Group Assignment 3

Daniel Cada (Questions 1 & 5), Albert Bargalló i Sales(Questions 3 & 4), Erik Schahine (Questions 2 & 5), and Moaaz Tameer Islam(Questions 3 & 4)

December 22, 2024

Question 1

The three-state Markov chain transition matrix:

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

1a) Transition Diagram:

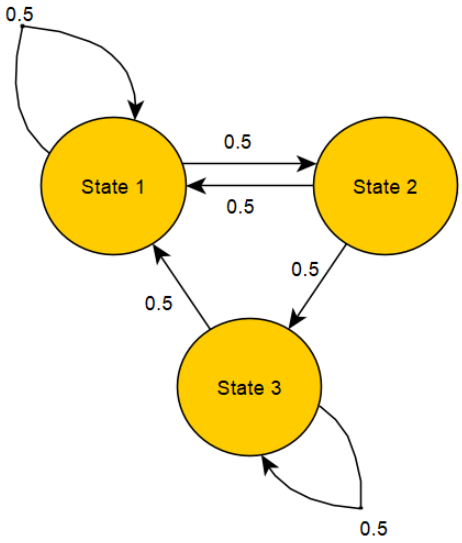

Figure 1: Transition Diagram

1b) Stationary distribution of π:

Writing out the equations:

$$\pi_1 = 0.5\pi_1 + 0.5\pi_2 + 0.5\pi_3, \quad \pi_2 = 0.5\pi_1, \quad \pi_3 = 0.5\pi_1 + 0.5\pi_3.$$

Solving these equations gives:

$$\pi_1 = \frac{1}{2}, \quad \pi_2 = \frac{1}{4}, \quad \pi_3 = \frac{1}{4}.$$

1c) Probability of being in state 2 at time 4, starting from state 1 at time 1:

To find $P(X_4 = 2 \mid X_1 = 1)$, compute the P^3 matrix:

$$P^3 = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

$P(X_4 = 2 \mid X_1 = 1) = 25$

1d) Expected time to reach state 3 for the first time, starting at state 1:

Let h_i be the expected hitting time to state 3 starting from state i. The equations are: We calculate the "hitting time" for each of the starting states:

$$h_3 = 0, \quad h_1 = 1 + 0.5h_1 + 0.5h_2, \quad h_2 = 1 + 0.5h_1 + 0.5h_3.$$

$$\begin{cases} h_1 = 1 + 0.5 \cdot h_1 + 0.5 \cdot h_2 \\ h_2 = 1 + 0.5 \cdot h_1 + 0.5 \cdot h_3 \\ h_3 = 0 \end{cases} \Rightarrow \begin{cases} h_1 = 6 \\ h_2 = 4 \\ h_3 = 0 \end{cases} \tag{1}$$

Thus, the expected time to reach state 3 starting at 1, is 6.

1e) Period of each state:

Since we have loops in the system, we can go back to any state in infinitely many different number of steps. Thus, each of the states are aperiodic.

Question 2

Trying to classify a binary outcome Y, where a classifier $g(X$ has been used for training, we are interested in estimating the following quantities:

$$\begin{aligned} \text{Precision: } & \mathbb{P}(Y = 1 \mid g(X) = 1) \\ \text{Recall: } & \mathbb{P}(g(X) = 1 \mid Y = 1) \end{aligned} \tag{2}$$

(a)

Precision and recall can be described as:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} = \frac{TP}{PP} \\ \text{Recall} &= \frac{TP}{TP + FN} = \frac{TP}{AP} \end{aligned} \tag{3}$$

where:

- Number of True Positives: TP $\Rightarrow g(X) = 1$ and $Y = 1$
- Number of False Positives: FP $\Rightarrow g(X) = 1$ and $Y = 0$
- Number of False Negatives: FN $\Rightarrow g(X) = 0$ and $Y = 1$
- Number of Predicted Positives: PP = TP + FP $\Rightarrow g(X) = 1$
- Number of Actual Positives: AP = TP + FN $\Rightarrow Y = 1$

