Assignment 2 Applied Mathematics

-Mohit Vaishnav

Google Page Rank Algorithm:

PageRank is one of the methods Google uses to determine a page's relevance or importance. Google models the algorithm in the link structure in the form of directed graph where the nodes are webpages and the links from one webpage to another form the edges which also shows the direction of movement. A ***Hyperlink Matrix*** is created which is represented with the weighted links in turns showing the likelihood that a link from webpage *j* is selected from *i*. Here a specific situation also occurs, known as a ***Dangling Node*** which does not link to other nodes. To tackle the situation of dangling nodes there are various options. One amongst them is using the probability distribution vector "*w*" which is a non-negative with summation as 1. Now the resulting matrix is a combination of *H* and the column vector multiplied with "*w*". For example, "*w"* will have all the entries as *1/N* if there are in total of *N + 1* node into the system. Now to model the overall behavior of the system, Google forms a matrix using some damping factor (mostly accepted to be 0.85 but still very less is known in literature) less than 1. A uniform "***personalization vector***" which is a row probability distribution is used in the G which sometimes leads to "***Link Spamming***". This is the practiced by search engine optimization experts who add more and more links in their clients webpage to increase the ranking. So it becomes one of the reasons for not disclosing much about the ***damping factor*** *a*. Hence the matrix *G* has all the elements <=1 whose summation is 1, so it is also known as "***Row Stochastic Matrix or Markov Matrix***". This leads to $\lambda = 1$ as one of the solution also known as dominant eigenvalue of G and the other $\lambda = \pi$ as the dominant left eigenvector.

To find the page rank score of the matrix which has billions of rows and columns, power method is used where the *G* is decomposed using the eigenvalues and eigenvector. To compute the score of the page rank, repeated computation of *G* is required using the page rank vector which at some stage reaches the stability condition and the system is known to be converged. Using the definition of G and calculating the value of matrix, we require only one matrix vector multiplication i.e. *H* which is in itself is usually a sparse matrix. Power of convergence of the matrix is given by the ratio of the highest to lowest of the eigenvalues.

Google also has a pagerank display feature which is an indication of the page rank. It ranges from 0 (lowest) to 10 (highest). Google is yet to disclose more about this *toolbar feature* but it is possibly based on a logarithmic score.

Eigenvalue problem solution:

Eigenvalue and the corresponding eigenvectors are calculated for the G matrix which is written in the form of

$$G = V\Lambda V^{-1}$$

Where,

$$\Lambda = \begin{bmatrix} \lambda & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda \end{bmatrix}$$

$$\pi^k = G\pi^{k-1}$$

$$\pi^k = G^k\pi^0$$

$$\pi^k = V\Lambda^k V^{-1}\pi^0$$

Because of the $\lambda < 1$, for very high value of k, system has only one eigenvalue- ignoring other values which tends to 0.

Let the matrix be:

$$A = \begin{bmatrix} A & B & C & D & E & \\ 0 & 1 & 1 & 1 & 0 & A \\ 1 & 0 & 1 & 0 & 1 & B \\ 0 & 1 & 0 & 0 & 1 & C \\ 0 & 0 & 0 & 0 & 1 & D \\ 0 & 0 & 1 & 0 & 0 & E \end{bmatrix}$$

$$H = \begin{bmatrix} 0 & 0.5 & 0.33 & 1 & 0 \\ 1 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.5 & 0 & 0 & 0.33 \\ 0 & 0 & 0 & 0 & 0.33 \\ 0 & 0 & 0.33 & 0 & 0 \end{bmatrix}$$

This is a Markov matrix and $\lambda = 1$.

Let the vector $\pi^0 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$ all values as 1/5 = 0.2

Mohit Vaishnav
Assignment 2
Applied Mathematics

$$\pi^1 = \begin{bmatrix} 0.37 \\ 0.33 \\ 0.17 \\ 0.07 \\ 0.07 \end{bmatrix}$$

$$\pi^2 = H\pi^1 = \begin{bmatrix} 0.29 \\ 0.44 \\ 0.19 \\ 0.02 \\ 0.05 \end{bmatrix}$$

.

.

$$\pi^9 = H\pi^8 = \begin{bmatrix} 0.28 \\ 0.38 \\ 0.21 \\ 0.02 \\ 0.07 \end{bmatrix}$$

This satisfies the criteria (computed using MATlab) and hence represents the PageRank of 5 webpages.

It is easily checked that $P\pi = \pi$. If we consider the i-th coordinate of the vector π^t as the probability of being on page i at a given time n, and hence π^t as the probability distribution of pages at time n, then it is also the probability distribution at time $n + 1$. For this reason, the vector π is called the *stationary distribution*. This stationary distribution allows to order the pages. In our example, we order the pages as B, A, C, E, D, and we declare B the most important page.

