THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

Advanced Predictive Modeling

Lecture 1: Matrices, matrices, matrices

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

https://psarkar.github.io/teaching

What is a matrix

- In many applications, one comes across $m \times n$ matrices.
- You can think of matrix as a rectangular array or a table with m rows and n columns, where every element is a real number.
- Lets go over some examples.

Examples: Recommender systems

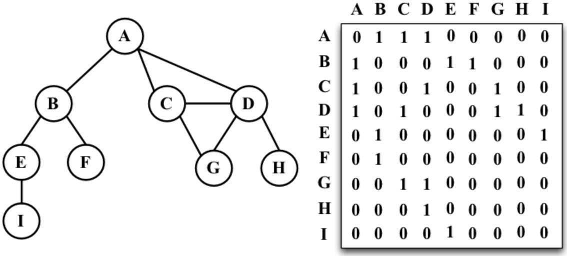
Alice	5	---	5	---	---	1
Bob	---	3	---	5	4	---
Reba	4	---	4	5	---	4

- Here, each row represents a user or customer.
- Each column represents a product
  - This can be a movie for Netflix
  - This can be a book for Amazon or Goodreads
  - This can be a product on Amazon
- The $(i, j)^{th}$ entry represents the rating provided by user i for product j
- Not all elements are observed, since not every customer has rated every product

Learning goals

- Matrix completion:
    - Given the “observed” entries, we want to infer the unobserved entries.
    - Helps in recommending new books/movies/music to users.
    - Typically, we pose this as an optimization problem, with suitable constraints on the learned matrix.

Examples: Adjacency matrices



	A	B	C	D	E	F	G	H	I
A	0	1	1	1	0	0	0	0	0
B	1	0	0	0	1	1	0	0	0
C	1	0	0	1	0	0	1	0	0
D	1	0	1	0	0	0	1	1	0
E	0	1	0	0	0	0	0	0	1
F	0	1	0	0	0	0	0	0	0
G	0	0	1	1	0	0	0	0	0
H	0	0	0	1	0	0	0	0	0
I	0	0	0	0	1	0	0	0	0

Figure 1: Courtesy: Oreilly.com

- Graphs or networks show up in a variety of machine learning and statistical applications.
- A graph consists of node set V and edge set E.
- An adjacency matrix A had rows and columns both corresponding to the nodes.
- A_{ij} has the weight of the edge from node i to node j.

Example: Political blogs data

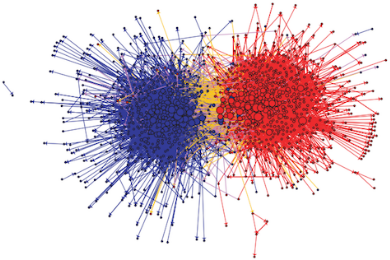


Figure 2: The political blogosphere and the 2004 U.S. election: divided they blog. L. Adamic and N. Glance

Example: Political blogs data

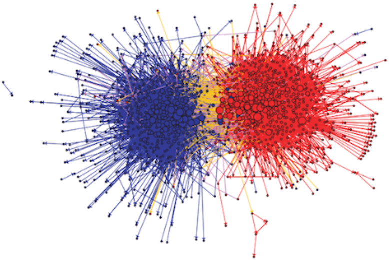


Figure 2: The political blogosphere and the 2004 U.S. election: divided they blog. L. Adamic and N. Glance

- Every node is a political blog.
- The link from blog i to j signify whether blog i points to blog j.
- This is a **directed** network.

Example: Political blogs data

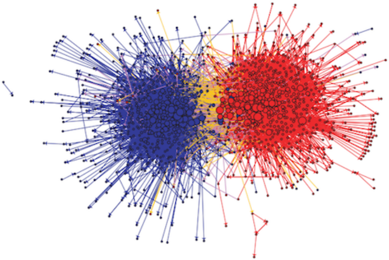


Figure 2: The political blogosphere and the 2004 U.S. election: divided they blog. L. Adamic and N. Glance

- Every node is a political blog.
- The link from blog i to j signify whether blog i points to blog j.
- This is a **directed** network.
- The colors signify the political orientation of a blog: blue for democratic and red for republican.

Learning goals

- Clustering
  - Given the “observed” network, can we learn which cluster each node belongs to?
  - Can we learn how many clusters are there in this network?
  - Inferring clusters is a key tool in exploratory data analysis and helps in a variety of applications like viral marketing, targeted advertising, etc.

Example: Facebook network

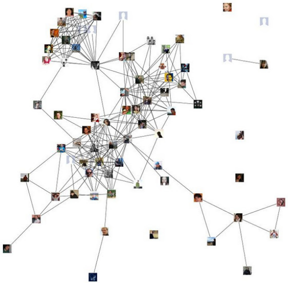


Figure 3: source: https://blog.revolutionanalytics.com/.

Example: Facebook network

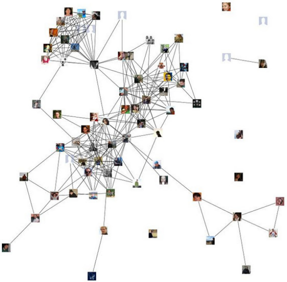


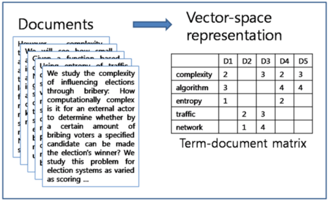
Figure 3: source: https://blog.revolutionanalytics.com/.

- Every node is a user.
- The link between node *i* to *j* signify whether *i* and *j* are.
- This is a **undirected** network.

Learning goals

- Link prediction
  - Can we predict future links?
  - This will lead to better friend recommendations in a social network.
  - This can also allow one to recommend which new blogs/news/products a user could be interested in.

Example: term-document matrix

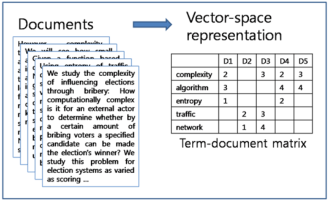


We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 4: source: Data Science authority, Quora.

Example: term-document matrix



We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure 4: source: Data Science authority, Quora.

- Every row corresponds to a term or work that appeared in a document, and each column represents a document.
- The $(i, j)^{th}$ entry represents the number of times word i appears in document j.
- A blank implies a zero, i.e. the word/term was not present in the document.

Learning goals

- Document representation
    - First task: just cluster documents into different topics
    - Second task: represent a topic as a mixture of topics and learn this mixture for each document.
- Document retrieval - given a query, can we retrieve the documents that are most relevant to it?