In the name of GOD.

Sharif University of Technology

Stochastic Process

Fall 2023

Hamid R. Rabiee

Homework 5	Hypothesis Testing, Markov Chains, Hidden Markov Model	Deadline : 1402/10/05

1. [15] (One real-world application) The World Wide Web is a large graph with N web pages as vertices (v_1, ..., v_N). There is a directed edge from a vertex v_i to another vertex v_j, if v_i's corresponding page has a hyperlink to that of v_j's. Google has developed an algorithm called Page Rank, to rank the web pages based on validity/importance. Please study about the Page Rank and explain how is it related to Markov Chains. How is it calculated? Why is the theory behind Page Rank calculation always applicable to it?

   Solution:
   Google has defined a kind of random walk through the web pages. Being at page i, if i does not have any hyperlink to another page, the walker jumps randomly to any of the N pages. Otherwise, it follows one of the existing hyperlinks with a uniform distribution. Thus, the walking is a Markov Chain, with pages being states. Assuming the edge matrix is denoted by E, transition probabilities would be:

$$P_{ij} = \begin{cases} \frac{1}{\sum_k E_{ik}} & \sum_k E_{ik} > 0 \\ \frac{1}{N} & \sum_k E_{ik} = 0 \end{cases} \tag{1}$$

   But, to make sure the transition matrix becomes aperiodic, Google has also considered a very small probability of r for random restart. Thus at each page, the walker randomly jumps to any other page with a probability of $0 < r << 1$, and follows the above mentioned policy with a probability of $1 - r$. Therefore, the real transition matrix is calculated as:

$$T_{ij} = r \times \frac{1}{N} + (1 - r) \times P_{ij} \tag{2}$$

   Page Rank is equal to the stationary distribution of the mentioned Markov Chain. It always exists and it is unique because the states graph is aperiodic. Intuitively, the more a page has been referred to by the other pages, the higher the probability of presence in it in a long-time walk, and the higher the validity and popularity of the page.

2. [15] What is the stationary probability of a random walker on an undirected unweighted connected graph with binary edge matrix of E (assuming it is not bi-partite)? Prove your answer.

   Solution:

   (a) If a connected graph is not bipartite, the graph would be aperiodic.

i. If the graph is not bipartite, we would have a cycle of odd length from any node to itself. Please check this link for proof.

ii. For each node of the graph, we can go from the node to one of its neighbors and return (a cycle of length 2), and we also have at least a cycle of odd length. So the period would be 1 and the graph is aperiodic.

(b) As the graph is aperiodic, it has a unique stationary distribution. So if we propose a stationary distribution and prove it satisfies the two conditions for the stationary distribution, then it is the solution!

i. $\sum_i \pi_i = 1$

ii. $T \times \pi = \pi \equiv \forall i : \pi_i = \sum_j \pi_j \times T_{ji}$

(c) Intuitively, we propose a solution of $\pi_i = \frac{deg_i}{2m}$, in which deg_i stands for the degree of the ith vertex and m stands for the number of edges in the graph which we know $\sum_i deg_i = 2m$.

i. $\sum_i = \sum_i \frac{deg_i}{2m} = \frac{2m}{2m} = 1$ ✓

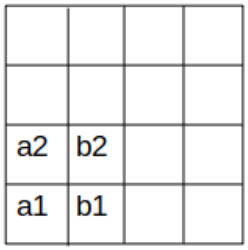
ii. $\sum_j \pi_j T_{ji} = \sum_j \frac{def_j}{2m} \frac{E_{ij}}{deg_j} = \frac{\sum_j E_{ij}}{2m} = \frac{deg_i}{2m} = \pi_i$ ✓

3. [15] Prove that the expected number of steps to return from a state to itself in a irreducible positive recurrent Markov Chain is equal to the inverse of the stationary probability of that state.

Solution:
Please refer to the last section of this link.

4. [15] Calculate the expected number of random moves a knight will take to return to the corner of a1 starting from the same place in a chess-board of size 4.

Solution:
Note that this is a connected, unweighted, undirected graph and there is a cycle of odd length from each state to itself. According to Question2, we only need to calculate the degree of each node for calculating stationary distribution. Due to asymmetry, we only need to calculate the possible number of moves for $a1$, $a2$, $b1$, and $b2$. For a L-shaped move, we would have:



- $a1$: 4
- $a2 = b1 = 5$ (again due to asymmetry)
- $b2 = 6$

The the stationary probability for $a1$ would be $\frac{4}{4+5+5+6} = \frac{1}{5}$. Then the expected number of steps would be 5 according to Question3.

