

CIS606 – Lecture 1

Woon Wei Lee, Jacob Crandall
Spring 2014, 10:00am-11:15am,
Mondays and Thursdays

For today:

- CIS606: welcome and logistics
- Collaborative filtering intro

Machine Learning, CIS606 - Logistics, and Welcome!

- Course syllabus: to be uploaded to Moodle
- Welcome to new instructor: Dr. Jacob Crandall
- Teaching Assistant: Baluyan Hayk
- Course grading breakdown (Spring 2013, but little change expected):
 - Midterm examination 15%
 - Final examination 25%
 - Class participation 5%
 - Projects 55% (12.5%, 12.5%, 30%)
- Informal components:
 - Laboratories/Assignments
 - Presentations
- Course aims:

“This course significantly extends the topics covered in the pre-requisite course, CIS501. The aim is to provide an in-depth treatment of a variety of important concepts, techniques, and algorithms in machine learning. Topics covered include linear regression, boosting, support vector machines, hidden Markov models, and Bayesian networks. The underlying theme in the course is statistical inference as it provides the foundation for most of the methods covered.”

Course Materials

- Text Books
 - **Course Texts**
 - Haykin, Simon. Neural Networks and Learning Machines. 3rd Edition. Prentice Hall, 2008. ISBN: 978-0131471399
 - Bishop, Christopher. Pattern recognition and machine learning. Springer, 2007. ISBN: 978-0387310732
 - **Additional Reading**
 - Hastie, T., R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition. New York, NY: Springer, 2009. ISBN: 978-0387848570.
 - Bishop, Christopher. Neural Networks for Pattern Recognition. New York, NY: Oxford University Press, 1995. ISBN: 9780198538646.
 - Duda, Richard, Peter Hart, and David Stork. Pattern Classification. 2nd ed. New York, NY: Wiley-Interscience, 2000. ISBN: 9780471056690.
 - MacKay, David. Information Theory, Inference, and Learning Algorithms. Cambridge, UK: Cambridge University Press, 2003. ISBN: 9780521642989

Cont'd

- “Learning outcomes”
 - Learn key concepts and algorithms in machine learning
 - The relationship between statistical inference and machine learning
 - Determine when, how and why particular algorithms work in given situations.
 - Use machine learning tools to perform key operations: regression, classification, clustering and visualization
 - Implement and test machine learning techniques
 - Communicate scientific and technical issues effectively.

Re-marking procedure

- Those wishing to dispute grades awarded for projects reports, should follow the procedure below:
 - First discuss with either of us to determine if (which is likely!) there is a logical reason behind the grade.
 - Grading of reports is, unfortunately, quite subjective
 - but in general we should be able to provide feedback particularly for cases where the grade is particularly bad (or good, if you really want to dispute a good grade ;-))
 - Apart from trivial cases, submit a request *in writing*, stating detailed reasons for your disputing the grade
 - No “ad-hoc” grade adjustments will be made
 - The *entire* report will be re-marked
 - if possible by a different instructor (applicable for projects 1 and 2)
 - For project 3, we will both be marking the reports so unfortunately an independent remark is not possible but we can still re-evaluate the case.
 - All such requests to be submitted within 2 days of receiving report grade

Collaborative filtering (CF)

- **Process of extracting information from very large datasets**
 - Combination of inputs from multiple agents, users, perspectives, etc..
 - Prediction of response of any one of these agents given the combined wisdom of the “crowd”
- **Drivers**
 - Availability of advanced computers, data mining algorithms, etc, but mainly...
 - The web! - pioneers include websites like Amazon, Google, CDNow, etc etc etc
- **Other interesting examples include:**
 - Pandora, Last.FM, Jango.com
 - Netflix, IMDB, allmovie.com
 - Stumbleupon
 - Facebook, Friendster
 - Etc., etc..
- **Broadly referred to as *Recommender Systems***
 - Accepts “preferences” from the user and returns with a set of recommendations
 - Generally a form of supervised learning but with very special conditions

masdarr

About 898,000 results (0.07 seconds)

Showing results for **masdar**. Search instead for [masdarr](#)

[Welcome to Masdar](#) ☆ 🔍

Masdar is a commercially driven enterprise that operates across the full spectrum of renewable energy and sustainable technology industry. ...