MATlab Code to Compute the PageRank.

```
N = input('Enter the number of webpages: ')
for(i = 1:N)
    for(j = 1:N)
        j
        i
        A(j,i) = input('Enter the values for i column: ')
    end
    H(:,i) = A(:,i)/(sum(A(:,i)));
    v(1,i) = 1/N;
end
%Solving Dangling Node Fix:
T(1:N, 1:N) = 1/N;
a = input('Enter the damping factor value: ')
G = a*A + (1-a)*T;
iter=input('Enter the number of iterations: ');
while(i<iter)
    v2 = G*v;
    v = v2;
    i = i + 1;
end
printf('PageRank for the matrix is: ')
v
```

Problem 2:

In a town called Computer Vision Village, the local newspaper The Computer Visionist has determined that a citizen who purchases a copy of their paper one day has 70% chance of buying the following day's edition. They have also determined that a person who does not purchase a copy of The Computer Visionist one day has 20% chance of purchasing it the next day. Records show that of the 1000 citizens of Computer Vision Village, exactly 750 purchased a copy of the newspaper on Day 0. To determine the appropriate amount of papers to press each day, the owner of The Computer Visionist, Mr Marr Rosenfeld, is interested the following types of questions:

1. If a person purchased a paper today, how likely is he to purchase a paper on Day 2? Day 3? Day n?
2. What sales figures can The Computer Visionist expect on Day 2? Day 3? Day n?
3. Will the sales figures fluctuate a great deal from day to day, or are they likely to become stable eventually?

Answer:

Let

g_1 be the group who buys the paper

g_2 be the group who do not buys the paper.

The above problem can be represented in the Markov matrix as:

$$g^1{}_1 = 0.7g^0{}_1 + .2g^0{}_2$$
$$g^1{}_2 = 0.3g^0{}_1 + .8g^0{}_2$$

$$G^1 = \begin{bmatrix} g11 \\ g12 \end{bmatrix}$$
$$M = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$$
$$G^0 = \begin{bmatrix} g01 \\ g02 \end{bmatrix}$$

$$G^1 = MG^0$$
$$G^2 = MG^1 = M^2G^0$$
.
.
$$G^k = MG^{k-1} = M^kG^0$$

To calculate the value of M^k we have to use diagonalization of matrix which is computed using the eigenvalues and eigenvector calculation.

Eigenvalues and Eigenvectors of this matrix are:

$\lambda = 1$

$v_1 = \begin{bmatrix} -0.55 \\ -0.83 \end{bmatrix}$

And

$\lambda = 0.5$

$v_2 = \begin{bmatrix} -0.71 \\ 0.71 \end{bmatrix}$

Hence,

$V = \begin{bmatrix} -0.71 & -0.55 \\ 0.71 & -0.83 \end{bmatrix}$

$V^{-1} = \begin{bmatrix} -0.85 & 0.57 \\ -0.72 & -0.72 \end{bmatrix}$

$\Lambda = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}$

$M^k = V\Lambda^k V^{-1}$

$$= \begin{bmatrix} 0.6 * (0.5^k) + 0.4 & -0.4 * (0.5^k) + 0.4 \\ -0.6 * (0.5^k) + 0.6 & 0.4 * (0.5^k) + 0.6 \end{bmatrix}$$

$\lim_{n\to\infty}(M)^k = \begin{bmatrix} 0.4 & 0.4 \\ 0.6 & 0.6 \end{bmatrix}$

Likelihood of purchasing on day 2, i.e. k =2

$M^2 = \begin{bmatrix} 0.55 & 0.30 \\ 0.45 & 0.70 \end{bmatrix}$

Sales on day 2,

$M^2G^0 = \begin{bmatrix} 0.55 & 0.30 \\ 0.45 & 0.70 \end{bmatrix} \begin{bmatrix} 750 \\ 250 \end{bmatrix}$

$= \begin{bmatrix} 488 \\ 512 \end{bmatrix}$

Likelihood of purchasing on day 2, i.e. k =3

$M^3 = \begin{bmatrix} 0.475 & 0.35 \\ 0.525 & 0.65 \end{bmatrix}$

Sales on day 2,

$M^3G^0 = \begin{bmatrix} 0.475 & 0.35 \\ 0.525 & 0.65 \end{bmatrix} \begin{bmatrix} 750 \\ 250 \end{bmatrix}$

$$= \begin{bmatrix} 444 \\ 556 \end{bmatrix}$$

Likelihood of purchasing on day n, i.e. k =n

$$M^n = \begin{bmatrix} 0.6 * (0.5^k) + 0.4 & -0.4 * (0.5^k) + 0.4 \\ -0.6 * (0.5^k) + 0.6 & 0.4 * (0.5^k) + 0.6 \end{bmatrix}$$

Sales on day n,

$$M^n G^0 = \begin{bmatrix} 350 * (0.5^n) + 400 \\ -350 * (0.5^n) + 600 \end{bmatrix}$$

As k → ∞ sales are likely to become stable as we can see from the above equation and they tends to 400 people buying the magazine and rest not purchasing it. With very large value of k, eigenvalue will become almost negligible and eventually have no effect on the stability of the system. Thereby the system reaches to stability.