Group Assignment 3

Kirti Barman, Sourav Sharan, Vyoma Shah, Aditya Khadkikar
Group 15 - Introduction to Data Science

January 16, 2025

Exercise 1. Consider a three state (1, 2, 3) Markov Chain with transition matrix

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

(a) Draw transition diagram.
(b) Find the stationary distribution π.
(c) Given that the chain is in state 1 at time 1, what is the probability that the chain is in state 2 at time 4?
(d) Given that the chain is in state 1 at time 1, what is the expected time until the chain is in state 3 the first time?
(e) What is the period of each state?

Solution. Solving for Markov chain transition matrix P,

(a) Transition Matrix:

The transition diagram connects states based on the probabilities given in P:

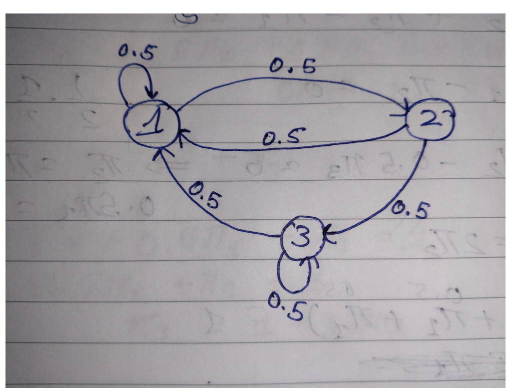

Figure 1: Transition Diagram

- State 1 transitions to itself (0.5) and to state 2 (0.5).
- State 2 transitions to state 1 (0.5) and state 3 (0.5).
- State 3 transitions to state 1 (0.5) and itself (0.5).

(b) Stationary Distribution:

The stationary distribution $\pi = (\pi_1, \pi_2, \pi_3)$ satisfies

$$\pi P = \pi$$

Solving $\pi P = \pi$ for the matrix given above, we get the following set of equations:

$$\pi_1 = 0.5\pi_1 + 0.5\pi_2 + 0.5\pi_3$$

$$\pi_2 = 0.5\pi_1$$

$$\pi_3 = 0.5\pi_2 + 0.5\pi_3$$

$$\pi_1 + \pi_2 + \pi_3 = 1(Normalization).$$

Solving gives,

$$\pi = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$$

(c) State Transition Probability:

Using the transition matrix raised to the power of 4, we get:

$$P^4 = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}$$

The probability of being in state 2 at time 4 starting from state 1 is found from looking at the i^{th} row (where i = the start destination = 1), and the j^{th} column (where j = the end destination = 2). This is found to be 0.25.

(d) Expected Time to State 3:

To find the expected time to reach state 3 given that we started at state 1, we set up a system of equations.

This aims at finding the expected number of steps from a state i to the destination, i being the state space which we have. Those are the unknown variables in the system of equations. It is set up like below:

$$m_{1\to 3} = 1 + p_{1\to 2}m_{2\to 3} + p_{1\to 1}m_{1\to 3}$$
$$m_{2\to 3} = 1 + p_{2\to 1}m_{1\to 3}$$
$$m_{3\to 3} = 0$$

From the above equations, $m_{3\to 3}$ is 0, as we are already at our desired destination node. We substitute in the values for the probabilities, as we know them from the transition matrix.

$$m_{1\to 3} = 1 + 0.5m_{2\to 3} + 0.5m_{1\to 3}$$
$$m_{2\to 3} = 1 + 0.5m_{1\to 3}$$

Moving the unknown terms to one side, we get:

$$0.5m_{1\to 3} - 0.5m_{2\to 3} = 1$$
$$-0.5m_{1\to 3} + m_{2\to 3} = 1$$

Solving the equations, we get $m_{1\to 3} = 6$, and $m_{2\to 3} = 4$.

Therefore, the expected time to reach state 3 from state 1 is 6.

(e) Period of Each State:

To calculate the period of each state, we can see the possible routes that can be taken to start from the given state, and return back.

For state 1, some of the possible routes are 1->1 (1 step), 1->2->1 (2 steps), 1->2->3->1 (3 steps). The greatest common divisor of these step counts is 1, so the period of state 1 is 1.