Therefore, we have the empirical versions:

$$\begin{aligned} \text{Empirical Precision} &= \frac{\sum_{i=1}^{n} 1(g(X_i) = 1 \wedge Y_i = 1)}{\sum_{i=1}^{n} 1(g(X_i) = 1)} \\ \text{Empirical Recall} &= \frac{\sum_{i=1}^{n} 1(g(X_i) = 1 \wedge Y_i = 1)}{\sum_{i=1}^{n} 1(Y_i = 1)} \end{aligned} \tag{4}$$

(b)

Let the random variable representing the cost of the decision $g(X)$ be:

$$C = \begin{cases} c, & \text{if } \quad g(X) = 1 \quad \text{and} \quad Y = 0 \quad \text{(FP)} \\ d, & \text{if } \quad g(X) = 0 \quad \text{and} \quad Y = 1 \quad \text{(FN)} \\ 0, & \text{otherwise (TP or TN)} \end{cases} \tag{5}$$

Adapting equation (2):

$$\begin{aligned} \mathbb{P}(Y = 0 \mid g(X) = 1) &= 1 - \text{Precision} \\ \mathbb{P}(g(X) = 0 \mid Y = 1) &= 1 - \text{Recall} \end{aligned} \tag{6}$$

From the Bayes' Theorem we know:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A) = \mathbb{P}(B) \cdot \mathbb{P}(A \mid B) \tag{7}$$

Then:

$$\begin{aligned} \mathbb{P}(FP) = \mathbb{P}(g(X) = 1 \text{ and } Y = 0) &= \mathbb{P}(Y = 0 \mid g(X) = 1) \cdot \mathbb{P}(g(X) = 1) \\ \mathbb{P}(FN) = \mathbb{P}(g(X) = 0 \text{ and } Y = 1) &= \mathbb{P}(g(X) = 0 \mid Y = 1) \cdot \mathbb{P}(Y = 1) \end{aligned} \tag{8}$$

The expected cost is:

$$\mathbb{E}[C] = c \cdot \mathbb{P}(FP) + d \cdot \mathbb{P}(\mathbb{FN}) \tag{9}$$

In terms of precision and recall:

$$\mathbb{E}[C] = c \cdot [(1 - \text{Precision}) \cdot \mathbb{P}(g(X) = 1)] + d \cdot [(1 - \text{Recall}) \cdot \mathbb{P}(Y = 1)] \tag{10}$$

(3)

Using Hoeffding's inequality to produce a confidence interval for the expected cost:

$$\begin{aligned} \epsilon &= \sqrt{\frac{\max(c,d)^2 \ln(\frac{2}{\delta})}{2n}} \\ \overline{C} &= \frac{1}{n} \sum_{i=1}^{n} C_i \end{aligned} \tag{11}$$

We obtain the following confidence interval for the expected cost:

$$[\, \overline{C} - \epsilon, \quad \overline{C} + \epsilon \,] \tag{12}$$

To compute confidence intervals for precision and recall we could use bootstrapping. First, the original dataset is resampled many times (B) with replacement to create simulated datasets. For each bootstrap sample, we compute precision and recall. Then, the bootstrap distributions for *Precision_1, Precision_2, ..., Precision_B* and for *Recall_1, Recall_2, ..., Recall_B* are calculated. Finally, using these distributions, we can obtain confidence intervals for precision and recall:

$$\begin{aligned} \text{CI Precision: } & [Precision_{\alpha/2}, \quad Precision_{-\alpha/2}] \\ \text{CI Recall: } & [Recall_{\alpha/2}, \quad Recall_{-\alpha/2}] \end{aligned} \tag{13}$$

i think it mainly asked about an anal

Question 3

Let $X = (X_1, X_2, \ldots, X_d)$ and $Y = (Y_1, Y_2, \ldots, Y_d)$ be two d-dimensional zero-mean, unit-variance Gaussian random vectors. Assume: $X_i, Y_i \sim \mathcal{N}(0, 1)$ are independent and identically distributed and X and Y are independent.

We aim to show that X and Y are nearly orthogonal by calculating their dot product $X \cdot Y$ and bounding the probability that $|\cos(\theta)| > \varepsilon$, where wee use cosine similarity $\cos(\theta) = \frac{X \cdot Y}{\|X\| \|Y\|}$.

