Mathematics of Image and Data Analysis
Math 5467

Lecture 8: PageRank

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467S21

Last time

- Spectral Clustering

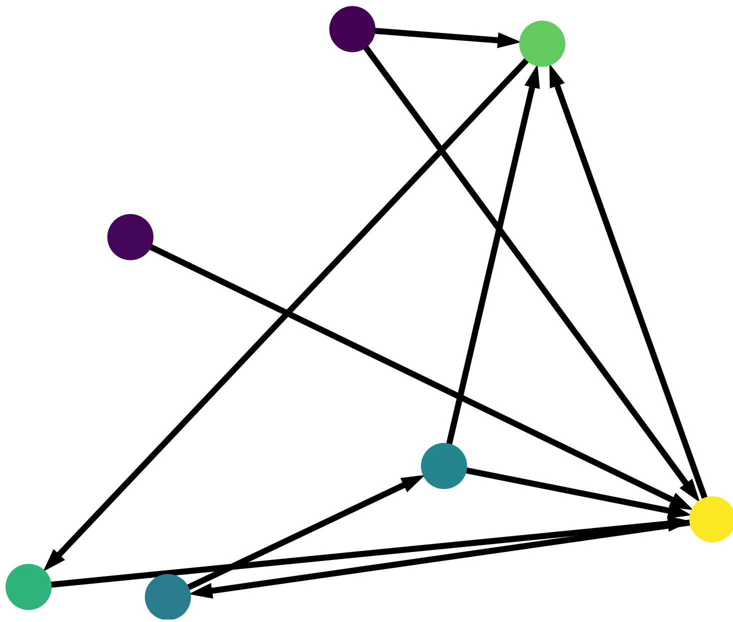
Today

- PageRank

PageRank

The PageRank algorithm ranks websites based on the link structure of the internet. It was used to sort Google search results until 2006, and has been used in

- Biology (GeneRank), chemistry, ecology, neuroscience, physics, sports, and computer systems...



PageRank

Main Idea: Take a random walk on the internet for T steps.

$$\text{Rank of site } i = \lim_{T\to\infty} \frac{1}{T}(\text{Number of times site } i \text{ is visited}).$$

Problem: Random walks can get stuck in disconnected components of the internet, and may never visit a given site i.

Solution: Every so often, the random walker teleports to a random site on the internet. The walker is called a random surfer.

Code demo

Mathematics of PageRank

To describe PageRank mathematically, we start with an adjacency matrix W

$$W(i,j) = \begin{cases} 1, & \text{if site } i \text{ links to site } j \\ 0, & \text{otherwise.} \end{cases}$$

We also have a probability transition matrix P for the random walk:

$$P(i,j) = \text{Probability of stepping from } j \text{ to } i.$$

Both P and W are $n \times n$ matrices, $n =$number of webpages.

Mathematics of PageRank

Clicking on a link at random from webpage j leads to the transition probabilities

$$P(i,j) = \frac{W(j,i)}{\sum_{k=1}^{n} W(j,k)}.$$

Exercise 1. Show that $P = W^T D^{-1}$, where D is the diagonal matrix with diagonal entries $D(i,i) = \sum_{j=1}^{n} W(i,j)$. $\triangle$

Random surfer

Let $\alpha \in [0, 1)$ be the random walk probability, and let $v \in \mathbb{R}^n$ be the teleportation probability distribution. That is, $v(i) \geq 0$ for all i, and $\sum_i v(i) = 1$.

Random surfer dynamics: When at website j, the random surfer chooses the next site as follows:

1. With probability α the surfer clicks an outgoing link at random, that is, the surfer navigates to website i with probability $P(i, j)$.

2. With probability $1 - \alpha$ the surfer teleports to website i with probability $v(i)$.

Teleportation

Teleportation distribution: Common choices are

- $v(i) = 1/n$ for all i (jump to a site uniformly at random).
- (Localized PageRank) $v(i) = \delta_{ij}$ (always jump back to site j).

Localized PageRank ranks all sites based on their similarity to site j.

