

Data Mining Techniques
Chapter 8
Nearest Neighbor Approaches: Memory-based Reasoning
and Collaborative Filtering

Memory-based reasoning (MBR) 2
MBR in practice 3
Rents example. 4
MBR challenges. 5
Case study on news stories 6
Measuring distance 7
Building distance functions 8
Distance functions continued 9
The combination function 10
Collaborative filtering 11
Collaborative filtering example 12

Memory-based reasoning (MBR)

- Reason from experience by recognizing similar examples from the past.
- Database of known records searched to find preclassified records similar to a new (unclassified) record:
 - the similar records (or neighbors) are used for classification and estimation.
- Requires:
 - *distance* function to calculate similarity of two records;
 - *combination* function to combine results from several neighbors.

© Iain Pardoe, 2006

2 / 12

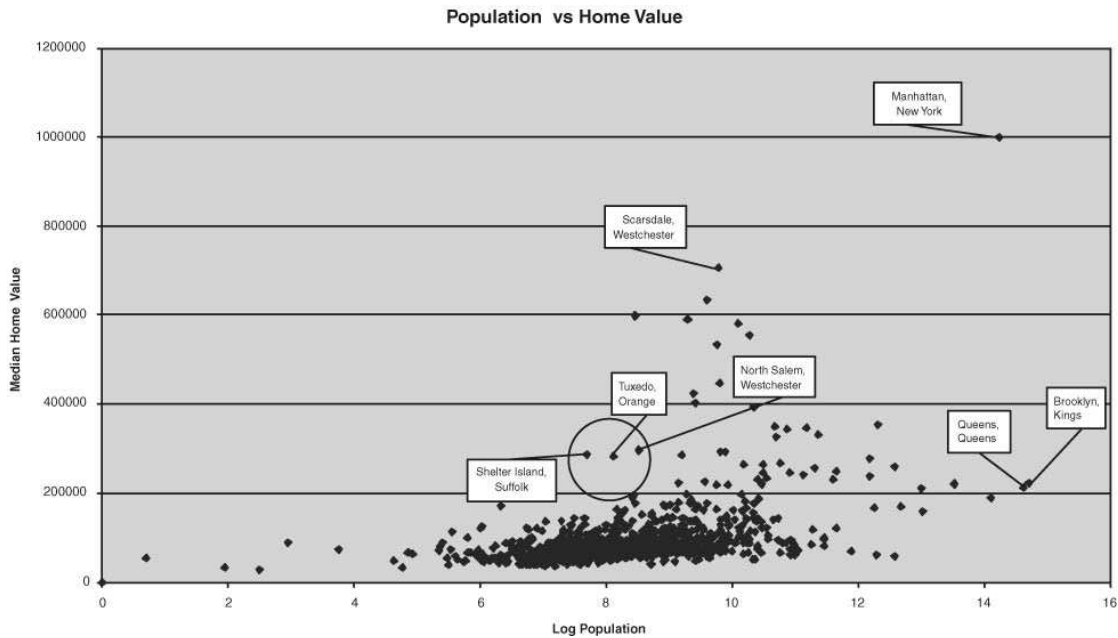
MBR in practice

- Advantages:
 - can use data “as is,” even data such as images and free-text;
 - able to adapt easily to new data.
- Disadvantages:
 - resource intensive (training sample *is* the model).
- Example: using MBR to estimate rents in Tuxedo, NY by using data on rents in similar towns (p. 259–262, beware some numerical mistakes).

© Iain Pardoe, 2006

3 / 12

Rents example



Town	Pop.	Med. rent	<\$500	\$500–750	\$750–1k	\$1–1.5k	>\$1.5k	No rent
SI	2228	\$804	3.1	34.6	31.4	10.7	3.1	17.0
NS	5173	\$1150	3.0	10.2	21.6	30.9	24.2	10.2

© Iain Pardoe, 2006

4 / 12

MBR challenges

- Choosing a *balanced* set of historical records for the training sample:
 - rather than a *random* sample, try to have equal numbers of records representing different categories.
- Finding distances from unknown case to all historical records, then picking k smallest distances can be very computer-intensive:
 - solution—replace training set with “cluster centers.”
- Determining the distance function, combination function, and number of neighbors to use.

© Iain Pardoe, 2006

5 / 12

Case study on news stories

- Classification codes are keywords describing content.
- Training set: nearly 50,000 news stories.
- Distance function: relevance feedback measures similarities of two documents.
- Combination function: most common codes for nearest neighbors.
- Choosing the number of neighbors: experiment between 1 and 11.
- Results compare computer predictions with human editors, e.g.:

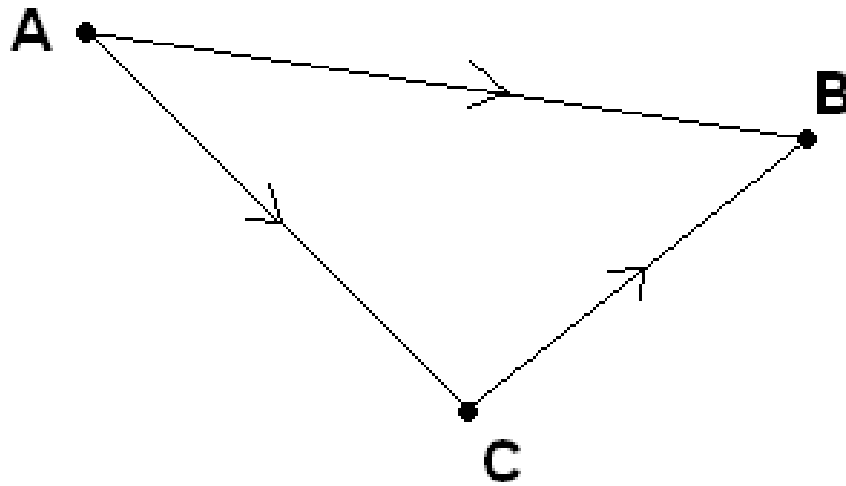
MBR	assigned	not	total	Human	assigned	not	total
correct	1760	440	2200	correct	1826	374	2200
incorrect	684			incorrect	249		
total	2444			total	2075		

