

11-741/11-641/11-441: Machine Learning for Text Mining

Introduction

Yiming Yang
Carnegie Mellon University
yiming@cs.cmu.edu

Instructor and Teaching Assistants

Faculty

- Yiming Yang
- yiming@cs.cmu.edu
- GHC 6717



TA's (office hours on the syllabus)

- Yuexin Wu
- Rose Catherine Kanjirathinkal



Outline

- Course Contents Overview
- Administrative Details

Big Data Era

New data are being created at an unprecedented rate.

Every hour we add

- 1 1,000 articles on Wikipedia.
- 236,000 videos on Youtube.
- 3,400,000 pictures on Flickr.
- 46 million posts on Twitter.

How do we handle such huge and diversified data?

Text Mining Techniques

- **Clustering (of documents, users, shopping items, ...)**
 - To discover latent groups, topics, social communities, etc.
- **Classification**
 - To label data using predefined classification taxonomies
- **Authority Detection**
 - To measure the social impacts of web sites, papers, people, etc.
- **Collaborative Filtering (Recommendation)**
 - To predict personal interests based on the tastes of similar users

Clustering web search results

The image shows a Google search results page for the query "China foreign policy". Red annotations highlight key elements:

- A query:** Points to the search bar containing "China foreign policy".
- 7.5 M results -- impossible to read all!:** Points to the result count "About 7,540,000 results (0.57 seconds)".
- Clusters of documents:** Points to a list of search results, including:
 - China's in the hot seat on 2 major foreign-policy issues** (Business Insider - Aug 27, 2016)
 - China's Potemkin Diplomacy** (Forbes - Aug 27, 2016)
 - It's Time for China Analysts to Stop Talking Past One Another** (Foreign Policy (blog) - Aug 24, 2016)
 - China Is Fueling a Submarine Arms Race in the Asia-Pacific** (Foreign Policy - Aug 26, 2016)
- View all:** Points to the "View all" link below the first cluster of results.

Clustering web search results (cont'd)

Google China foreign policy

All News Videos Images Shopping More Search tools

About 7,540,000 results (0.57 seconds)

clustered articles by topic

China's in the hot seat on 2 major fo
Business Insider - Aug 27, 2016
As Chinese President Xi Jinping prepares for his first month, two major foreign-policy challenges loom.
China's Potential Diplomacy
Forbes - Aug 27, 2016
Diplomats never too late to curb NK nukes
The New York Times Herald - 7 hours ago
[View all](#)

North Korea opens its doors to the newly renovated central zoo
The Independent - 20 hours ago
But one of the most popular attractions might come as a surprise to foreign visitors. North Korea's new zoo-1.jpg ... Former South Korean President Kim Dae-jung, who pursued a sort of détente with Pyongyang called the "Sunshine Policy" presented the Jindo ...

Pyongyang's zoo big on dogs
Northwest Arkansas News - 9 hours ago
But one of the most popular attractions might surprise foreign visitors. Just across from the former South Korean President Kim Dae-jung, who pursued a sort of détente with Pyongyang the "Sunshine policy," presented the North with a Jindo ...

North Korea says it's now able to nuke US
Fox 32 Chicago - Aug 27, 2016
... are defensive in nature. Thursday marks the anniversary of the "Military First" policy in North Korea was on full display. ... North ...

'How could our country lie so completely?': meet the North Korean defector
The Guardian - Aug 27, 2016
Yet, as a young bureaucrat rising in the system, Park had no desire to leave; not until one summer of 1999, when he received a message from a Chinese man. He had ... "The police interrogate me, asking again and again if I listened to ...

