Group Assignment 3

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

January 20, 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 Diagram:

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

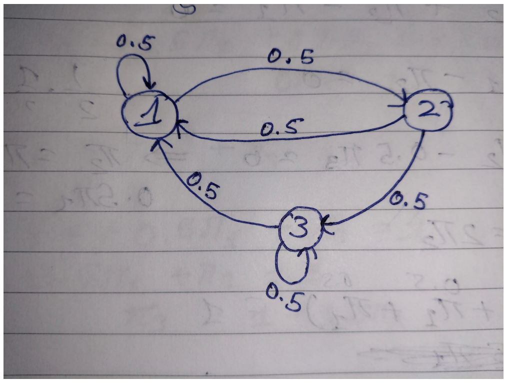

Figure 1: Transition Diagram

- 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

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

The probability of being in state 2 at time 4 starting from state 1 is 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:

$$m_{1\to3} = 1 + p_{1\to2}m_{2\to3} + p_{1\to1}m_{1\to3}$$
$$m_{2\to3} = 1 + p_{2\to1}m_{1\to3}$$
$$m_{3\to3} = 0$$

From the above equations, $m_{3\to3}$ 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.

$$m_{1\to3} = 1 + 0.5m_{2\to3} + 0.5m_{1\to3}$$
$$m_{2\to3} = 1 + 0.5m_{1\to3}$$

Moving the unknown terms to one side, we get:

$$0.5m_{1\to3} - 0.5m_{2\to3} = 1$$
$$-0.5m_{1\to3} + m_{2\to3} = 1$$

Solving the equations, we get $m_{1\to3} = 6$, and $m_{2\to3} = 4$.

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 the first state 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:**

$$\text{Precision} = \frac{N(Y = 1 \cap g(X) = 1, m)}{m},$$

where m represents the total subset of points from the predictions by the model g(x) in which a 1 was predicted (TP + FP). The above equation is based on the Long Term Relative Frequency, aiming to estimate the probability equation of precision empirically.

$$\text{Recall} = \frac{N(g(x) = 1 \cap Y = 1, n)}{n},$$

where n represents the number of points from the testing dataset that match the label 1 (TP + FN). Similar to precision, this is the empirical formula, and can be used to estimate the probability equation for recall.

Cost Function:

The cost of running the test is c when a battery is not deteriorated, however g(x) predicted the opposite. This refers to the model returning false positives (FP). Secondly, the cost d occurs when the battery is actually deteriorated, but predicted as functional (test not run), hence corresponding to the false negatives (FN).

A cost function can be defined as having a combination of the empirical precision $\hat{P}$ and recall $\hat{R}$ depending on what the placeholders for c and d are. If one has a higher cost than the other, it can be weighted more heavily, for it being a more significant part of the cost function. This is with a parameter $0 \leq \alpha \leq 1$. The expectation of the random variable C representing the cost function, is in terms of of the tradeoff combination as (1-Precision) and

(1-Recall) is so that the overall cost function can be made with respect to the false negatives and false positives. The precision and recall values both initially have the true positives in the numerator.

$$E[C] = \alpha(1-\hat{P}) + (1-\alpha)(1-\hat{R}).$$
$$E[C] = \alpha\frac{TP+FP-TP}{TP+FP} + (1-\alpha)\frac{TP+FN-TP}{TP+FN}.$$
$$= \alpha(\frac{N(g(X)=1 \cap Y=0, m)}{m}) + (1-\alpha)(\frac{N(g(X)=0 \cap Y=1, n)}{n}),$$

where 1 represents the label ”deteriorated”, and 0 represents that the battery is ”not deteriorated”.

Confidence Intervals:

With $\hat{P}$ being the empirically estimated precision, confidence intervals are found using the binomial proportion interval:

$$P \in \left[\hat{P} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{m}}, \hat{P} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{m}}\right].$$

where m is the same value as used in the empirical precision formula (the number of predictions that had label 1).

For recall, in a similar manner, the same confidence interval is described as the following:

$$R \in \left[\hat{R} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{R}(1-\hat{R})}{n}}, \hat{R} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{R}(1-\hat{R})}{n}}\right],$$

where n is the number of points in the testing dataset that are labelled 1. □

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.

