Group Assignment 3

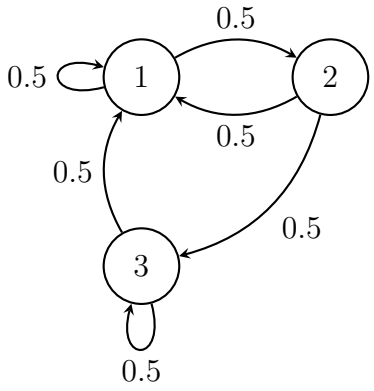
Holger Swartling, Nir Teyar

20-12-2024

Solution Q1

(a) Transition Diagram

The transition diagram of the Markov chain that was given by the matrix looks as following when illustrated.



(b) Stationary Distribution π

The stationary distribution is satisfied by the following procedure:

$$\pi P = \pi, \quad \text{where } \pi = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}, \quad \pi_1 + \pi_2 + \pi_3 = 1.$$

By expanding π we can attain the following relationship: P = π, we obtain:

$$\begin{aligned} \pi_1 &= 0.5\pi_1 + 0.5\pi_2 + 0.5\pi_3, \\ \pi_2 &= 0.5\pi_1, \\ \pi_3 &= 0.5\pi_2 + 0.5\pi_3. \end{aligned}$$

Given the second equation we attain, $\pi_2 = 0.5\pi_1$. We later use substitution for the following $\pi_2 = 0.5\pi_1$ into the third equation:

$$\pi_3 = 0.5(0.5\pi_1) + 0.5\pi_3 \implies \pi_3 = 0.5\pi_1.$$

The normalization equation that is described above is then used like so:

$$\pi_1 + 0.5\pi_1 + 0.5\pi_1 = 1 \implies 2\pi_1 = 1 \implies \pi_1 = 0.5.$$

We will finally substitute $\pi_1 = 0.5$ into $\pi_2 = 0.5\pi_1$ and $\pi_3 = 0.5\pi_1$:

$$\pi_2 = 0.5(0.5) = 0.25, \quad \pi_3 = 0.5(0.5) = 0.25.$$

This gives us the answer, which is the stationary distribution:

$$\pi = \begin{pmatrix} 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

(c) Probability of State 2 at Time 4

The initial state vector is $\mathbf{v}_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$. The state vector at time t is given by:

$$\mathbf{v}_t = \mathbf{v}_1 P^{t-1}.$$

We compute P^3:

$$P^2 = P \cdot P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

$$P^3 = P^2 \cdot P = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix} \cdot \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

$$\mathbf{v}_4 = \mathbf{v}_1 P^3 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

The probability of being in state 2 at time 4 is 0.25.

(d) Expected Time to Reach State 3

We define $T_{1\to 3}$ as the expected time for our chain to reach state 3 when starting from state 1:

$$T_{1\to 3} = 1 + P(1 \to 1)T_{1\to 3} + P(1 \to 2)T_{2\to 3}.$$

Which is almost the same for $T_{2\to 3}$:

$$T_{2\to 3} = 1 + P(2 \to 1)T_{1\to 3}.$$

We then replace the transition probabilities:

$$T_{1\to 3} = 1 + 0.5T_{1\to 3} + 0.5T_{2\to 3},$$

$$T_{2\to 3} = 1 + 0.5T_{1\to 3}.$$

Use equation 2:

$$T_{2\to 3} = 1 + 0.5T_{1\to 3}.$$

We subtitute $T_{2\to 3}$ into the our first equation:

$$T_{1\to 3} = 1 + 0.5T_{1\to 3} + 0.5(1 + 0.5T_{1\to 3}).$$

Further, simplify the given expression and then rearrange it:

$$T_{1\to 3}(1 - 0.75) = 1.5 \implies T_{1\to 3} = \frac{1.5}{0.25} = 6.$$

Thus, $T_{1\to 3} = 6$ and $T_{2\to 3} = 1 + 0.5(6) = 4$.

