Google Page Rank Lecture 35

In the early days of the internet, search engines weren't very good. They mostly functioned by considering only the number of word matches in a query. So if you wanted a recipe for crab tacos and searched "crab taco", you might get very bad web pages, such as one which said only "crab taco crab taco...." 400,000 times (hey, anyone can put anything on the internet).

The idea of Larry Page and Sergey Brin, 2 Stanford grad students, was to find a metric which valued popular pages over weird webpages in untravelled corners of the internet. We start from the beginning to develop the math necessary to do this.

Probability

Defn A probability on a finite set S is a function $p: S \to \mathbb{R}$ such that

(i) $0 \leq p(s) \leq 1$ for all $s \in S$ (the probability of anything happening is between 0 and 1)

(ii) $\sum_{s \in S} p(s) = 1$ (the probability of something happening is 1)

ex1 Suppose we are drawing cards from a deck, and consider 2 outcomes: we draw a numbered card (2-10) or a royal card (A, King, Queen, Jack). Assuming a very well shuffled deck, we model this as follows:

$S = \{\text{numbered}, \text{royal}\}$, $p(\text{numbered}) = \dfrac{\#\text{ numbered cards}}{\#\text{ total cards}} = \dfrac{36}{52} = \dfrac{9}{13}$

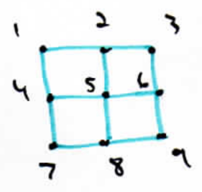
$p(\text{royal}) = \dfrac{\#\text{ royal cards}}{\#\text{ total cards}} = \dfrac{16}{52} = \dfrac{4}{13}$.

Defn A **Markov process** is a sequence of probabilistic events $S_0, S_1, S_2, \ldots$ with probabilities $P_0, P_1, P_2, \ldots$ such that P_{k+1} depends only on P_k, and this does not depend on k.

To make precise the meaning of "P_{k+1} depends only on P_k" is beyond the scope of this lecture. We provide an illustrative example.

ex2 (Drunken physicist on a 3×3 grid)

Suppose we have a 3×3 grid, and an agent who starts at the center; we number the positions 1–9, so $S_0 = 5$. At every time step, she takes a step in a direction on the grid in a random direction. So, $p_0(5) = 1$ and $p_0(i) = 0$ for $i \neq 5$ (we assign him to 5 with probability 1). Then she takes a step ~~two~~ 2, 4, 6, 8 with equal probability; $p_1(2) = p_1(4) = p_1(6) = p_1(8)$. Suppose we observe ~~her~~ take a step to 2. Then her next step is to 1, 3 or 5, each with probability $1/3$. They key point is that ~~your~~ her next step only depends on her previous step; nothing before it.



Matrix Notation

If we label the possible states $1 \ldots n$, then we can write S ~~as a probability~~ P as a vector. So, for example, in ex1, we ~~can~~ let numbered be outcome 1 and royal be outcome 2, and $P = \begin{pmatrix} 9/13 \\ 4/13 \end{pmatrix}$. In ex2, the state ~~at~~ at time step k is given by one of the nine positions. So

$$P_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0 \\ \frac14 \\ 0 \\ \frac14 \\ 0 \\ \frac14 \\ 0 \\ \frac14 \\ 0 \end{pmatrix}.$$

We will see how to compute the further vectors $P_2, P_3, \ldots$ as soon.

Defn The <u>transition matrix</u> of a Markov process $P_0, P_1, \ldots$ is the matrix $T = (t_{ij})$ where t_{ij} is the probability that $s_{k+1} = i$ given that $s_k = j$ (written $P_{k+1}(i \mid s_k = j)$ and read "the probability that the $k+1$th position is i given that the kth position is j").

ex 2 continued So, in example 2, our transition matrix is

$$T = \begin{pmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{pmatrix}$$

e.g. t_{58} = probability of taking a step to 5 given that we're currently on 8

The transition matrix thus gives the probability of transitioning from one state to another. In fact,

$$P_1 = T P_0, \quad P_2 = T^2 P_0, \quad P_3 = T^3 P_0, \ldots, \quad P_k = T^k P_0.$$

Page Rank

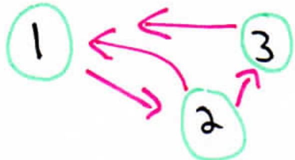
We define the page rank of a webpage through the probability that a randomly walking internet surfer "eventually" lands on a given webpage.

Defn A <u>directed graph</u> (or <u>digraph</u>) is a set of vertices $V = \{1, 2, \ldots, n\}$ and edges $E = \{(a_1, b_1), \ldots, (a_m, b_m)\}$.

We define a directed graph of the internet by labeling the websites $1 \ldots n$, and adding and edge from a_i to b_i if a has a link to b.

ex Consider an internet with 3 websites 1, 2, & 3.

We draw the internet graph as

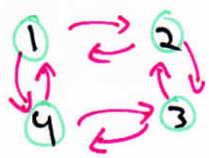


This means there is a link from 1 to 2, 2 to 1, 2 to 3, and 3 to 1.

Idea) We take a step along an edge at random, and let

$P_\infty = \lim_{k \to \infty} T^k P_0$ (independent of P_0). We then use $(P_\infty)_i$ to define the page rank of website i.

Problem) This limit might not exist. For example, consider the following graph. Set $P_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, so we start at 1.



$T = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$. We compute $P_1 = TP_0 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}$, $P_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$,

$P_3 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}$, $P_4 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$, ...

and this never settles down (~~cannot~~ it never converges). The problem is at odd time steps we will always be at an even position; at an even time step, we will always be at an odd position.

To fix this, we apply laziness, or "damping." We choose a damping factor between 0 and 1, and add a probability of $1-d$ that we ~~[illegible]~~ stay put. So, for example, if we choose a damping factor of $2/3$ with our previous internet,

$$T = \begin{pmatrix} \frac13 & \frac13 & 0 & \frac13 \\ \frac13 & \frac13 & \frac13 & 0 \\ 0 & \frac13 & \frac13 & \frac13 \\ \frac13 & 0 & \frac13 & \frac13 \end{pmatrix} = \frac13 \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Then, using the same P_0,

$$TP_0 = \begin{pmatrix} \frac13 \\ \frac13 \\ 0 \\ \frac13 \end{pmatrix}, \quad T^2 P_0 = \begin{pmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{pmatrix}, \quad T^3 P_0 = \begin{pmatrix} 7/27 \\ 7/27 \\ 6/27 \\ 7/27 \end{pmatrix}$$

We ~~can~~ can see that $T^k P_0$ is converging rapidly to $\begin{pmatrix} \frac14 \\ \frac14 \\ \frac14 \\ \frac14 \end{pmatrix}$.

So, we define page rank by choosing a damping factor d and set ~~[illegible]~~. $P_0 = \frac{1}{N}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, and define the page rank of a website as its entry in $P_\infty = \lim_{k \to \infty} T^k P_0$.

ex 3 (Continued) So for the previous internet with a damping factor of $\frac34$, we get the transition matrix

$$T = \begin{pmatrix} \frac14 & \frac38 & \frac34 \\ \frac34 & \frac14 & 0 \\ 0 & \frac38 & \frac14 \end{pmatrix}.$$

We compute $T^{16} P_0 = T^{16} \cdot \frac13 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

$$T^{16} \cdot P_0 = \begin{pmatrix} .3999995 \\ .4000005 \\ .2 \end{pmatrix}.$$