Let X and Y be two independent random vectors, where each component X_i and Y_i (for $i = 1, 2, \ldots, d$) is drawn from a standard normal distribution $\mathcal{N}(0, 1)$. Thus, we have:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_d \end{pmatrix}$$

The dot product $X^\top Y$ is given by:

$$X^\top Y = \sum_{i=1}^{d} X_i Y_i$$

Distribution of the Dot Product

Since X_i and Y_i are independent standard normal random variables, the product $X_i Y_i$ follows a distribution known as the product of two independent normal variables. Specifically, the mean of $X_i Y_i$ is:

$$\mathbb{E}[X_i Y_i] = \mathbb{E}[X_i]\mathbb{E}[Y_i] = 0 \cdot 0 = 0$$

The variance of $X_i Y_i$ can be calculated as follows:

$$\mathrm{Var}(X_i Y_i) = \mathbb{E}[X_i^2]\mathbb{E}[Y_i^2] = 1 \cdot 1 = 1$$

Thus, the sum $\sum_{i=1}^{d} X_i Y_i$ is a sum of d independent random variables, each with mean 0 and variance 1. Therefore, by the Central Limit Theorem, as d becomes large, $X^\top Y$ will be approximately normally distributed:

$$X^\top Y \sim \mathcal{N}(0, d)$$

To analyze the probability that $|X^\top Y|$ exceeds ϵ, we standardize the variable:

$$Z = \frac{X^\top Y}{\sqrt{d}} \sim \mathcal{N}(0, 1)$$

Bound the probability:

$$P(|X^\top Y| > \epsilon) = P\left(\left|\frac{X^\top Y}{\sqrt{d}}\right| > \frac{\epsilon}{\sqrt{d}}\right)$$

Using the properties of the standard normal distribution, we can express this probability as:

$$P(|Z| > \frac{\epsilon}{\sqrt{d}}) = 2P(Z > \frac{\epsilon}{\sqrt{d}})$$

Using the tail bound for the standard normal distribution, we have:

$$P(Z > x) \leq \frac{1}{2}e^{-\frac{x^2}{2}} \quad \text{for } x > 0$$

Thus,

$$P(Z > \frac{\epsilon}{\sqrt{d}}) \leq \frac{1}{2}e^{-\frac{\epsilon^2}{2d}}$$

Final Bound

Combining these results, we find:

$$P(|X^\top Y| > \epsilon) \leq 2 \cdot \frac{1}{2}e^{-\frac{\epsilon^2}{2d}} = e^{-\frac{\epsilon^2}{2d}}$$

Conclusion

We have shown that the probability that the dot product $X^\top Y$ exceeds ϵ is bounded by:

$$P(|X^\top Y| > \epsilon) \leq e^{-\frac{\epsilon^2}{2d}}$$

This indicates that as d increases. □

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.

Let u_i be a non-zero vector in $\mathbb{R}^n$. The outer product $u_i u_i^\top$ is an $n \times n$ matrix. The rank of a matrix is defined as the dimension of the column space (or row space) of the matrix.

1. **Column Space:** The columns of $u_i u_i^\top$ are all scalar multiples of the vector u_i. Specifically, the j-th column of $u_i u_i^\top$ is given by

$$(u_i u_i^\top) j = u_i (u_i^\top)_j = u_i (uij),$$

   where u_{ij} is the j-th component of u_i. Thus, all columns are in the direction of u_i.

2. **Rank:** Since all columns are multiples of u_i, the column space of $u_i u_i^\top$ is spanned by the single vector u_i. Therefore, the rank of $u_i u_i^\top$ is 1.

Thus, we conclude that:

$$\text{rank}(u_i u_i^\top) = 1.$$

Now, consider the matrix $U = \sum_{i=1}^r u_i u_i^\top$.

1. **Linear Independence:** If the vectors $u_1, u_2, \ldots, u_r$ are linearly independent, then the outer products $u_i u_i^\top$ will contribute to the rank of U. Each $u_i u_i^\top$ adds a new dimension to the column space of U.

2. **Rank of the Sum:** The rank of a sum of matrices is at most the sum of their ranks. Since each $u_i u_i^\top$ has rank 1, we have:

$$\text{rank}(U) \leq \sum_{i=1}^r \text{rank}(u_i u_i^\top) = r.$$

3. **Achieving Rank r:** If $u_1, u_2, \ldots, u_r$ are linearly independent, then the rank of U is exactly r. This is because the column space of U will be spanned by the vectors $u_1, u_2, \ldots, u_r$, which are r linearly independent vectors.

Thus, we conclude that:

$$\text{rank}(U) = r.$$

SVD on U

The Singular Value Decomposition (SVD) of a matrix U can be expressed as:

$$U = V \Sigma V^\top,$$

where:

- V is an orthogonal matrix whose columns are the left singular vectors of U,
- Σ is a diagonal matrix containing the singular values of U.