(e) Period of Each State

1. State 1: The chain can return to state 1 in a cycle of length 1 ($1 \to 1$), a cycle of length 2 ($1 \to 2 \to 1$) and a cycle of length 3 ($1 \to 2 \to 3 \to 1$). Since the cycle can self-transition ($3 \to 3$) an arbitrary number of times, there exist cycles of length n for every $n \geq 3$. The greatest common divisor is 1 which is the case for all chains that include a state that can self-transition.
   Period = 1

2. State 2: The chain can return to state 2 in a cycle of length 2 ($2 \to 1 \to 2$) and a cycle of length 3 ($2 \to 3 \to 1 \to 2$). In addition since it can also self-transition in both state 1 and state 3 there are cycles of length n for every $n \geq 3$
   Period = 1

3. State 3: The chain can return to state 3 in a cycle of length 1 ($3 \to 3$), a cycle of length 3 ($3 \to 1 \to 2 \to 3$) and cycles of length n for every $n \geq 3$ due to the self-transition to state 1.
   Period = 1

Solution Q2

(a) Empirical Precision and Recall

Precision and recalls, $g(X)$, defined as:

$$\begin{aligned} \text{Precision} &= P(Y = 1 \mid g(X) = 1), \\ \text{Recall} &= P(g(X) = 1 \mid Y = 1). \end{aligned}$$

Using testing data and probabilities we can estimate empirically. The empirical definitions are as follows:

- **Empirical Precision:**

$$\begin{aligned} \text{Precision} &= \frac{\text{Number of true positives}}{\text{Number of predicted positives}} \\ &= \frac{\sum_{i=1}^{n} \mathbb{1}\{g(X_i) = 1, Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{g(X_i) = 1\}}, \end{aligned}$$

  $\mathbb{1}\{\cdot\}$ is the indicator function that is equal 1 if the condition is true and 0 otherwise. Numerator counts instances in which the given classifier correctly predicts $Y = 1$, and denominator counts all instances where the classifier predicts $g(X) = 1$.

- **Empirical Recall:**

$$\begin{aligned} \text{Recall} &= \frac{\text{Number of true positives}}{\text{Number of actual positives}} \\ &= \frac{\sum_{i=1}^{n} \mathbb{1}\{g(X_i) = 1, Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\}}. \end{aligned}$$

  We know that the numerator is the same as for precision, and the denominator counts the total number of positives in the dataset (i.e., instances where $Y = 1$).

Example Illustration: We will assume that the testing dataset has the following confusion matrix:

	$Y = 1$	$Y = 0$
$g(X) = 1$	TP	FP
$g(X) = 0$	FN	TN

Here:

- TP (True Positives): Cases where $g(X) = 1$ and $Y = 1$.
- FP (False Positives): Cases where $g(X) = 1$ but $Y = 0$.
- FN (False Negatives): Cases where $g(X) = 0$ but $Y = 1$.
- TN (True Negatives): Cases where $g(X) = 0$ and $Y = 0$.

Using the confusion matrix, precision and recall can be expressed as:

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

The given metrics provide insights into the performance of the classifier, specifically its ability to avoid false positives (precision) and false negatives (recall).

(b) Expected Cost of Decision $g(X)$

To analyse the cost associated with the classifier's decisions, we define the random variable C as follows:

$$C = \begin{cases} c & \text{if } g(X) = 1 \text{ and } Y = 0, \\ d & \text{if } g(X) = 0 \text{ and } Y = 1, \\ 0 & \text{otherwise.} \end{cases}$$

This setup give us the folllowing two types of costs:

- **False Positive Cost (c):** When the classifier predicts $g(X) = 1$, battery deteriorated, but the battery is not deteriorated ($Y = 0$). This means that a cost of c occurs, representing the expense of an unnecessary test.
- **False Negative Cost (d):** When the classifier predicts $g(X) = 0$, battery healthy, but the battery is actually deteriorated ($Y = 1$). This means that a cost of d occurs, representing the damage caused by battery failure.

We calculate the expected cost as follows:

$$\mathbb{E}[C] = c \cdot P(g(X) = 1 \mid Y = 0) \cdot P(Y = 0) + d \cdot P(g(X) = 0 \mid Y = 1) \cdot P(Y = 1).$$

Rewrite this in terms of precision and recall:

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

Here:

- $P(g(X) = 1)$ represents the proportion of instances where the classifier predicts $g(X) = 1$, estimated as:
$$P(g(X) = 1) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(X_i) = 1\}.$$
- $P(Y = 1)$ represents the proportion of actual positives in the dataset, estimated as:
$$P(Y = 1) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\}.$$

(c) Confidence Intervals for Expected Cost, Precision, and Recall

In order to estimate confidence intervals for precision, recall, and expected cost, we may use the bootstrap method. The bootstrap makes use of a resampling-based approach that is able to resample multiple estimate through sampling the data with replacement.

