Data Acquisition, Storage, Linkage

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Things to remember

- Check access to data
- Next week
    - Assignment
        - Weekly review
        - ACS Data loading
    - Readings
- Proposal guidelines

Data Acquisition Challenges

- Political
- Internal Awareness
- Legal/Contractual
- Ethical
- Technical

Technical (challenges)

- How should you get data?
  - API access
  - Flat files
  - Database dumps
- How much should it be processed before you get it?
- How do you build a repeatable data acquisition pipeline?
- When do you collect new data?

Data Storage

- Use Databases whenever possible
  - Types of databases
- Deidentification
  - hashing

Data (Record) Linkage

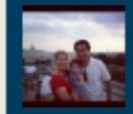
OBAMA BIDEN

Matt --

This is cool:

You can see exactly how many people named Matt have already voted.

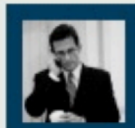
Take a look at that, then share it with some people you know -- like Izzy, Megan, and Burson who live in crucial battleground states -- so they can see how many people with their names have voted, and then look up their polling place.


Izzy Ortega


Megan Drapa


Burson Snyder


Matthew Gagnon


Eric Cantor


Zach Jason

Goals

- Determine if pairs of *records* describe the same entity

- Main applications:
  - *Joining* two different data sources
  - *Removing duplicates* from a single data source

Record Linkage: Synonyms

- (data) matching
- merge/purge
- duplicate detection
- de-duping
- reference matching
- co-reference/anaphora resolution

Factors to consider

- Deduping or Linkage
  - 1-1 or 1-many or many-1
- Rule-based or ML based
  - Do you have labeled training data?
- Domain specific or generic similarity metrics?
- Evaluation metric
  - Precision or recall
  - Task-specific - Implications on future analysis (bias for example)

Approaches

- Exact matching
- Rule-based
- Probabilistic linkage

Common reasons for mismatches

- Case (capital, lower case, etc.)
- Nicknames
- Prefixes
- Suffixes
- Initials
- Punctuation
- Spaces
- Digits
- Transpositions
- Abbreviations

When are two records about the same entity?

- Examples of possible similarity metrics
  - Edit distance
  - Soundex

“Fuzzy” Matching System

- Apply set of cascading rules
- Assign confidence score based on which rules fire

How do we not compare every pair?

- How do we avoid looking at |A| * |B| pairs?
- *Blocking:* choose a smaller set of pairs that will contain all or most matches.
  - Simple blocking: compare all pairs that “hash” to the same value (e.g., same Soundex code for last name, same birth year)
  - Extensions (to increase *recall* of set of pairs):
    - Block on *multiple* attributes (soundex, zip code) and take union of all pairs found.
    - *Windowing:* Pick (numerically or lexically) *ordered* attributes and sort (e.g., sort on last name). The pick all pairs that appear “near” each other in the sorted order.

Machine Learning based Record Linkage

- Generate training data
  - Label pairs as match/no match
- Generate features over each pair
  - Distance metrics over different attributes (fname, lname, dob, etc.)
  - Tfidf scores
- Build and evaluate classifiers

One-off versus recurring matching

- Unique identifiers: persistence?
- What do we do with new or changed pairs?

Things to remember

- Check access to data
- Next week
    - Assignment
        - Weekly review
        - ACS Data loading
    - Readings
- Proposal guidelines