36-350 Data Mining

Tom Minka

Day 4

How to simplify your life

by abstracting your variables

Last time: histograms

Histograms require abstraction

- Words (100,000 bins, excluding digits, punctuation)
- Colors (millions, reduced to 64)
- Sub-images (zillions, reduced to 6,500)
- Call details (reduced to time/location bins)
- UNIX commands (reduced to application name)

Abstracting variables

- Most data mining methods require some kind of data abstraction
- Data is often more detailed than needed
  - Too many categories, too much precision
- Extra subdivisions can hide trends
  - focusing you on noise, sampling variation
  - E.g. day to day sales, phone call time/duration
- Want to reduce the number of values without jeopardizing results

Preprocessing in data mining

- Big part of many mining tools
- Active area of research
- Several books on it, most discuss it
- Data given to statistics students is already preprocessed to right level of abstraction

Reasons to abstract

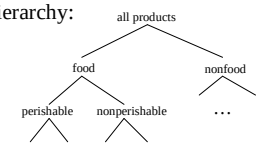
- Simpler description, visualization
- Faster computation
- Better density estimation
- Easier discrimination of groups

Example

- Supermarket wants to learn buying profile of different customers, from receipts
- Store carries thousands of products
- A fraction will have never been bought by a given customer
- Inventory changes over time
- How would you represent the information?

Product hierarchy

- Can't use histogram over specific products
- Can divide products according to some level in a product hierarchy:



- Make per-customer histograms over some level of abstraction

Other possibilities

- Client is interested in store areas customer visits
  - Group products by aisle
- Client wants to compare weekend vs. weekday shopping
  - Histogram on (product type, day)

Conceptual hierarchies

- Important first step in data mining
- Based on domain knowledge, experts
- "What are the ways this variable could be abstracted?"
- Appropriate level depends on question
  - May need to change levels during data mining

Examples

- Address:
  - Street - City – State – Region - Country
- Date/time:
  - Absolute: Day since 1970, month since, year
  - Relative: day of year, month of year, season
  - Holiday, weekend, business hours
- Price: expensive, moderate, inexpensive
- Occupation

Abstracting text

- Reduce numbers to "#"
- Match proper names
- Remove stopwords ("the")
- Stemming ("walk" vs. "walking")
- Synonyms, senses ("bank" vs. "bank")
- Topics (agriculture, economy, …)
- Part of speech, grammatical position

Abstracting colors

- No single hierarchy possible
- Perceptual studies have established 3D neighborhood structure
- Hue most important dimension
- Abstract on brightness, then hue