Latent Semantic Indexing

Jean Mark Gawron

San Diego State University

March 12, 2015

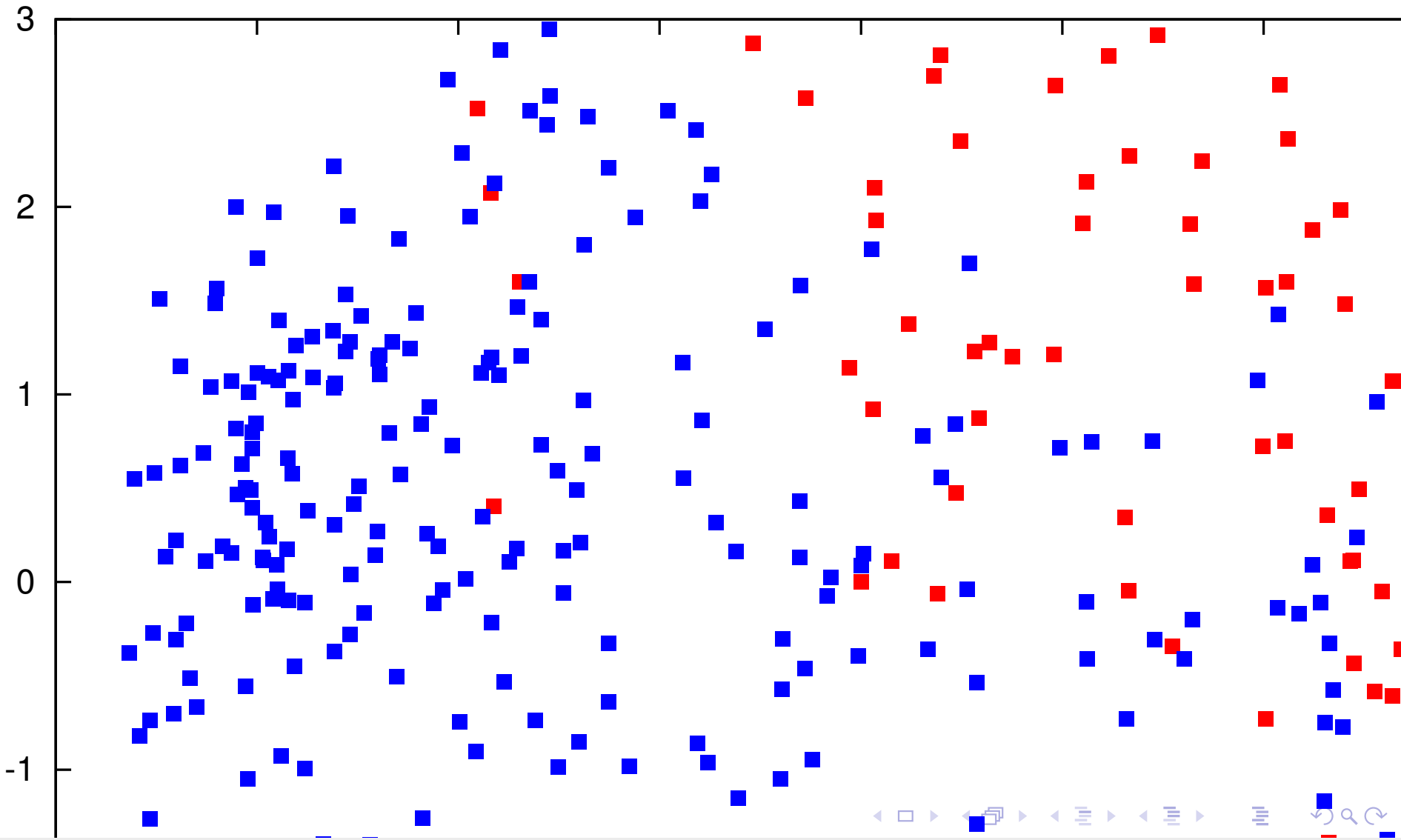
Outline

Latent Semantic Indexing

LSI

LSI is a technique for reducing the dimensionality of matrix representations of relations LSI was introduced in Deerwester (1990) as a strategy for information retrieval (IR). as solution to certain classic problems of document retrieval.

House data freduced to 2D



Bills

	Bill
V1.	handicapped-infants
V2.	water-project-cost-sharing
V3.	adoption-of-the-budget-resolution
V4.	physician-fee-freeze
V5.	el-salvador-aid
V6.	religious-groups-in-schools
V7.	anti-satellite-test-ban
V8.	aid-to-nicaraguan-contras
V9.	mx-missile
V10.	immigration
V11.	synfuels-corporation-cutback
V12.	education-spending
V13.	superfund-right-to-sue
V14.	crime

Representation of a member

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	n	y	n	y	y	y	n	n	n	y	NA	y	y
2	n	y	n	y	y	y	n	n	n	n	n	y	y
				...									