©Yiming Yang Lecture Notes Fall 2016 7

Common Clustering Algorithms

- K-means
- Single-pass
- Bottom-up Hierarchical (Agglomerative)
- Top-down Hierarchical
- Bipartite Reinforcement
- Probabilistic (soft) Clustering
- Spectral Clustering

K-means Clustering of Text Data

- Input: matrix X (n documents by m unique words)

Sparse! $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & \cdots & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & x_{ij} & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & \cdots & \cdots & x_{nm} \end{bmatrix}$

Term (points to x_{12})

Document (points to x_{2m})

Term weight (TF*IDF) (points to x_{ij})

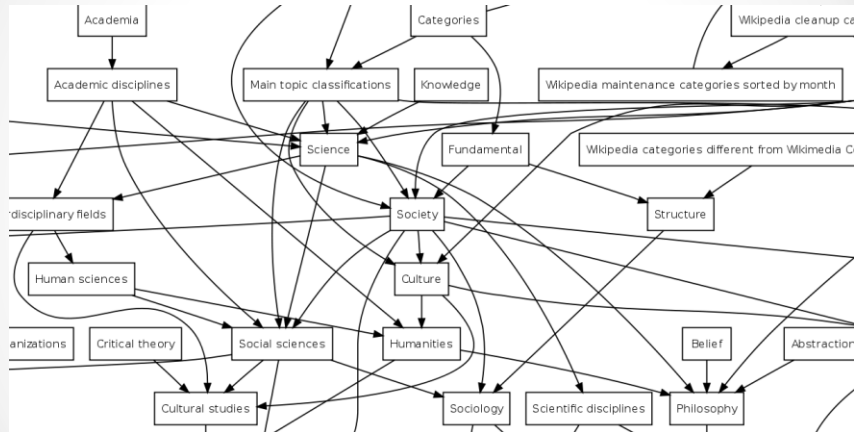
- Output: K clusters of documents (rows) or words (columns)

Text Classification



The screenshot shows the Wikipedia article for 'Text mining'. The article is titled 'Text mining' and is part of the Wikipedia encyclopedia. It includes a summary of the concept, its applications, and a list of references. The article is written in a clear, concise style, typical of Wikipedia. The left sidebar contains navigation links such as 'Main page', 'Contents', 'Featured content', 'Current events', 'Random article', 'Donate to Wikipedia', and 'Wikipedia store'. The bottom of the page features a 'Categories' section with links to 'Artificial intelligence applications', 'Data mining', 'Computational linguistics', 'Data analysis', 'Natural language processing', and 'Statistical nat'.

Wikipedia categories form a graph



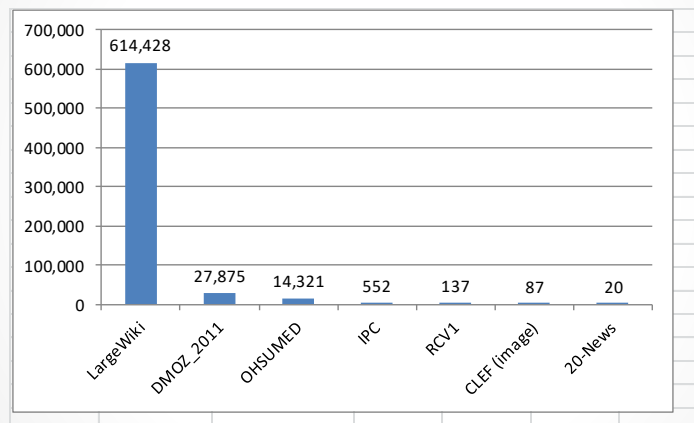
4.45 million articles in English Wikipedia, ~500,000 categories