The dot product of X and Y is:

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

by assumption

Since X_i and Y_i are independent standard normal random variables, each product $X_i Y_i$ is the product of two independent $\mathcal{N}(0, 1)$ variables. The expected value of each term is:

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

The variance of $X_i Y_i$ is:

$$\text{Var}(X_i Y_i) = \mathbb{E}[(X_i Y_i)^2] - (\mathbb{E}[X_i Y_i])^2 =$$

$$\text{Var}(X_i)\text{Var}(Y_i) + \text{Var}(X_i)\mathbb{E}({X_i}^2) + \text{Var}(Y_i)\mathbb{E}({X_i}^2) = 1 \cdot 1 + 0 + 0 = 1$$

Since $X \cdot Y$ is a sum of d independent, identically distributed random variables $X_i Y_i$, the sum has mean zero and $\sum_{i=1}^{d} 1 = d$ variance:

$$X \cdot Y \sim \mathcal{N}(0, d).$$

The cosine of the angle θ between X and Y is defined as:

$$\cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|}.$$

Each $\|X\|^2$ and $\|Y\|^2$ follows a $\chi^2(d)$ distribution since they are the sum of squared normal variables. Thus we have that $\mathbb{E}(\|X\|^2) = d$ (since the expectd value of a chi squared distribution is the number of degrees of freedom, in this case d) and for large d, the norms concentrate around $\sqrt{d}$

$$\|X\| \approx \sqrt{d}, \quad \|Y\| \approx \sqrt{d}.$$

Substituting, we find that:

$$\cos(\theta) \approx \frac{X \cdot Y}{d}.$$

Since $X \cdot Y \sim \mathcal{N}(0, d)$, we have:

$$\cos(\theta) \sim \mathcal{N}\left(0, \frac{1}{d}\right).$$

As d grows large, $cos(\theta)$ approaches $\mathcal{N}(0, 0)$. i.e converges in probability to 0. This means that X and Y become "nearly" orthogonal.
To bound $P(|\cos(\theta)| > \varepsilon)$, note that:

$$P(|\cos(\theta)| > \varepsilon) = P\left(\left|\frac{X \cdot Y}{d}\right| > \varepsilon\right) = P\left(|X \cdot Y| > \varepsilon d\right).$$

Since $X \cdot Y \sim \mathcal{N}(0, d)$, let $Z = \frac{X \cdot Y}{\sqrt{d}} \sim \mathcal{N}(0, 1)$. Then:

$$P(|X \cdot Y| > \varepsilon d) = P\left(|Z| > \varepsilon\sqrt{d}\right).$$

For a standard normal random variable $Z \sim \mathcal{N}(0, 1)$, the probability can be approximated as:

$$P(|Z| > x) = 2 \cdot (1 - \Phi(x)) \approx \frac{2}{\sqrt{2\pi}x} e^{-x^2/2} \quad \text{for large } x.$$

Substituting $x = \varepsilon\sqrt{d}$, we get:

$$P(|X \cdot Y| > \varepsilon) \approx \frac{2}{\sqrt{2\pi}\varepsilon\sqrt{d}} e^{-\frac{\varepsilon^2 d}{2}}.$$

This probability decays exponentially with d. For large d, the cosine of the angle between X and Y becomes arbitrarily small, implying once again that X and Y are "nearly" orthogonal.

Question 4

We are given that u_i is an $n \times 1$ unit-length vector, meaning $\|u_i\| = 1$, and the vectors $u_1, \ldots, u_r$ are linearly independent.

Part (a): $u_i u_i^T$

To show that $u_i u_i^T$ is rank one, we analyze its structure:

- u_i is an $n \times 1$ column vector.
- u_i^T is a $1 \times n$ row vector.

The product $A_i = u_i u_i^T$ results in an $n \times n$ matrix. Each entry of this matrix is the product of corresponding elements from u_i and u_i^T.

- Every column of A_i is a scalar multiple of u_i.
- Every row of A_i is a scalar multiple of u_i^T.

Since the matrix is entirely generated by the vector u_i, its column space is spanned by u_i, meaning the matrix has only one linearly independent column. Thus, the rank of A_i is 1.

