CIS4301 Notes: Data Mining

Ryan Roden-Corrent

April 11, 2014

1 Frequent Item Sets

Frequent Item Sets are an artifact of the data. For example:
What is the higest selling product in the DB?

2 Associative Rule Mining

For example, one notices from the DB that whenever beer is sold, diapers are sold.
$Beer \rightarrow Diapers$ | Represents association from Beer to Diapers
A store could draw the that men who are sent out to buy diapers also decide to buy beer, and decide to place diapers and beer in close proximity.

2.1 Collaborative Filtering

Suppose you have a table of Users and Movies. For each row, you know what movie each user has watched.

Table 1: Example Table

User1	Independence Day	Godfather	I am Legend
User2			I am Legend

Might consider that User2 would like to watch Independence Day or The Godfather because User1 also watched Independence Day.

2.1.1 The Netflix Prize

Netflix offered $1M to the person who could come up with the most efficient algorithm for performing such a computation.

3 Similarity

3.1 Jaccard distance

$JaccardDistance = \frac{|A \cap B|}{|A \cup B|}$
Usually used for boolean values

Table 2: Jaccard Computation

1	0	1	0
1	0	0	0

3.2 Euclidean Distance

$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 ...}$

Table 3: Jaccard Computation

price	tax
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$

3.3 Clustering

Visualize a 2D dataset on a plane - clusters may become obvious. These represent data items that are significantly closer to some set of values than others.

3.3.1 K-means Clustering

Find **centroids**for each cluster. For each point in the data set, assign it to its closest **centroid** (using some distance calculation like Euclidean or Jaccard). Adjust centroids to minimize the distances within each cluster. Then compute a new centroid, and keep repeating until the centroids no longer move.
A demo is available HERE