• @Yiming Yang, Lecture on Web-scale Classification

7/29/2014

• 11

Large-scale Classification Challenge

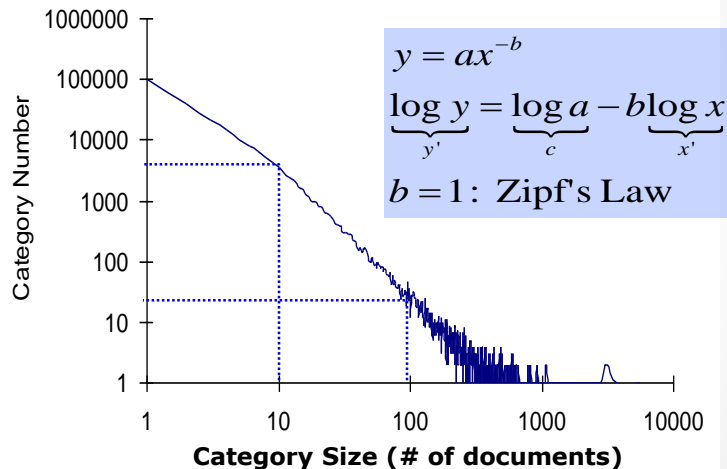


• ©Yiming Yang

Lecture Notes Fall 2016 12

The Power Law Phenomenon

(Skewed distribution of categories over documents)



©Yiming Yang

Lecture Notes Fall 2016

• 13

Many Classification Methods

- Rule-based Expert Systems (Hayes, 1990)
- Regression (linear, polynomial, logistic, ...) (Furh'91; Yang'92)
- Nearest Neighbor methods (Creedy'92; Yang'94)
- Naïve Bayesian probabilistic methods (Lewis'92)
- Decision trees, symbolic rule induction (Apte'94)
- Neural networks, logistic regression (Wiener'95)
- Error Correcting Output Coding (Kong & Dietterich'95)
- Rocchio-style (Lewis et al., 1996)
- Support Vector Machines (Joachims'98)
- Boosting or bagging (Schapire'98)
- Hierarchical Language Modeling (McCallum'98)
- First Order Inductive Learning (Slattery'99)
- ...

©Yiming Yang

Lecture Notes Fall 2016

• 14

Statistical Classification: A Toy Example

(Graph in ESL by Hastie et al.)

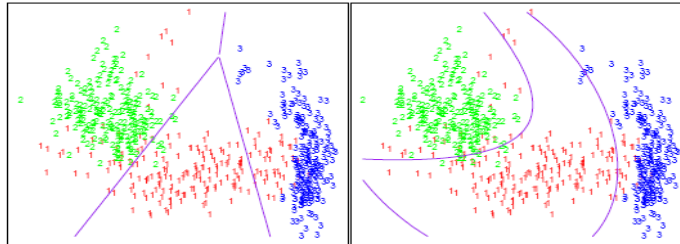


Figure 4.1: The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_{12}, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.

©Yiming Yang

Lecture Notes Fall 2016

● 15

Text Mining Techniques

- **Clustering** (of documents, users, shopping items, ...)
 - To discover latent groups, topics, social communities, etc.
- **Classification**
 - To label data using predefined classification taxonomies
- **Authority Detection**
 - To measure the social impacts of web sites, papers, people, etc.
- **Collaborative Filtering**
 - To predict personal interests based on the tastes of similar users

● ©Yiming Yang

Lecture Notes Fall 2016

● 16

★★★★★ = Must See
 ★★★★★ = Will Enjoy
 ★★★★★ = It's OK
 ★★★★★ = Fairly Bad
 ★★★★★ = Awful

movielens

helping you find the *right* movies

Welcome yiming@cs.cmu.edu ([Log Out](#))
 You've rated **56** movies.
You're the 28th visitor in the past hour.

[Home](#) | [Find Movies](#) | [Q&A \(new\)](#) | [Preferences](#) | [Help](#)

Shortcuts

Basic Search

Title:

All Genres ▾
All Dates ▾

Domain: All movies ▾

Tag:

☐ Use selected buddies!
☒ Exclude your ratings
☒ Exclude movies without predictions

Search!