Given that $U = \sum_{i=1}^{r} u_i u_i^\top$:

1. **Eigenvalues and Eigenvectors:** The eigenvalues of U are the singular values. Since U is a sum of rank-1 matrices, the non-zero eigenvalues of U will correspond to the squared norms of the vectors u_i. Specifically, if u_i has norm $\|u_i\|$, then the singular values σ_i are given by:
$$\sigma_i = \|u_i\| \quad \text{for } i = 1, 2, \ldots, r.$$
2. **Constructing V:** The columns of V can be taken as the normalized vectors $\frac{u_i}{\|u_i\|}$ for $i = 1, 2, \ldots, r$. If $r < n$, the remaining columns of V can be filled with orthogonal vectors to complete the orthogonal basis.

Thus, the SVD of U can be summarized as:

$$U = V \Sigma V^\top,$$

where V contains the normalized vectors and Σ contains the singular values corresponding to the vectors u_i. □

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.

$X \sim \text{Uniform}(B_1)$: X is uniformly distributed over the unit ball $B_1 \subset \mathbb{R}^d$. This means that X is a random vector with a uniform distribution inside the unit ball $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$.

$Y = \|X\|_2$: Y is the ℓ_2-norm (Euclidean norm) of X.

We need to: 1. Find the distribution function of Y. 2. Analyze $\ln(1/Y)$ and calculate $\mathbb{E}[\ln(1/Y)]$.

Uniform Distribution in B_1: The volume element in $\mathbb{R}^d$ is proportional to r^{d-1} in spherical coordinates, where $r = \|x\|_2$. For X uniformly distributed over B_1, the probability density function (PDF) is constant inside the unit ball:

$$f_X(x) = \begin{cases} \frac{1}{\text{Vol}(B_1)} & \text{if } \|x\|_2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\text{Vol}(B_1)$ is the volume of the unit ball in $\mathbb{R}^d$, given by:

$$\text{Vol}(B_1) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2} + 1\right)}.$$

Distribution of Y: The cumulative distribution function (CDF) of Y, $F_Y(y)$, is:

$$F_Y(y) = P(Y \leq y) = P(\|X\|_2 \leq y).$$

Since $\|X\|_2 \leq y$ corresponds to the ball of radius y (B_y), the probability is proportional to the volume of this smaller ball:

$$F_Y(y) = \begin{cases} \frac{\text{Vol}(B_y)}{\text{Vol}(B_1)} & \text{if } 0 \leq y \leq 1, \\ 1 & \text{if } y > 1. \end{cases}$$

The volume of B_y scales as y^d, so:

$$\text{Vol}(B_y) = y^d \cdot \text{Vol}(B_1).$$

Thus:

$$F_Y(y) = \begin{cases} y^d & \text{if } 0 \leq y \leq 1, \\ 1 & \text{if } y > 1. \end{cases}$$

PDF of Y: The PDF of Y, $f_Y(y)$, is the derivative of $F_Y(y)$:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} d \cdot y^{d-1} & \text{if } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z = \ln(1/Y) = -\ln(Y)$. Since $Y > 0$, Z is well-defined.

PDF Transformation: The relationship between Z and Y implies $Y = e^{-Z}$. Using the change of variables formula:

$$f_Z(z) = f_Y(y) \left| \frac{dy}{dz} \right| = f_Y(e^{-z}) \cdot e^{-z}.$$

Substitute $f_Y(y) = d \cdot y^{d-1}$ for $0 \leq y \leq 1$, which means $e^{-z} \in (0,1)$, or $z > 0$:

$$f_Z(z) = d \cdot (e^{-z})^{d-1} \cdot e^{-z} = d \cdot e^{-dz}, \quad z \geq 0.$$

Distribution of Z: Z follows an exponential distribution with rate parameter d:

$$f_Z(z) = d \cdot e^{-dz}, \quad z \geq 0.$$

Calculating $\mathbb{E}[\ln(1/Y)]$

Since $Z = \ln(1/Y)$, $\mathbb{E}[\ln(1/Y)] = \mathbb{E}[Z]$. For $Z \sim \text{Exp}(d)$, the expected value is:

$$\mathbb{E}[Z] = \frac{1}{d}.$$

Final Results

1. **Distribution Function of Y:**

$$F_Y(y) = \begin{cases} y^d & \text{if } 0 \leq y \leq 1, \\ 1 & \text{if } y > 1. \end{cases}$$

2. **PDF of Y:**

$$f_Y(y) = \begin{cases} d \cdot y^{d-1} & \text{if } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

3. **PDF of $\ln(1/Y)$:**

$$f_Z(z) = d \cdot e^{-dz}, \quad z \geq 0.$$

4. **Expected Value of $\ln(1/Y)$:**

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