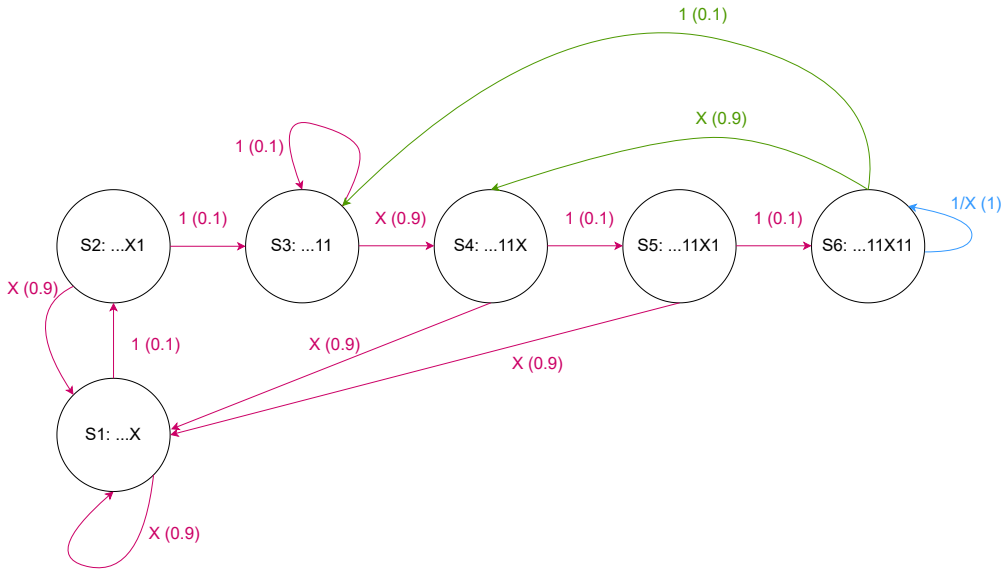
5. [25] A lucky box receives a dollar from the player and returns a prize of 1000$ with a probability! Everyone thinks the machine tosses a coin to decide whether to give the prize, but it actually has a secret mechanism inside! It has a short memory of 5 digitis to keep the last 5 randomly selected digits and if the sequence turns out as 11[*anything_but_*1]11, it passes the prize. Everytime a new digit is randomly selected and appended to the previously generated sequence. Knowing the digits are selected based on a uniform distribution, and considering the initial state of the memory to be 00000, calculate:

   (a) The probability of at least one win in the first 10, 100, and 10000 moves respectively?

   (b) The average profit of the owner in a long time run?

   Please use a computer for your matrix operations.

   Solution:
   There are two possible transitions for each state of the graph. Appearance of 1, or any other digit that we call X. The state graphs for the two parts of this question are a little bit different. The general scheme is as follows. The pink links are shared between the parts (a) and (b), while the blue links complete the part (a) and the green ones part (b).



   (a) p(1) = 0.1, p(X) = 0.9
   The transition matrix becomes:

$$T = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 0.1 & 0 \\ 0.9 & 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

   We need: $T^N_{0,5}$, which is 5.37e-04 for 10, for 8.52e-03 100, and 0.9986 for 10000.

   (b) We have the following set of equations according to the properties of stationary probability:

$$\begin{cases} 0.9\pi_1 + 0.9\pi_2 + 0.9\pi_4 + 0.9\pi_5 = \pi_1 => \pi_1 = 9(\pi_2 + \pi_4 + \pi_5)(I1) \\ 0.1\pi_1 = \pi_2(I2) \\ \pi_3 = 0.1\pi_2 + 0.1\pi_3 + 0.1\pi_6(I3) \\ \pi_4 = 0.9\pi_3 + 0.9\pi_6(I4) \\ \pi_5 = 0.1\pi_4(I5) => I1, I2, I5 => \pi_1 = 990\pi_5(I6), \pi_1 = 99\pi_4(I7) \\ \pi_6 = 0.1\pi_5(I8) => I8, I6 => 9900\pi_6 = \pi_1(I9) \\ I2, I3, I9 => 9 \times 9900\pi_3 = 991\pi_1 \\ \sum_i \pi_i = 1 => \pi_1 + 0.1\pi_1 + \frac{991}{9 * 9900}\pi_1 + \frac{1}{99}\pi_1 + \frac{1}{990}\pi_1 + \frac{1}{9900}\pi_1 = 1 => \pi_1 = 0.891 \\ => \pi_2 = 0.0891, \pi_3 = 0.00991, \pi_4 = 0.009, \pi_5 = 0.0009, \pi_6 = 0.00009 \end{cases} \tag{4}$$