Select Buddies

☐ Test Buddy

There are **22698** movies matching your search:
 Movies without a prediction are **Not Shown**
 Movies you've rated are **Not Shown**
 You've sorted by: **Prediction or Rating**
[Show Printer-Friendly Page](#) | [Download Results](#) | [Permalink](#)

Tags Related to Your Search: sci-fi (2895), based on a book (2831), comedy (2763), action (2701), atmospheric (2476), (about tags)

Page 1 of 1514 1 2 3 4 ... 1514 next Skip to page # Go

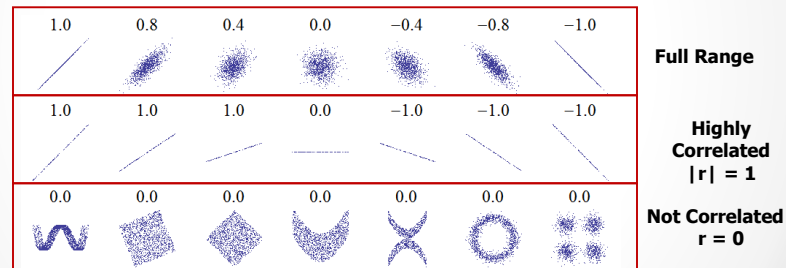
Prediction or Rating ↕	Your Rating	Movie Information	Wishlist List
★★★★★	Not seen ▾	Blood Diamond (2006) DVD info imdb flag Movie Tuner Action, Adventure, Crime, Drama, Thriller, War - English, Mende, Afrikaans [add tag] Popular tags: journalism 📺🔊🔊 politics 📺🔊🔊 action 📺🔊🔊	<input type="checkbox"/>
★★★★★	Not seen ▾	Elizabeth (1998) DVD VHS info imdb flag Movie Tuner Drama - English, French [add tag] Popular tags: true story 📺🔊🔊 British 📺🔊🔊 historical 📺🔊🔊	<input type="checkbox"/>
★★★★★	Not seen ▾	Hoosiers (a.k.a. Best Shot) (1986) DVD info imdb flag Movie Tuner Drama, Romance	<input type="checkbox"/>

• 1514
Lecture Notes Fall 2016 • 17

CF: How do we measure the similarity of the tastes b/w two users (e.g., x and y)?

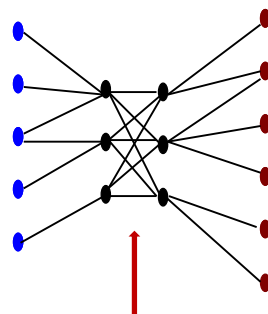
Pearson's Correlation Coefficient (PCC)

$$r_{xy} = z_x \cdot z_y \quad -1 \leq r_{xy} \leq 1$$



(a graph from Wikipedia)

Matrix Factorization: Introducing Hidden Layers



Dimension-reduced Representation of Users and Items

Topic Coverage

<http://nyc.lti.cs.cmu.edu/classes/11-741/f16/index.html>

- High-dimensional vectors & sparse matrices (1 lecture + HW1)
- Clustering (3 lectures + HW2)
- Link Analysis (2 lectures + HW3)
- Collaborative Filtering (3 lectures + HW4)
- Classification (3 lectures + HW5)
- Learning to Rank (2 lectures + HW6)
- Significance Testing (3 lectures + HW7)
- Others (SVD, Matrix Factorization, SGD, Deep Learning, 4 lectures)
- CPP (Capstone Project Proposal) (12 topics, more details later)

Administrative Details

Okay, that's the content ... now for the administrative stuff

Support System: Resources

- Textbooks (available at the bookstore)
 - Primary: [Introduction to Information Retrieval](#) (IR), Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, Cambridge University Press. 2008
 - Reference: [Pattern Recognition and Machine Learning \(ML\)](#), Christopher M. Bishop, Springer 2006
- Course materials via URL's
 - <http://nyc.lti.cs.cmu.edu/classes/11-741/f16/index.html> for homework assignments, lecture notes, copies of papers (when necessary) and data with restricted access to .cmu.edu (or via VPN)
 - <https://piazza.com/class/irv5bgvra6b33u>
 - <http://www.cmu.edu/blackboard/>