Null-Space of $u_i u_i^T$

We know that $A_i x = u_i(u_i^T x)$. Suppose x is orthogonal to u_i, meaning $u_i^T x = 0$. Then the null-vector equation comes out as:

$$A_i x = u_i(u_i^T x) = u_i(0) = 0$$

This is the only solution, as x is assumed to be non-zero in the null-vector equation. Similarly for non-zero u_i values, the value of $u_i u_i^T \neq 0$. So the only way the equation becomes 0 is if $u_i^T x = 0$. The null-space then can be defined as the subspace spanned by this combination, i.e where x is orthogonal to the u_i^T. Also known as the perp! Short for perpendicular complement :D

Range of $u_i u_i^T$

The matrix $A_i = u_i u_i^T$ maps any vector x to:

$$A_i x = u_i(u_i^T x)$$

Since the result is always a scalar multiple of u_i, the range of A_i consists of all vectors in the direction of u_i. Thus, the range is:

$$\text{Range}(A_i) = \{u_i c : c \in \mathbb{R}\}$$

Thus, the range of $u_i u_i^T$ is the subspace spanned by u_i.

Conclusion:

- The matrix $u_i u_i^T$ is rank one.
- The null-space is $\{x \in \mathbb{R}^n : u_i^T x = 0\}$, the orthogonal complement of u_i.
- The range is span(u_i), the subspace spanned by u_i.

Part (b): Verifying $U = \sum_{i=1}^{r} u_i u_i^T$ is a Rank r Matrix

To verify that $U = \sum_{i=1}^{r} u_i u_i^T$ is rank r, consider the following:
1. Each matrix $u_i u_i^T$ is rank 1, with its range spanned by u_i.
2. Since the vectors $u_1, \ldots, u_r$ are linearly independent by assumption, their spans are distinct and non-overlapping.

Proof The column space of U is the span of all vectors u_i, meaning:

$$\text{Range}(U) = \text{span}(\{u_1, \ldots, u_r\})$$

Since $\{u_1, \ldots, u_r\}$ are linearly independent, the dimension of this space is r. Therefore, the rank of U is r.

Additionally, any linear combination of u_i that results in a zero vector would require the coefficient to be 0, as stated at the beginning of this question.

Part (c): Singular Value Decomposition

Singular Value Decomposition (SVD) is a fundamental matrix factorization technique that decomposes a matrix A into three components:

$$A = U\Sigma V^T \tag{14}$$

Where U is an $m \times m$ orthogonal matrix of left singular vectors, Σ is an $m \times n$ diagonal matrix of singular values & V^T is the transpose of an $n \times n$ orthogonal matrix of right singular vectors. Our matrix, which is made as the outer product of 2 unit vectors, will likely decompose with fixed properties.

Case i

The answer is that it will not always be the case. Because the right singular vectors are equal to the eigenvectors of U^TU. But we can use the fact that the matrix will always be symmetric, which means that the left and right singular vectors are the same, i.e the eigenvectors of U. Using this information, any case where u_i is not orthogonal to its complement means it won't be equal to its singular vector. Because SVD always requires it's singular vectors to be orthogonal to each other. Counter-example below:

$$u_1 = \begin{bmatrix}1\\0\end{bmatrix}, \quad u_2 = \begin{bmatrix}1\\1\end{bmatrix}$$

The matrix $U = u_1u_1^T + u_2u_2^T$ is:

$$U = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}$$

The eigenvectors of this matrix, which are the singular vectors, are not u_1 and u_2, but the orthogonal eigenvectors of U.

Case ii: Orthogonal Vectors

When $u_1, \ldots, u_r$ are mutually orthogonal unit vectors, the matrix $U = \sum_{i=1}^{r} u_iu_i^T$ becomes a diagonal matrix. In this case:

$$\text{Singular Values of } U = \{1, 1, \ldots, 1\} \ (r \text{ times}) \tag{15}$$

The right singular vectors in the SVD will exactly match the original orthogonal vectors $u_1, \ldots, u_r$, as expected from our theory in part i).