The rate of benefit would be : $1 - \pi_3 * 1000\$ = 0.91\$pergame$

6. [15] A medical company performs researches to assess the effectiveness of a newly discovered medication over a special disease. 50 diseased people are randomly selected and divided to two groups of 25. One group, names the case group, receive the medication while the other group, named as control group, receive something irrelevant. After a month, the severity of the case group has been observed to be 4.2 with a standard deviation of 1.5 and the other group 5.8 with a standard deviation of 1.8. Can the company conclude that the medication has been effective with a significance level of 0.01? Assess the results using the p-value approach with a significance level of 0.5.

   Solution:

Step 1: The hypothesis statement is H_0: $\mu = \$1{,}240$ versus H_1: $\mu \neq \$1{,}240$.

Observe that μ represents the true-but-unknown mean for November. The comparison value \$1,240 is the known traditional value to which you want to compare μ.

Do not be tempted into using H_1: $\mu < \$1{,}240$. The value in the data should not prejudicially influence your choice of H_1. Also, you should not attempt to second-guess the researcher's motives; that is, you shouldn't try to create a story that suggests that the researcher was looking for smaller costs. In general, you'd prefer to stay away from one-sided alternative hypotheses.

Step 2: Level of significance $\alpha = 0.05$.

The story gives no suggestion as to the value of α. The choice 0.05 is the standard default.

Step 3: The test statistic will be $t = \sqrt{n}\,\dfrac{\bar{x} - \mu_0}{s}$. The null hypothesis will be rejected if $|t| \geq t_{\alpha/2;n-1}$. If $|t| < t_{\alpha/2;n-1}$ then H_0 will be accepted or judgment will be reserved.

At this point it would be helpful to recognize that the sample size is small; we should state the assumption that the data are sampled from a normal population.

In using this formula, we'll have $n = 15$, $\mu_0 = \$1{,}240$ (the comparison value), and $\bar{x} = \$1{,}080$ and $s = \$180$ will come from the sample. The value $t_{\alpha/2;n-1}$ is $t_{0.025;14} = 2.145$.

The "judgment will be reserved" phrase allows for the possibility that you might end up accepting H_0 without really believing H_0. This happens frequently when the sample size is small.

Step 4: Compute $t = \sqrt{15}\,\dfrac{\$1{,}080 - \$1{,}240}{\$180} \approx -3.443$.

Step 5: Since | -3.443 | = 3.443 > 2.145, the null hypothesis is rejected. The November cases are significantly different.

> Plugging in the numbers and reaching the "reject" decision are routine. Observe that we declare a *significant* difference. The word *significant* has jargon status; specifically, it means that a null hypothesis has been rejected.

This discussion did not request a *p*-value. However, we can use the value 3.443 in the *t* table to make a statement. Using the line for 14 degrees of freedom, we find that

$$t_{0.005;14} = 2.977 < 3.443 < 3.787 = t_{0.001;14}$$

we see that H_0 would have been rejected with $\alpha = 0.01$ (for which $\alpha/2 = 0.005$) and would have been accepted with $\alpha = 0.002$ (for which $\alpha/2 = 0.001$). Thus we can make the statement $0.002 < p < 0.01$. Some users might simply write $p < 0.01$ ** , using the ** to denote significance at the 0.01 level.

> You can use Minitab to get more precise *p*-values. Use **Calc** ⇒ **Probability Distributions** ⇒ **t** and then fill in the details
>
> ⊙ **Cumulative probability**
> **Degrees of freedom:** 14
> **Input constant:** 3.443
>
> Minitab will respond with this:
>
> ```
>        x     P( X <= x )
> 3.4430          0.9980
> ```
>
> The excluded probability to the right is 1 - 0.9980 = 0.0020. The same probability appears below -3.443, so the *p*-value should be given as 0.0040.

Some people simply prefer confidence intervals as a method of summarizing. Here the 95% interval for μ is $\bar{x} \pm t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}$, which is $\$1{,}080 \pm 2.145\frac{\$180}{\sqrt{15}}$. Numerically this is \$1,080 ± \$100 or (\$980 to \$1,180). It might be noted that the comparison value \$1,240 is outside this interval, consistent with the fact that H_0 was rejected at the 5% level.