Recall: $1760/2200 = 80\%$, $1826/2200 = 83\%$

Precision: $1760/2444 = 72\%$, $1826/2075 = 88\%$

Measuring distance

- Non-negative number, $d(A, B) \geq 0$.
- Identity, $d(A, A) = 0$.
- Commutative, $d(A, B) = d(B, A)$.
- Triangle inequality, $d(A, B) \leq d(A, C) + d(C, B)$.



Building distance functions

- Quantitative variables, one at a time:
 - absolute difference, $|A - B|$;
 - standardized, $|A - B| / \text{max. difference}$;
 - alternatively, $|A - B| / \text{sd}$;
 - squared difference, $(A - B)^2$.
- Qualitative/categorical, one at a time:
 - 1 if match, 0 otherwise.
- Merge individual distances together:
 - Manhattan (city block), $d_{X_1}(A, B) + d_{X_2}(A, B) + d_{X_3}(A, B)$;
 - Standardized, divide by maximum;
 - Euclidean, $\sqrt{d_{X_1}(A, B)^2 + d_{X_2}(A, B)^2 + d_{X_3}(A, B)^2}$.
- Example: class exercise.

© Iain Pardoe, 2006

8 / 12

Distance functions continued

- Distances for other data types, e.g., zip codes:
 - $d(A, B) = 0$ if zip codes are identical;
 - $d(A, B) = 0.1$ if first three digits are identical;
 - $d(A, B) = 0.5$ if first digits are identical;
 - $d(A, B) = 1$ if first digits are different.
- Incorporating customer behavior, e.g., offer history:
 - $d(A, B) = 0$ if A and B responded to the offer;
 - $d(A, B) = 0.1$ if A and B were sent the offer, but neither responded;
 - $d(A, B) = 0.2$ if A and B were not sent offer;
 - $d(A, B) = 0.3$ if A and B were sent the offer, but only one responded;
 - $d(A, B) = 1$ if one was sent the offer and the other was not;

© Iain Pardoe, 2006

9 / 12

The combination function

- Ask the neighbors for the answer, e.g., average for a quantitative target.
- Basic approach—democracy (winner is the category with the most votes).
- Generalization—weighted voting (closer neighbors have stronger votes than neighbors farther away).
- For classification problems, use proportion of neighbors voting for the winner as a “probability.”

© Iain Pardoe, 2006

10 / 12

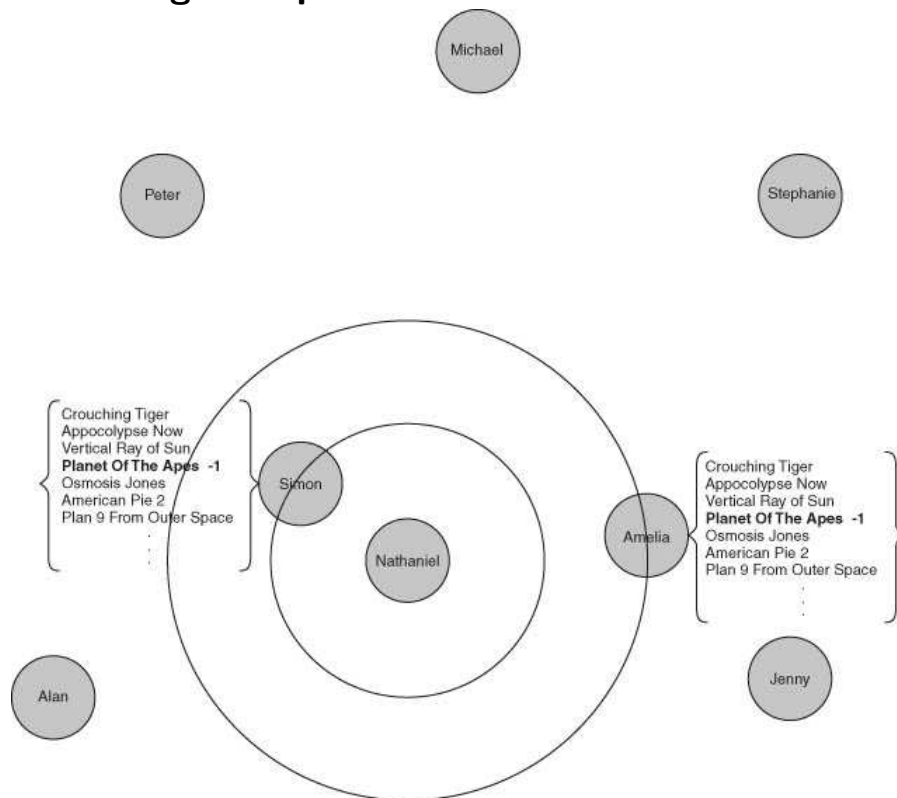
Collaborative filtering

- A nearest neighbor approach to making personalized recommendations (AKA social information profiling).
- Three steps:
 1. Build profiles—customers rate a selection of items.
 2. Compare profiles—measure distances between customers based on their profiles.
 3. Make predictions—combine profiles of nearby customers to predict ratings for items a particular customer has not yet rated.
- Examples:
 - Textbook p. 284 (next slide).
 - Netflix.

© Iain Pardoe, 2006

11 / 12

Collaborative filtering example



© Iain Pardoe, 2006

12 / 12