Case Studies A

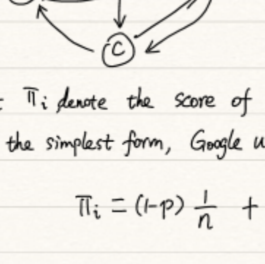
2020年3月25日 上午 12:18

Case Studies A

PageRank

Problem: Rank the webpages?

Data: The linkages of webpages, e.g.



A, B, …, F are webpages.

"Ⓔ→Ⓕ" means webpage Ⓔ has a hyperlink to Ⓕ, and so on.

Model: Let π_i denote the score of node i.

In the simplest form, Google uses the following model

$$\pi_i = (1-p)\frac{1}{n} + p\sum_{j\in N(i)} \frac{\pi_j}{L_j}, \quad \cdots\cdots (PR)$$

$i=1,2,\ldots,n$

where n — total number of nodes

$N(i)$ — number of nodes pointing to node i.

L_j — the number of outbound links of node j.

p — a parameter satisfying $0<p<1$

Actually, π_i is the probability representing the likelihood that a preson randomly clicking on links will arrive at webpage i. There are two terms in π_i:

$(1-p)\frac{1}{n}$ — the person will, with probability $1-p$, open an arbitrary new webpage with equal probability. so he/she opens webpage i with prob. $(1-p)\frac{1}{n}$.

$p\sum_{j\in N(i)}\frac{\pi_j}{L_j}$ — the preson will, with probability p, click links an arbitrary link on the current webpage with equal probability. So, the probability that the person arrives webpage i is

π_i is always smaller than 1.

$$p\sum_{j\in N(i)} \frac{\pi_j}{L_j}$$

π_j — the probability the current webpage is j.

L_j — there are L_j outbounds of j.

So the conditional probability from $j\to i$ is $\frac{\pi_j}{L_j}$

Example: Consider the example above. The equation (PR) is

$$\begin{bmatrix}\pi_A\\ \pi_B\\ \pi_C\\ \pi_D\\ \pi_E\\ \pi_F\end{bmatrix} = (1-p)\begin{bmatrix}1/6\\1/6\\1/6\\1/6\\1/6\\1/6\end{bmatrix} + p \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0 & 0 & 1/2 & 0 & 1/3 & 1\\ 1/2 & 0 & 1/2 & 0 & 1/3 & 0\\ 0 & 1/2 & 0 & 1 & 0 & 0\\ 1/2 & 1/2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1/3 & 0\end{bmatrix}\end{array} \begin{bmatrix}\pi_A\\ \pi_B\\ \pi_C\\ \pi_D\\ \pi_E\\ \pi_F\end{bmatrix}$$

e.g., C has two outbound links (to A and B)

normalized adjancency matrix (s.t. column sum is 1)

This is a linear equation on vector $\pi\in\mathbb{R}^6$. We can use, e.g., LU decomposition, to solve it. If $p=0.85$, then the solution is

$$\pi = \begin{bmatrix}0.189\\0.242\\0.305\\0.208\\0.025\\0.032\end{bmatrix}.$$

85%在現有webpage開一條新link
15%關掉所有頁面開一個新Page

So the ranking is

Rank	node	Score	in	out
1	C	0.305	2	2
2	B	0.242	3	2
3	D	0.208	2	1
4	A	0.189	3	2
5	F	0.032	1	1
6	E	0.025	0	3

how many Inbounds link & outbound link

- Why C ranks higher than B?
- Why B ranks higher than A?

D與B都是有多的inbound link,但他們同時link出去C,證明C也很重要

In general, Equation (PR) can be written in matrix form as

$$\pi = \frac{1-p}{n}\cdot\mathbb{1} + pA\pi,$$

$$A: a_{ij} = \begin{cases} 0 \ if\ j\notin N(t) \\ \frac{1}{L_j} if\ j\in N(i)\end{cases}$$

where $\pi\in\mathbb{R}^n$ is the score vector, $\mathbb{1}=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\in\mathbb{R}^n$, and A is the normalized adjancency matrix with each column sum 1. We obtain

$$(I-pA)\pi = \frac{1-p}{n}\mathbb{1}, \quad \cdots\cdots$$

which can be solved by LU decomposition

Question: Prove that the solution $\mathbb{1}^T\pi=1$ and $\pi_i\ge 0\ \forall i$. (i.e., prove that the solution π is indeed a probability distribution).

Open questions: ① How to improve the model (PR)? pagerank

② Any other possible applications of (PR)?

Things to Note:
1. $I - pA$ is non − singular if $p \neq 1$ (Proof = out c.)
2. $0 \le \pi \le 1$ and $\pi^T\mathbb{1} = 1$

Image Decomvolution

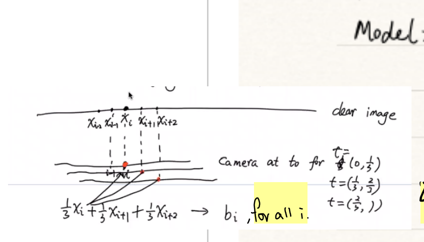
拍照片防手抖

Problem: Recover the clear image from a blurred one.

Model: For simlicity, we consider 1D example, only one line i.e. the unknown clear image is a vector $x\in\mathbb{R}^n$, where x_i, $i=1,\cdots,n$ is gray level at pixel i.