Steps to Compute Confidence Intervals:

1. **Resample the data:** We generate B bootstrap samples through resampling the testing data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with replacement.
2. **Compute metrics for each sample:** In each bootstrap sample, calculate precision, recall, and expected cost by using the same formulas described earlier.
3. **Obtain the confidence intervals:** Each metric, sort the B bootstrap estimates and take the $\alpha/2$ and $1 - \alpha/2$ percentiles to form the confidence interval.

Example: Lets suppose that we generate $B = 1000$ bootstrap samples and generate the following results for precision:

- Lower bound (2.5th percentile): 0.75,
- Upper bound (97.5th percentile): 0.85.

Then, the 95% confidence interval for precision is:

$$CI_{\text{Precision}} = [0.75, 0.85].$$

In the same way confidence intervals can be computed for recall and expected cost.

Solution Q3

Properties of Gaussian Random Vectors

We know from the exercise that X and Y are d-dimensional random vectors.

- $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathcal{N}(0, I_d)$, where I_d is the $d \times d$ identity matrix.
- X and Y are independent.

The dot product $Z = X \cdot Y = \sum_{i=1}^{d} X_i Y_i$. Each X_i and Y_i are independent Gaussian random variables with mean 0 and variance 1.

Distribution of the Dot Product Z

The dot product is to be written as:

$$Z = \sum_{i=1}^{d} X_i Y_i.$$

Since $X_i \sim \mathcal{N}(0, 1)$ and $Y_i \sim \mathcal{N}(0, 1)$, their product $X_i Y_i$ has:

- Mean: $\mathbb{E}[X_i Y_i] = \mathbb{E}[X_i] \cdot \mathbb{E}[Y_i] = 0$.
- Variance: $\text{Var}(X_i Y_i) = \mathbb{E}[X_i^2] \cdot \mathbb{E}[Y_i^2] = 1 \cdot 1 = 1$.

Since Z is a sum of d independent random variables $X_i Y_i$, the Central Limit Theorem implies that for large d, Z approximately follows a Gaussian distribution:

$$Z \sim \mathcal{N}(0, d).$$

Bounding $P(|Z| > \epsilon)$

To bound the probability that $|Z| > \epsilon$, we make use of the given properties of the standard normal distribution. First, we go ahead and standardize Z:

$$Z_{\text{std}} = \frac{Z}{\sqrt{d}} \sim \mathcal{N}(0, 1).$$

Probability is given by:

$$P(|Z| > \epsilon) = P\left(\left|\frac{Z}{\sqrt{d}}\right| > \frac{\epsilon}{\sqrt{d}}\right).$$

We make use of the complementary cumulative distribution function of the standard normal distribution:

$$P\left(|Z_{\text{std}}| > \frac{\epsilon}{\sqrt{d}}\right) = 2\Phi\left(-\frac{\epsilon}{\sqrt{d}}\right),$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Interpretation

For large d, the dot product Z is approximately the Gaussian with a mean 0 and variance d. The probability that Z exceeds a given fixed threshold ϵ decreases as d increases. This tells us that X and Y are nearly orthogonal in high-dimensional spaces.

Solution Q4

(a) Verifying Rank-One Property of $u_i u_i^T$

The given matrix $u_i u_i^T$ is as seen a matrix $n \times n$ formed as the outer product of the vector u_i with itself:

$$u_i u_i^T = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{in} \end{pmatrix} \begin{pmatrix} u_{i1} & u_{i2} & \cdots & u_{in} \end{pmatrix}.$$

Properties:

- $u_i u_i^T$ is symmetric because $(u_i u_i^T)^T = u_i u_i^T$.
- $u_i u_i^T$ is a rank one because it is given by the outer product of a single vector. This means every column is a scalar multiple of u_i which means they are all linearly dependent. The only nonzero eigenvalue in this case is $\|u_i\|^2 = 1$.

Null Space: The null space of $u_i u_i^T$ is a part of all vectors orthogonal to u_i, i.e.,

$$\text{Null}(u_i u_i^T) = \{v \in \mathbb{R}^n \mid u_i^T v = 0\}.$$

Range: The range of $u_i u_i^T$ is the span of its columns. Since all columns are linearly dependent the range is just the span of u_i, i.e.,

$$\text{Range}(u_i u_i^T) = \text{Span}(\{u_i\}).$$

(b) Verifying Rank-r Property of U

The given matrix U is given by:

$$U = \sum_{i=1}^{r} u_i u_i^T.$$

Since we know from assumption that $u_1, \ldots,$ and u_r are linearly independent, we also know that the rank of U is equal to r:

- Each $u_i u_i^T$ adds one rank to U because u_i is independent of the other vectors.
- The total matrix U has rank r because it is formed by summing r rank-one projections onto linearly independent subspaces.