The PageRank vector

For $k \geq 0$ define

$$x_k(i) = \text{Probability that the random surfer is at page } i \text{ on step } k.$$

Definition 2. The **PageRank** vector x is

$$x(i) = \lim_{k \to \infty} x_k(i),$$

provided the limit exists.

Transition probabilities

To see how x_k transitions to x_{k+1} requires some probability. We condition on the location of the surfer at step k, and on the outcome of the coin flip, to obtain

$$x_{k+1}(i) = (1-\alpha)v(i) + \alpha \sum_{j=1}^{n} P(i,j)x_k(j).$$

$(Px_k)(i)$

We can write this in matrix/vector form as

(1) $$x_{k+1} = (1-\alpha)v + \alpha P x_k.$$

If x_k converges to a vector x as $k \to \infty$, then x should satisfy

$$x = (1-\alpha)v + \alpha P x.$$

Question: Does x_k converge as $k \to \infty$, and if so, how quickly does it converge?

Analysis of PageRank

We consider the PageRank equation

$$x = (1 - \alpha)v + \alpha P x. \tag{2}$$

Lemma 3. *Let $v \in \mathbb{R}^n$ and $0 \leq \alpha < 1$. Then there is a unique vector $x \in \mathbb{R}^n$ solving the PageRank equation (2). Furthermore, the following hold.*

(i) We have $\sum_{i=1}^n x(i) = \sum_{i=1}^n v(i)$. ← if $\sum v(i) = 1$ then $\sum x(i) = 1$

(ii) If $v(i) \geq 0$ for all i, then $x(i) \geq 0$ for all i.

↑ if $v(i)$ are probabilities so are $x(i)$.

The ℓ_1-norm

Recall $\|x\| = \sqrt{\sum x(i)^2}$ Euclidean norm

$\|x\| = \|x\|_2$

It will be more convenient to work in the ℓ_1-norm $\|\cdot\|_1$ defined by

$$\|x\|_1 = \sum_{i=1}^{n} |x(i)|.$$

In the ℓ_1-norm, the transition matrix P is non-expansive.

Proposition 4. *We have* $\|Px\|_1 \le \|x\|_1$.

Proof: Recall $\sum_{i=1}^{n} P(i,j) = 1$ (*)

probability of $j \to i$

$$\|Px\|_1 = \sum_{i=1}^{n} |Px(i)| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} P(i,j)\, x(j) \right|$$

$$\le \sum_{i=1}^{n} \sum_{j=1}^{n} P(i,j)\, |x(j)|$$

Sum over i $\quad (*) = \sum_{j=1}^{n} |x(j)| = \|x\|_1$ $\blacksquare$

Proof: (of lemma 3) $\quad x = (1-\alpha)v + \alpha P x$

Write as $\quad \boxed{Ax = v}$ where

$$A = (1-\alpha)^{-1}(I - \alpha P) \qquad \alpha < 1$$

Since $\quad Ax = (1-\alpha)^{-1}(x - \alpha P x)$

Claim: $\text{Ker}(A) = \{0\}$ for $0 \le \alpha < 1$

To see this, let $z \in \text{Ker}(A)$, $Az = 0$

or $\cancel{(1-\alpha)^{-1}}(z - \alpha P z) = 0$

$$z = \alpha P z$$

Thus,

$$\|z\|_1 = \|\alpha P z\|_1$$

$$= \alpha \|Pz\|_1$$

Prop 4 $\leq \alpha \|z\|_1$

$$\Rightarrow \quad (1-\alpha)\|z\|_1 \le 0$$

$$\Rightarrow \quad (1-\alpha)\|z\|_1 = 0$$

So either $z=0$ or $\alpha=1$

and $\alpha<1$ so $z=0$ ✓

Hence for every $v \in \mathbb{R}^n$ there exists a unique $x \in \mathbb{R}^n$ solving