We also assume the blurring is a moving average, i.e. the observed blurred image $b\in\mathbb{R}^n$ satisfies

$$b_i = (x_{i-1}+x_i+x_{i+1})/3 \qquad i=1,2,\cdots,n.$$



上述這張圖是比喻在t=(0,1)的時候手抖了一下收到的Image跟之前1/3不同，所以b_i(clear image)的形成就有三個term

general啲嘅形式係 $\sum_{j=-S}^{S} K_j x_{i-j} = b_i, i=1,\ldots,n$

而有啲相機可以估計K_j出黎

所以成個問題就變成solve Ax=b

where A=

$$\begin{bmatrix} \ddots & & & & & & \\ & K_s \cdots & K_1 & K_0 & K_{-1} & \cdots K_{-s} & \\ & & & & & & \ddots \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

For $i=1$, we have $b_1 = (x_0 + x_1 + x_2)/3$

For $i=n$, we have $b_n = (x_{n-1} + x_n + x_{n+1})/3$

where x_0, x_{n+1} are not defined yet. we assume $x\in\mathbb{R}^n$ is periodic, i.e., $x_1\cdots x_n | x_1\ x_2 \cdots x_n | x_1\ x_2 \cdots$

$x_0 = x_n$ and $x_{n+1} = x_1$ — Called boundary condition.

Therefore, the model is

Proof of above things to note:

$0\le\pi\le1$ use Perron-Frobenius theorem.

$\underbrace{\pi^T\mathbb{1}}_{\sum_{i=1}^{n}\pi_i} = 1$

Proof:

$$(I-pA)\pi = \frac{1-p}{n}\mathbb{1}$$

$$\mathbb{1}^T(I-pA)\pi = \frac{1-p}{n}\mathbb{1}^T\mathbb{1} \quad (=n)$$

$$\mathbb{1}^T\pi - p\mathbb{1}^TA\pi = 1-p$$

$$\mathbb{1}^T\pi = (1-p) + p\mathbb{1}^TA\pi$$

$$\mathbb{1}^T\pi = 1-p+p\mathbb{1}^T\pi \quad (\because \mathbb{1}^TA = \mathbb{1}^T \because \text{Column sum of } A \Rightarrow 1)$$

if $p\neq1$, then $\mathbb{1}^T\pi = 1$

To improve the pagerank model, maybe we could improve by getting data to guess the probability of entering each page.

$$\begin{cases} b_1 = (x_1+x_2+x_n)/3 \\ b_i = (x_{i-1}+x_i+x_{i+1})/3, \quad i=2,\cdots,n-1 \\ b_n = (x_1+x_{n-1}+x_n)/3 \end{cases}$$

In matrix form,

$$Ax=b,$$

where

$$A = \frac{1}{3}\begin{bmatrix} 1 & 1 & & & 1 \\ 1 & 1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 1 & & & 1 & 1 \end{bmatrix}$$

The de-blur is to solve x from $Ax=b$.

In general, many blurring can be modeled as $Ax=b$.

We can solve Ax=b by LU decomposition
However, deblurring by solving Ax=b is unstable:
a small noise in b will lead to a large error in the sol.
因為兩個原因:1.因為x_i係可以alternative的，
令到b的value可以始終等於某個value
令到Ax=b有infinitely many sol.
2.因為影相過程可能有noise
如果係很短的曝光時間底下
noise will be amplified 造成很大誤差

數學的表達方式是:

$A\tilde{x} = \tilde{b}$ 雖然$\|b-\tilde{b}\|_2$很小，

$Ax = b$ 但$\|x-\tilde{x}\|_2$可以很大.

Solution: However, A is singular (i.e., not invertible), and there are infinitely many solutions of $Ax=b$.

To have a unique solution, we use Tikhonov regularization, where we solve

加個懲罰項，令x的range不至於太大或者太細

$$\min_x \|Ax-b\|_2^2 + \alpha\|x\|_2^2, \quad \cdots\cdots (TK)$$

Here $\alpha>0$ is called regularization parameter.

In other words, we find an $x\in\mathbb{R}^n$ s.t.

① $\|Ax-b\|_2^2$ is small.

② $\|x\|_2^2$ is small.

α balances the relative importance of ① and ②.

Larger $\alpha \Longrightarrow \|Ax-b\|_2^2$ is larger, and $\|x\|_2^2$ is smaller.

smaller $\alpha \Longrightarrow \|Ax-b\|_2^2$ is smaller, and $\|x\|_2^2$ larger.

To solve the minimization (TK), we take gradient of the objective function and set it to 0. We obtain (TK) is equivalent to solve

$$2A^T(Ax-b) + 2\alpha x = 0$$

$$\Updownarrow$$

$$(A^TA+\alpha I)x = A^Tb.$$

Fact: $A^TA+\alpha I$ is symmetric positive definite.

proof. i) $(A^TA+\alpha I)^T = (A^TA)^T + \alpha I^T = A^TA+\alpha I$

ii) $\forall y\neq0$, $y^T(A^TA+\alpha I)y = y^TA^TAy + \alpha y^Ty$

$= \|Ay\|_2^2 + \alpha\|y\|_2^2 > 0$ (Since $\|y\|_2>0$, $\alpha>0$) ∎

Therefore, we can use Cholesky decomposition to solve

$$(A^TA+\alpha I)x = A^Tb.$$