[www.masdar.ae/](#) - [Cached](#) - [Similar](#)

[Careers](#)

[Contact us](#)

[Masdar City](#)

[Masdar Institute of Science and ...](#)

[Board Members](#)

[Introduction](#)

[Zayed future energy prize](#)

[Masdar, Global Energy Company ...](#)

is near.

Chicken Little (2005) [More at IMDbPro](#)


 **G** 81 min - [Animation](#) | [Adventure](#) | [Comedy](#)
- 4 November 2005 (USA)

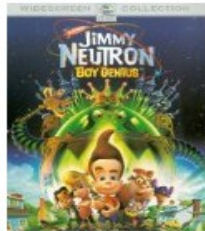
★★★★★☆☆☆☆☆ **5.8/10**


Users: (18,247 votes) 230 reviews | Critics: 145 reviews
Metascore: **48/100** (based on 32 reviews from Metacritic.com)

After ruining his reputation with the town, a courageous chicken must save the people of his fellow village when

Recommendations

 [Meet the Robinsons \(2007\)](#)

 [Jimmy Neutron: Boy Genius \(2001\)](#)

 [Space Chimps \(2008\)](#)

Cont'd

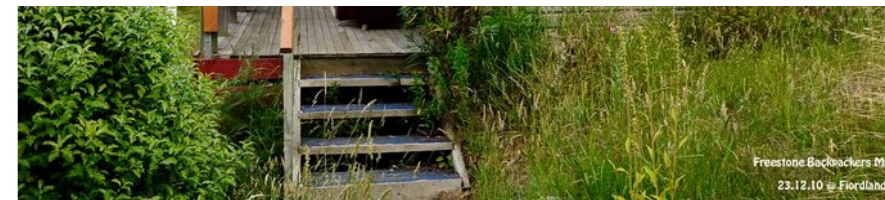
- **How are preferences collected or transmitted?**

- Explicit ratings:
 - i.e.: Users consciously enter their ratings
 - Ratings, “likes/dislikes” (thumbs up/down buttons)
 - Tags, Hyperlinks, Citations in Academic publications
 - “Who would you rather be stuck on a desert island with.. X or Y?”... etc.
 - Most accurate retrieval of user sentiment
- Implicit ratings
 - Based on the behaviour of the user
 - Web searches, phrases and statistics (multiple visits, etc)
 - Co-occurrence patterns
 - Blogging comments, pingbacks, trackbacks
 - Sentiment extraction using NLP
 - Citations in academic publications
 - Involves inference of user preference - may need post-processing, pattern recognition.

I think only one addition would make this book even better: have a list of references at the end of each chapter to find more information about certain topics. For programming can point readers to Lincoln Stein's book on CGI.pm for more details.

Was this review helpful to you? [\(Reset this\)](#)

2 of 2 people found the following review helpful:



A double room provided by the Freestone Backpackers which is really worth the price!

Added January 22 · Like · Comment

From the album
Fiordland

[Share](#)
[Tag This Photo](#)

Cont'd

- **Characteristics/Challenges -**
 - CF can be regarded as a supervised learning problem, but has a very unique set of problems and challenges:
 - Very large sets of data – order of millions or 100x millions of problems.
 - “MovieLens” dataset has 10M ratings, 10,000+ movies × 70,000+ users
 - “NetFlix prize” dataset has 100M+ ratings (!), with ~500,000 users and 17,000+ movies
 - Sparsity – *severe* missing data problems
 - Not all (or none of the) users had entered ratings for all items
 - There is “pattern” to the sparsity → new users will have very little data to go by, resulting in poor performance
 - New products/databases would take a very long time to reach acceptable performance.
 - Constant evolution – new users, new tags are constantly being added – need fast updating rules!

Techniques for CF

- **Standard formulation**

- In a typical scenario, there is:
 - A list of users, $U=\{u_1, u_2, \dots, u_m\}$
 - A list of items, $I=\{i_1, i_2, \dots, i_n\}$
 - Each user a has a list, I_a , of items for which ratings are available, and a corresponding rating, $r_{a,i}$ for each item in I_a

- Matrix representation (depicted right)
 - In data mining context, common to represent in the form of an $m \times n$ matrix

- **Goal is normally one of:**

1. To provide a *Prediction*, $P_{a,j}$ of the rating that that user would provide to item i_j
(Given that $i_j \notin I_a$)
2. To provide a *Recommendation* for user a .
This is typically a list of N items with highest probability of being “liked” by the user.