(c) Singular Value Decomposition (SVD) of U

(i) Are $u_1, \ldots, u_r$ the right singular vectors? No, the vectors $u_1, \ldots, u_r$ are not the right singular vectors of U in every case. The right singular vectors are given by the eigendecomposition of U^TU or UU^T and may not align with $u_1, \ldots, u_r$ unless additional conditions are imposed. $u_1, \ldots, u_r$ are defined as linearly independent but they are not guaranteed to be orthogonal.

Example: If $u_1, \ldots, u_r$ are not orthogonal, the right singular vectors will depend on the Gram-Schmidt process applied to $u_1, \ldots, u_r$ to construct an orthogonal basis.

Let $U = u_1u_1^T + u_2u_2^T$, where $u_1, u_2 \in \mathbb{R}^n$ are linearly independent but not orthogonal.

Suppose:

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

Then:

$$U = u_1u_1^T + u_2u_2^T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

The eigenvectors of U (and thus the right singular vectors) are obtained by solving $(U - \lambda I)v = 0$. They are:

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Since $u_1 \neq v_1$ and $u_2 \neq v_2$ they are not the right singular vectors.

(ii) Orthogonal Case: If $u_1, \ldots, u_r$ are orthogonal, then:

- The singular values of U are all 1.
- The SVD of U aligns with $u_1, \ldots, u_r$, and they are then the right singular vectors.

In our case:

$$U = \sum_{i=1}^{r} u_iu_i^T \quad \text{with singular values } \sigma_i = 1 \text{ for all } i = 1, \ldots, r.$$

Solution Q5

(a) Distribution Function of Y

The random vector X is uniformly distributed in the unit ball $B_1 \subset \mathbb{R}^d$, so $\|X\|_2 \leq 1$ with 1 as a probability. The random variable $Y = \|X\|_2$ gives us the Euclidean norm of X, which takes values in $[0, 1]$.

To find the cumulative distribution function of Y, $F_Y(y)$, note that $F_Y(y)$ is the probability that $\|X\|_2 \leq y$:

$$F_Y(y) = P(Y \leq y) = \frac{\text{Volume of the ball of radius } y}{\text{Volume of the unit ball}}.$$

The volume of a ball of radius y in $\mathbb{R}^d$ is given by:

$V_d(y) = C_d \cdot y^d$, where $C_d = \dfrac{\pi^{d/2}}{\Gamma(d/2+1)}$ is a constant depending on the dimension.

The volume of the unit ball is $V_d(1) = C_d$.

$$F_Y(y) = \frac{V_d(y)}{V_d(1)} = y^d \quad \text{for } y \in [0, 1].$$

The probability density function of Y is the derivative of the CDF:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = d \cdot y^{d-1}, \quad y \in [0, 1].$$

(b) Distribution of $\ln(1/Y)$

Define $Z = \ln(1/Y)$. To find the PDF of Z, we use the transformation method:

$$Z = \ln(1/Y) \implies Y = e^{-Z}.$$

The Jacobian of the transformation is given by:

$$\frac{dY}{dZ} = -e^{-Z}.$$

Substitute to the given PDF of Y:

$$f_Z(z) = f_Y(e^{-z}) \cdot \left| \frac{dY}{dZ} \right| = d \cdot (e^{-z})^{d-1} \cdot e^{-z} = d \cdot e^{-dz}, \quad z \geq 0.$$

Which gives us $Z = \ln(1/Y)$ follows an exponential distribution with rate parameter d:

$$Z \sim \text{Exponential}(d).$$

(c) Expectation of $\ln(1/Y)$

Using the Distribution of Y:

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

Substitute in $\ln(1/y) = -\ln(y)$:

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

Using the given substitution $u = \ln(y)$, $y = e^u$, and $dy = e^u\, du$:

$$\mathbb{E}[\ln(1/Y)] = -d \int_{-\infty}^0 u \cdot e^{du}\, du.$$

Using integration by parts gives us:

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

Using the Distribution of Z: Since $Z \sim \text{Exponential}(d)$, the mean of an exponential random variable with rate d is $1/d$:

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

Contribution statement

The assignment as a whole was done by both group members. The group members did the assignment first individually and then checked answers with each other upon completion, thus resulting in approximately 50% contribution from each member.