Question 5

5a) Distribution function of Y

To calculate the probability distribution, we can look at the cumulative distribution function (CDF) of the function. Since we are working within a hypersphere in the *n-th* dimension, we can calculate the CDF by looking at the probability that our uniformly chosen point p is within or outside of any radius r of our hypersphere. The volume of a hypersphere in any dimension is given by the formula:

$$V_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \cdot r^n \tag{16}$$

The CDF is the probability of our point being r units away from the center of the unit ball or less. This is equivalent with taking the ratio between our hypersphere of radius r and the unit ball:

$$\text{CDF} = F(r) = \frac{\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \cdot r^n}{\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \cdot 1^n} \quad \text{for} \quad 0 \le r \le 1 \tag{17}$$

Leaving us with just the following:

$$\begin{cases} 0 & r < 0 \\ r^n & 0 \le r \le 1 \\ 1 & r > 1 \end{cases} \tag{18}$$

To get the probability density function, we just derive the CDF, giving us the final PDF:

$$\text{PDF} = f(r) = n \cdot r^{n-1} \tag{19}$$

5b) Distribution of $ln(\frac{1}{Y})$:

We can define $z = ln(\frac{1}{Y}) = -ln(Y)$. To compute the distribution of Z, we can calculate the *CDF* of Z using Y.
For $Z \le z$, we have:

$$ln(\frac{1}{Y}) \le z \Rightarrow \frac{1}{Y} \le e^z \Rightarrow e^{-z} \tag{20}$$

$$F_Z = P(Z \le z) = P(Y \ge e^{-z}) \tag{21}$$

Using the CDF we calculated in 5a; r^n we can define the probability $P(Y \geq e^{-z})$ as:

$$P(Y \geq e^{-z}) = 1 - F_Z(e^{-z}) = 1 - (e^{-z})^{r^n} = 1 - e^{-r^n z} \quad \text{for} \quad z \geq 0 \tag{22}$$

Therefore, the CDF F_Z is:

$$F_Z = \begin{cases} 0 & z < 0, \\ 1 - e^{-nz} & z \geq 0. \end{cases} \tag{23}$$

Differentiating F_Z with respect to z, we find the probability density function (PDF) of Z, f_Z:

$$f_Z = \frac{d}{dz} F_Z(z) = ne^{-nz} \tag{24}$$

5c

We calculate $\mathbb{E}[\ln(1/Y)]$ using two methods: from the distribution of Y and from the distribution of $Z = \ln(1/Y)$.
The expectation is given by:

$$\mathbb{E}[\ln(1/Y)] = \int_0^1 \ln(1/y) f_Y(y)\, dy,$$

where the PDF of Y is:

$$f_Y(y) = ny^{n-1}, \quad y \in [0, 1].$$

$$\mathbb{E}[\ln(1/Y)] = \int_0^1 \ln(1/y) n y^{n-1}\, dy.$$

Using the logarithmic identity $\ln(1/y) = -\ln(y)$, we rewrite the expectation as:

$$\mathbb{E}[\ln(1/Y)] = -n \int_0^1 \ln(y) y^{n-1}\, dy.$$

The integral $\int_0^1 \ln(y) y^{n-1}\, dy$ can be computed with integration by parts by differentiating $ln(y)$ and integrating y^{n-1}:

$$\int_0^1 \ln(y) y^{n-1}\, dy = -\frac{1}{n^2}.$$

Substituting this result:

$$\mathbb{E}[\ln(1/Y)] = -n \cdot \left(-\frac{1}{n^2}\right) = \frac{1}{n}.$$

Thus:

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

We can also use the Distribution of $Z = \ln(1/Y)$
The PDF of Z is:

$$f_Z(z) = ne^{-nz}, \quad z \in [0, \infty).$$

The expectation is:

$$\mathbb{E}[Z] = \int_0^\infty z f_Z(z)\, dz.$$

$$\mathbb{E}[Z] = \int_0^\infty z n e^{-nz}\, dz.$$

The integral $\int_0^\infty ze^{-nz}\,dz$ can be computed by partial integration, taking the derivative of z twice, and integrating e^{-nz} twice.

$$\int_0^\infty ze^{-nz}\,dz = \frac{1}{n^2}.$$

$$\mathbb{E}[Z] = n \cdot \frac{1}{n^2} = \frac{1}{n}.$$

Using both methods, we find:

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