester; you cannot use results you have developed in previous semesters.

Projects can be done by you as an individual, or in teams of two students. The course TA will consult with you on your ideas, but of course the final responsibility to define and execute an interesting piece of work is yours.

Your project will be worth 20% of your final class grade, broken into four main deliverables:

- **Project proposal** (1% of the course grade)
- **Project milestone report** (5% of the course grade) (** 4 pages maximum **, including references) describing the results of your first experiments by the milestone due date (see above). Note that, as with any conference, the page limits are strict. Papers over the limit will not be considered.
- **Final project writeup** (9% of the course grade) preferably in ACM format (** 8 pages maximum, 4 pages minimum **, including references; page limit is strict)
- **Final project presentation** (last week in-class) (5% of the course grade)

**Project Proposal:**

You must turn in a brief project proposal (** 1 page maximum **) on the due date (see above), in class. A list of suggested projects and data sets are posted below. Read the list carefully. You are encouraged to use one of the suggested data sets, because we know that they have been successfully used for machine learning in the past. If you prefer to use a different data set, we will consider your proposal, but you must have access to this data already, and present a clear proposal for what you would do with it.

Project proposal format: Proposals should be 1 page maximum. Include the following information:

- Your name and Andrew ID on top of the page
- Project title
- Data set
- Project idea. This should be approximately two paragraphs.
- Papers to read. Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.
- Teammate: Will you have a teammate? If so, whom? Maximum team size is two students.
- What will you complete by the project milestone due date? Experimental results of some kind are expected here.

**Project Writeups:**

Your write-ups should include the information detailed below, in approximately the order given. Your write-up need not have corresponding sections or bullet points, but course staff should be able to find the information without searching too hard. Be as precise/specific as you can.

Note: The mid-way report will be a relatively incomplete version of the final write up. It should include similar sections and address similar questions, but need not contain all the details. Think of the mid-way report as a preliminary version of the final draft. It is more of a status report, including preliminary results, issues that you are facing in developing your project, and how you plan to modify your approach to tackle some of those issues moving forward.

- **Introduction/Motivation/Problem Definition (15%)**
  - What is it that you are trying to solve/achieve? Who cares and why does it matter?
  - Identify, define, and motivate the problem that you are addressing.
  - How (precisely) will a machine learning solution address the problem?

- **Data Understanding and Preparation (15%)**
  - Identify and describe the data (and data sources) that will support machine learning to address the problem.
  - Include various aspects of the data such as its size (GB/MB/TB/etc), type(s), format, etc.
  - Specify how these data are integrated to produce the format required for machine learning.

- **Methodology (30%)**
  This is where you give a detailed description of your primary contributions. It is especially important that this part be clear and well written so that we can fully understand what you did.
  - How did you approach the problem? What challenges did you face? In what (unique) ways did you handle those challenges?
  - Specify the type of model(s) built and/or information/knowledge extracted.
Discuss choices for machine learning algorithm: what are other alternatives, and what are their pros and cons (in the context of the problem and as compared to your proposed solution)?

Discuss why and how this model should “solve” the problem (i.e., improve along some dimension of interest).

- **Evaluation and Results (30%)**
  We are interested in seeing a clear and conclusive set of experiments which successfully evaluate the problem you set out to solve. Make sure to interpret the results and talk about what we can conclude and learn from your approach.
  - How do you evaluate your machine learning solution to the specific question(s) you have addressed?
  - What do these evaluation methods tell you about your solution?
  It is not so important how well your method performs but rather, (a) how thorough and careful your evaluation is, and (b) how interesting and clever your results and findings are.

- **Style and writing (10%)**
  Overall writing, grammar, organization, figures and illustrations.

You are suggested to use the ACM format to write your project reports (8 pages maximum, 4 pages minimum, including references; this page limit is strict).

**Project Presentations:**

- Think of this as an oral version of your final project writeup.
- Present your work in a meaningful and interesting flow (e.g., motivation, problem definition, data description, challenges, proposed methods, results and their interpretation).
- Make sure to include enough details and background of your methodology (similar to a conference talk).
- See here and here for some how-to on giving a good/bad talk.
- Be prepared to ask (tough) questions to other project groups.
- We will spend (the last) 2 lectures on project talks. Depending on the number of project groups, each group will be given 5-8 minutes including questions.

**Datasets for Project:**

We provide a long list of potential data sources for your project right here. The project is open-ended and you are expected to come up with your own project description and problem definition. In addition to your technical approach, we will evaluate your creativity in formulating an interesting and important problem for the project.

---

_Last modified by Leman Akoglu, Dec 2017_
Course Policies

LECTURES

- All devices such as laptops, cell phones, noisy PDAs, etc. should be turned off for the duration of the lectures and the recitations, because they may distract other fellow students.
- Please come to all lectures on time and leave on time, again so that there are no distractions to the classmates.