Converting to numbers

The vote vector for two members:

$$\begin{array}{cccccccccccccccc} -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 0 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 0 \end{array}$$

The similarity of 2 members

$$\text{sim}(M[1],\ M[2]) = \sum_i M[1,i] * M[2,j]$$

What this says is that For any given bill i, we do

$$M[1,i] * M[2,i]$$

This is the product of $M[1]$'s vote on bill i with $M[2]$'s vote on bill i. Given the possible vote values are 0, 1, -1, we have :

	r_2 vote		
r_1 vote	-1	0	1
-1	1	0	-1
0	0	0	0
1	-1	0	1

Dot product

$$M[1] \cdot M[2] = \sum_j M[1][j] * M[2][j]$$

The dot product of vector $M[1]$ with vector $M[2]$

Similarity

Similarities as dot products

With centered data:

1. Covariance is a dot product
2. Correlation is a dot product

Dot product is not always the answer, but it does cover a lot of cases sensibly.

How to use it

1. Convert data (by operations such as centering) into a form in which dot product captures similarity intuitions.
2. This is essentially what we did with the house voiting data.

	r_2 vote		
r_1 vote	-1	0	1
-1	1	0	-1
0	0	0	0
1	-1	0	1

A shared vote makes a positive contribution to the similarity score; a disagreement makes a negative contribution.

Reducing dimensions

Suppose

$$v_1 \cdot v_2 > v_3 \cdot v_4$$

Then v_1 is more similar to v_2 than v_3 is to v_4.

Now suppose we reduce all these vectors to a 2D representation with some transformation of the data we'll call T. Then we'd like it to be the case that the similarity relations are preserved as much as possible. That is, we'd like it to be the case that:

$$T(v_1) \cdot T(v_2) > T(v_3) \cdot T(v_4)$$

as much as possible.

Similarity matrix

imagine an even bigger table than the one we started with, that represents all the similarity relations in our original data. That is, we want a table S such that

$$S[i,j] = M[i,*] \cdot M[j,*] \tag{1}$$

In terms of our our original data set with 435 house members and 16 votes, this is a much larger 435 x 435 table in which each cell (i,j) represents the similarity in the voting record of house member i with house member j.

Statement of problem

Dimensionality reduction preserving similarity

We want a table S_2 which is based on M_2, a 2-dimensional representation of the data in M in the following way:

$$S_2[i,j] = M_2[i] \cdot M_2[j].$$

And we want the similarity values in S_2 to match those in S as well as possible.

Our goal in reducing dimensions

How do we measure the similarity of two tables?

Discrepances	X	$S - S_2 = X$
Sum of squares	$\| X \|$	$\| X \| = \sqrt{\sum_i^M \sum_j^N X[i,j]^2}$
Goal		The S_2 that minimizes $\| X \|$

What about the picture?

Is our goal in looking for similarity-preserving transformation going to help us draw a picture? Now that I know X and Y are similar, where do I put them?

The cosine theorem

$$\text{dist}^2(v_1, v_2) = \mid v_1 \mid + \mid v_2 \mid -2v_1 \cdot v2$$

Suppose vectors all have the same length:

$$\text{dist}^2(v_1, v_2) = 2(k^2 - v_1 \cdot v_2)$$

So when dot product (similarity) gets bigger, the distance grows smaller.

Enter matrix products

1. Linear algebra is about large tables consisting of ordered sequences of vectors of any size filled with real numbers, like our vote table M.
2. They're called **matrices**
3. A compact way of stating how e got the similarity table:

$$S = MM'$$

S is the **matrix product** of M with M' Here M' denotes the **transpose** of M, that is, the matrix you get by exchanging the rows and columns of M.

$$M[i, j] = M'[j, i]$$

So if M is a matrix with i row and j columns, M' is a matrix with j rows and i columns.

Definition/Example

The matrix product MN can be defined in terms of dot products of rows of M with columns of N. The (i, j) cell of the MN is the dot product of the ith row of M with the jth column of N

$$MN[i, j] = M[i, *] \cdot N[*, j]$$

Here we use $M[i, *]$ for the ith row of M, and $N[*, j]$ for the j row of N. So

$$S[i, j] = M[i, *] \cdot M'[*, j]$$

And since M' just exchanges the rows and columns of M, this is:

$$S[i, j] = M[i, *] \cdot M[j, *] \tag{2}$$

This is our original definition of S.

Summary

1. M yields similarity table S.
2. The values in S are computed by taking dot products of the rows in M.
3. There is a compact algebraic expression for defining S as the matrix product of M with its transpose.
4. We want a mapping of the 16-dimensional (16 column) data in M to a two column matrix M_2 which is going to yield similarity table S_2.
5. We want to minimize the Frobenius Norm of S and S_2. (a measure of the discrepancy).

SVD

Any matrix can be factored into a matrix product of 3 other matrices, which is called its SVD:

So what?

We replace S with a diagonal matrix S_2 with only the two largest Eigenvalues and 0';s everywhere else, we get an approximation of X we'll call X_2, which has rank 2. An important theorem due to Eckhart and Young (1963) shows that this approximation is the closest rank 2 approximation of X, in the sense that the discrepancy between X and X_2 has a smaller Frobenius norm than the discrepancy between X and any other rank 2 matrix. It turns out that the SVD of the similarity matrix of X is very closely related to the SVD of its simarity matrix S:

$$S = XX' = TS^2T'$$

And the solution to our dimensionality reduction problem, the best rank 2 approximation of S, is therefore

$$S_2 = TS_2^2T'.$$