Data Preprocessing

Class Project: Any questions?

Class “homework” (recap from previous lecture)

Please research one-sentence definitions of:

- Collaborative Filtering
- Recommender Systems
- Filter Bubble
- Serendipity in Recommender Systems
- Difference between Fake News and Hate Speech

p.s. Please don’t ask me where to search them. Be smart ;)

Solutions (may or may not be right)

Collaborative Filtering: "Oh, you like that? Then you'll probably like this." It's how machines guess your next favorite thing based on what similar users enjoyed.

Recommender Systems: Algorithms that suggest what you might like next, from movies to products, based on your preferences and behaviors.

Filter Bubble: The cozy little online world where you only see stuff you agree with, because algorithms think differing opinions might hurt your delicate feelings.

Serendipity in Recommender Systems: When algorithms surprise you with unexpected recommendations that you end up loving, like stumbling upon a hidden gem.

Difference between Fake News and Hate Speech: Fake news is false information spread intentionally or unintentionally, while hate speech targets and attacks individuals or groups based on attributes like race, religion, or gender.

An attribute is a property or characteristic of an object

A collection of attributes describes an object

Object is also known as record, point, case, sample, entity, or instance

Attribute values are numbers or symbols assigned to an attribute for a particular object.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Different types of attributes

1. Nominal

   Examples: ID numbers, eye color, zip codes

2. Ordinal

   Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

3. Interval

   Examples: calendar dates

4. Ratio

   Examples: temperature in Kelvin, length, time, counts

Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Document Data

Each document becomes a `term' vector

- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

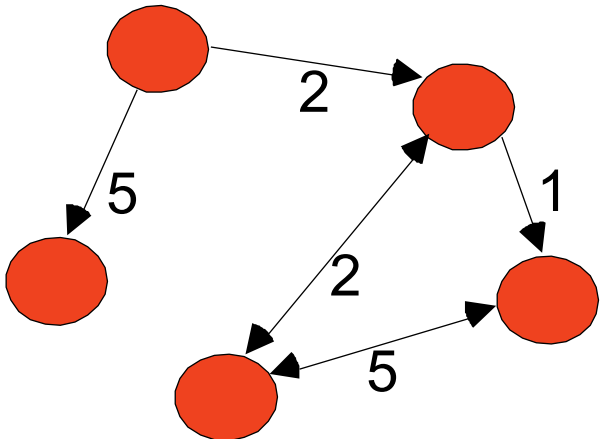
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

A special type of record data, where each record (transaction) involves a set of items.

For example: consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

Graph Data



Data Quality Problems

1. Noise and outliers
2. Missing values
3. Duplicate data
4. Wrong data