$$x = (1-\alpha)v + \alpha P x.$$

To prove (i): $\sum_{i=1}^{n} x(i) = \sum_{i=1}^{n} v(i)$

$$\sum_{i=1}^{n} x(i) = \sum_{i=1}^{n} \left((1-\alpha) v(i) + \alpha \sum_{j=1}^{n} P(i,j) x(j) \right)$$

$$= (1-\alpha) \sum_{i=1}^{n} v(i) + \alpha \sum_{j=1}^{n} x(j) \underbrace{\sum_{i=1}^{n} P(i,j)}_{=1}$$

$$= (1-\alpha) \sum_{i=1}^{n} v(i) + \alpha \sum_{i=1}^{n} x(i)$$

$$\Rightarrow \quad (1-\alpha) \sum_{i=1}^{n} x(i) = (1-\alpha) \sum_{i=1}^{n} v(i) \quad \checkmark$$

Since $\alpha < 1$

To prove (ii): If $v(i) \geq 0$ for all i then $x(i) \geq 0$ for all i

Assume $v(i) \geq 0$ for all i

$$|x(i)| = \left| (1-\alpha) v(i) + \alpha \sum_{j=1}^{n} P(i,j) x(j) \right|$$

$$\leq (1-\alpha) v(i) + \alpha \sum_{j=1}^{n} P(i,j) |x(j)|$$

$\sum_{i=1}^{n} P(i,j) = 1$

$$\sum_{i=1}^{n} |x(i)| \leq (1-\alpha) \sum_{i=1}^{n} v(i) + \alpha \sum_{j=1}^{n} |x(j)|$$

$$(1-\alpha) \sum_{i=1}^{n} |x(i)| \leq (1-\alpha) \sum_{i=1}^{n} v(i) = (1-\alpha) \sum_{i=1}^{n} x(i)$$

↑ part (i)

$$\Longrightarrow x(i) \geq 0 \quad \blacksquare$$

Eigenvector problem

Remark 5. When v is a probability distribution, it is common to re-write the PageRank problem (2) as an eigenvector problem

$$x(i) \geq 0$$

$$P_\alpha x = x$$

$$\mathbf{1}^T x = 1 = \sum_{i=1}^{n} x(i)$$

where

$$P_\alpha := (1-\alpha)v\mathbf{1}^T + \alpha P.$$

Note

$$x = (1-\alpha)v + \alpha P x$$

$$= (1-\alpha)v\mathbf{1}^T x + \alpha P x$$

$$= \underbrace{\left((1-\alpha)v\mathbf{1}^T + \alpha P\right)}_{P_\alpha} x$$

$$\Longrightarrow P_\alpha x = x$$

Convergence of the PageRank iteration

Let $v \in \mathbb{R}^n$ and $0 \leq \alpha < 1$. Let x_k satisfy the PageRank iteration

$$x_{k+1} = (1 - \alpha)v + \alpha P x_k,$$

and let x be the unique solution of the PageRank problem

$$x = (1 - \alpha)v + \alpha P x.$$

Theorem 6. *We have*

(3) $$\|x_k - x\|_1 \leq \alpha^k \|x_0 - x\|_1.$$

Since $0 \leq \alpha < 1$, this is convergence of $x_k \to x$ with a linear convergence rate of α.

Proof:

$$x = (1-\alpha)v + \alpha P x$$

$$x_k = (1-\alpha)v + \alpha P x_{k-1}$$

$$\|x_k - x\|_1 = \|\alpha(Px_{k-1} - Px)\|_1$$

$$= \alpha \|P(x_{k-1} - x)\|_1$$

Prop 4

$$\leq \alpha \|x_{k-1} - x\|_1$$

By induction ...

$$\leq \alpha^k \|x_0 - x\|_1 \quad \blacksquare$$

Since $\alpha < 1$, $\alpha^k \to 0$ as $k \to \infty$

Power iteration

Remark 7. In the eigenvector formulation discussed above, the PageRank iteration $x_{k+1} = P_\alpha x_k$ is basically the power iteration to find the largest eigenvector of P_α The normalization step is not needed since $\|x_k\|_1 = 1$ for all k.

Personalized PageRank for image retrieval (.ipynb)