Course Organization

- ☐ **Lectures**
 - Problems (sub areas), Methods, Algorithms, Evaluations
- ☐ **Hands-on experience (homework)**
 - 5 programming assignments (Clustering, Link Analysis, etc.)
 - 2 problem solving exercises (basic matrix algebra, significance tests)
- ☐ **Weekly reading summaries**
 - Required for 11-741 students only (**but everyone is highly encouraged**)
- ☐ **Mid-term Exam: No**
- ☐ **Final Exam**
 - **Closed-book, no cheating sheets, no arrangement for exam-time exception**
- ☐ **CPP (Capstone Project Proposal)**
 - Team work (presentations + 4-page write up)

Difference among 741, 641 and 441

(More details in the syllabus)

□ 11-741

1. Count as a PhD-level course (12 units) in LTI
2. Reading summary per week is required
3. CPP work/peer-review is required; HW 2-7

□ 11-641

1. Count as a MS-level course (12 units) but not PhD-level in LTI
2. Reading summary is not required
3. CPP work/peer-review is required; HW 1-6

□ 11-441

1. Count as UG course (9 units)
2. Reading summary is not required
3. CPP work is not required, but CPP peer-review is required; HW 1-6

HW	11-741 Weight	11-641 Weight	11-441 Weight
HW0	0%	0%	0%
HW1	0%	5%	10%
HW2	5%	10%	10%
HW3	10%	10%	10%
HW4	12.5%	12.5%	15%
HW5	12.5%	12.5%	15%
HW6	10%	10%	10%
HW7	5%	0%	0%

©Yiming Yang

25

Grading

- Students in 11-741 (PhD level course) will be graded by 55% on 6 homework assignments (2-7), **17.5% CPP, 17.5% final exam** and 10% on reading summaries.
- Students in 11-641 (MS level course) will be graded by 60% on 6 homework assignments (1-6), **20% CPP and 20% final exam**. Reading summaries are not required but encouraged.
- Students in 11-441 (UG level course) will be graded by 70% on 6 homework assignments (1-6), **20% final exam, and 10% on peer-reviews of the CPP work** (including oral presentations and written reports) by the students in 11-741 and 11-641. CPP work, except the peer-review part, is not required for 11-441 students. Reading summaries are also not required but encouraged.

©Yiming Yang

Lecture Notes Fall 2016

26

CPP (Capstone Project Proposal)

1. Students are divided into teams, each focuses on one topic
2. Each student submits 3 candidate topics by **TBD**
3. Instructor/TA's assign one topic (**3 papers**) per team
4. Presentation on literature overview + new idea + proposed thorough work
5. Write-up proposal (**4 pages**)
6. Peer-reviewed evaluations on 4-6 (50%) + by TA's (50%) on 4-6 and review questions/comments
7. The entire team will receive the same evaluation score, so good coordination among team members is important for quality work.

Homework Policies

- All homework must be submitted via Blackboard
 - Due by 11:59 pm of the due date
- Late homework:
 - Deduct 10% for each day late.

Don't fall behind.

Cheating, Copying, Plagiarism, Etc

- You must be the author of everything that you submit for a grade
- Revising or modifying someone else's work does not make you the author
- It is okay to discuss homework with other students, share ideas, experience, and lessons learned
- Turn in the signed form (otherwise you will not be graded)

http://nyc.lti.cs.cmu.edu/classes/11-741/f15/Policy/cheating_policy_form.pdf

Cheating, Copying, Plagiarism, Etc

Penalties

- Usually failure of the course
- Possibly expulsion from the graduate program

If you are having problems meeting your deadlines

... submit the assignment late, or don't submit it at all

- Being late or taking a zero just lowers your grade
- Cheating causes you to fail the course