PREREQUISITES

Students are expected to have the following background:

- Basic knowledge of probability and statistics
- Basic knowledge of linear algebra and algorithms
- Working knowledge of basic computing principles
- Basic programming skills at a level sufficient to write a reasonably non-trivial computer program in a language of preference

ASSIGNMENTS

- Assignments are due at the *beginning of lecture* on the due date.
- The due date of assignments are posted at the assignments page.
- Assignments will be posted on Canvas.
- Students should submit the programming part of assignments electronically via Canvas, and return 'knitted' print-outs with answers in class.

Important Note: As we reuse problem set questions, covered by papers and webpages, we expect the students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not google for answers. The purpose of problem sets in this class is to help you think about the material, not just give us the right answers. Therefore, please restrict attention to the books mentioned on the front page when solving problems on the problem set. If you do happen to use other material, it must be acknowledged clearly with a citation on the submitted solution.

Questions and requests

- You should use Piazza for all your questions about the assignments and the course material. Instructor and TA(s) will do their best to answer your questions timely.
- Regrade requests should be done in writing/email,
  - within 2 days after graded assignments are distributed
  - to the TA that graded the question, and specifying
    - the question under dispute (e.g., 'HW1-Q.2.b')
    - the extra points requested (e.g., '2 points out of 5')
    - and the justification (e.g., 'I forgot to divide by variance, but the rest of my answer was correct')
  - In the remote case there is no satisfactory resolution, please contact the instructor.

Homework pick-up information

- You may pick up graded homeworks etc., from the course admin
  - Mrs. Adrienne McCorkle, HBH 2250
  - 9:00-11:30am, 1:30-4:30pm every weekday
  - with photo-id (for your privacy protection)

Late policy

- No delay penalties, for medical/family/etc. emergencies (bring written documentation, like doctor's note).
- Each student is granted '4 slip days' total for the whole course duration, to accommodate for coinciding
deadlines/interviews/etc. That is, no questions asked, if the total delay is 4 days or less.

- You can use the extension on any assignment during the course. For instance, you can hand in one assignment 4 days late, or 4 different assignments 1 day late each.
- Late days are rounded up to the nearest integer. For example, a submission that is 4 hours late will count as 1 day late.
- After you have used up your slip days, any assignment handed in late will be marked off 25% per day of delay.

- To use slip days:
  - upload your homework on Canvas to mark the time of submission
  - make sure to return hard copy next time in class
  - note down on the front page of your hard copy submission: count of slip days you used, as well as the count of slip days left

**Collaboration**

You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge any help from other students or non-class sources will receive NO credit. Copied work will receive NO credit. Any and all violations will be reported to the Heinz College administration and may appear in the student’s transcript.

**Academic integrity**

All students are expected to comply with CMU's policy on academic integrity. Please read the policy and make sure you have a complete understanding of it.

**EMAIL**

Piazza should be used for general course and assignment related questions. For other types of questions (e.g., to report illness, request various permissions) please contact the instructor directly via email.

Please make sure to include '95828' in the subject line of your email.

**AUDITING**

Auditing is not allowed. Only those students who are officially enrolled to take the course for credit are allowed to sit in class.

---

_Last modified by Leman Akoglu, Dec 2017_
Resources

A Few Useful Things to Know about Machine Learning

CheatSheets for R

https://github.com/writezhe/R_for_MLPS_cheatsheets provides a list of "cheatsheets" introducing R for the course, including commands, brief examples, and references necessary for each lecture.

Linear Algebra Review (Prerequisite)

Text:

- Linear Algebra Review
- Linear Algebra and Matrices

Slides:

- Linear Algebra Review

Videos:

- Essence of Linear Algebra

Probability Review (Prerequisite)

Text:

- Probability Cheatsheet
- Probability Reminders
- Review of Basic Concepts in Probability by Padhraic Smyth
- Review of Probability (book chapter)
- Probability Theory Review (optional)

Slides:

- Probability for Data Miners (1 to 25) by Andrew Moore
- Some Probability and Statistic (1 to 28) by David Blei
- Review of Probability by William Cohen

Tools (some open-source)

- Tableau for information visualization (free under academic program)
- Trifacta for data wrangling
- Prefuse for information visualization
- Cytoscape for network analysis
- WEKA: open-source machine learning toolkit
- scikit-learn Python libraries for machine learning
- SAS-Enterprise Miner for data analytics
- List of useful software

ML/KD contests and datasets

- Kaggle
- KDD Cup
Datasets for Data Science
Large collection of network datasets
Awesome Public Datasets
UCI datasets
ProPublica Data Store
NYC Open Data and Taxi Data
Boston taxi data
Graph data at SNAP and KONECT
Outlier Detection Data Sets (ODDS)
UCR Time Series Classification Archive
Stance identification dataset for fake news detection
LastFM
Foursquare check-ins
AOL query logs
Product review datasets
  - Amazon product reviews and more
  - Yelp Data Challenge
  - Online reviews from SNAP
Amazon product data
Amazon question/answer data
Stack Exchange Data Dump
Physician referral data
Google public datasets
List of large datasets open to public
U.S. Government's open data
Million Song Dataset
Quandl - a dataset search engine
Free 'big data' sources'
AWS Public Datasets
Academic data: Microsoft Academic Data and DBLP