For state 2, possible routes are 2->1->2 (2 steps), 2->3->1->2 (3 steps). The greatest common divisor of 2 and 3 is 1, so the period of state 2 is 1.

Lastly, for state 3, possible routes are 3->3 (1 step), 3->1->2->3 (3 steps). This is sufficient, as the greatest common divisor of 1 and 3 is 1, so the period of state 3 is also 1.

Hence, all states have a period of 1.

□

Exercise 2. Assume that we are trying to classify a binary outcome (Y), i.e., our data is of the form ($(X, Y) \sim F_{X,Y}$), where ($Y \in 0, 1$) and ($X \in \mathbb{R}^d$). We have used data to train a classifier ($g(X)$). We can evaluate the performance of the classifier using i.i.d. testing data, ($(X_1, Y_1), \ldots, (X_n, Y_n)$). We are interested in estimating the following quantities:

Precision: ($P(Y = 1|g(X) = 1)$)
Recall: ($P(g(X) = 1|Y = 1)$)

(a) Write down the empirical version of the precision and recall.
(b) Let us now think that the variable Y denotes if a battery's health has deteriorated or not, and let X denote a bunch of constructed health indicators about the battery. If the model g(X) predicts that the battery has deteriorated you need to run a test to confirm this. The cost of running the test is c when the battery is not deteriorated. On the other hand, if the battery is in fact deteriorated and the test is not run, the battery will die during use and the cost of this is d. Define a random variable representing the cost of the decision g(X) and write down the formula for the expected cost in terms of the precision and recall.
(c) Advanced question: can you produce a confidence interval for the expected cost? What about the precision and the recall?

Solution. **Precision and Recall:**

$$\text{Precision} = \frac{TP}{TP + FP}.$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Cost Function:

A cost function can be defined as:

$$C(P, R) = \alpha(1 - P) + (1 - \alpha)(1 - R).$$

The expected cost is:

$$E[C] = \alpha \frac{FP}{TP + FP} + (1 - \alpha)\frac{FN}{TP + FN}.$$

Confidence Intervals:

Confidence intervals for precision use the binomial proportion interval:

$$P \in \left[\hat{P} - z \cdot \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}, \hat{P} + z \cdot \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}\right].$$

Similar calculations apply for recall. □

Exercise 3. Show that two d-dimensional zero-mean, unit-variance Gaussian random vectors X and Y are nearly orthogonal by calculating their dot product $X^\top Y$ and bounding the probability that it exceeds ϵ.

Solution. **Dot Product Distribution:** The dot product $X^\top Y \sim \mathcal{N}(0, d)$.

Bounding Probability:

Using Gaussian tail bounds:

$$P(|X^\top Y| > \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2d}\right).$$

This shows that X and Y are nearly orthogonal for large d. □

Exercise 4. Prove that $u_i u_i^\top$ is rank 1, $U = \sum_{i=1}^{r} u_i u_i^\top$ is rank r, and perform SVD on U.

Solution. **Rank of $u_i u_i^\top$:** The outer product $u_i u_i^\top$ spans a 1-dimensional subspace, so it has rank 1.

Rank of $U = \sum_{i=1}^{r} u_i u_i^\top$: If $u_1, \dots, u_r$ are orthonormal, U spans an r-dimensional subspace, giving $\mathrm{rank}(U) = r$.

SVD of U: The singular value decomposition of U is $U = U \Sigma U^\top$, where U contains the singular vectors. □

Exercise 5. For $X \sim \text{Uniform}(B_1)$ (unit ball) and $Y = \|X\|_2$, find the distribution function of Y, analyze $\ln(1/Y)$, and calculate $E[\ln(1/Y)]$.

Solution. **Distribution Function of Y:** The cumulative distribution function of Y is:

$$F_Y(y) = y^d, \quad 0 \leq y \leq 1.$$

PDF of $\ln(1/Y)$: By denoting a random variable Z $= \ln(1/Y)$, upon simplifying for the RHS, we get that $Y = \frac{1}{e^Z}$. Using $Y = e^{-Z}$, the PDF is:

$$f_Z(z) = de^{-dz}.$$

Expected Value: The expected value is:

$$E[\ln(1/Y)] = \frac{1}{d}.$$

□