Introduction to Data Science group assignment 3

Max Callenmark Kasper Lindkvist Eskil Worm Forss
Rami Abou Zahra Oskar Ådahl

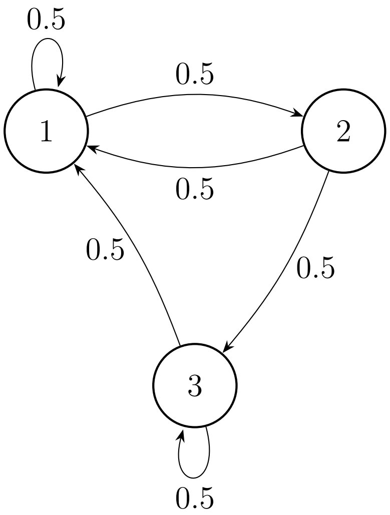
Division Of Work

Task 1: Max Callenmark
Task 2: Kasper Lindkvist
Task 3: Rami Abou Zahra
Task 4: Oskar Ådahl
Task 5: Eskil Worm Forss

1)

a)

The transition diagram is given by:



b)

Since the Markov chain is irreducible and aperiodic there is exist a unique stationary distribution π satisfying:

$$\pi P = \pi \quad \text{and} \quad \sum_{x \in \mathbb{X}} \pi(x) = 1$$

To find π we start by finding the eigenvalues of P^T:

$$\det(P - \lambda I) = 0 \implies \det(P - \lambda I) = -\lambda^3 + \lambda^2 = -\lambda^2(\lambda - 1) = 0 \implies \lambda_1 = 0 \quad \text{and } \lambda_2 = 1$$

We then want to find the eigenvector corresponding to the eigenvalue $\lambda_2 = 1$:

$$(P - \lambda_2 I)v = 0 \implies v = (2, 1, 1)^T$$

Using this eigenvector we can verify that it satisfies the first condition:

$$(2 \quad 1 \quad 1) \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix} = (2 \quad 1 \quad 1)$$

But we don't full fill the second condition, but since it is an eigenvector we can scale with $\frac{1}{4}$. So the stationary distribution is given by :

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

c)

Given that the chain is in state 1 at time 1 then the initial distribution vector is given by $\pi = (1 \quad 0 \quad 0)$. Since we are interested in what happens at time 4 we need to make 3 transitions. The distribution after 3 distributions is obtained by:

$$\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} P^3 = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}^3 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

So the probability that the chain is in state 2 at time 4 is $\dfrac{1}{4}$

d)

Let T_3 be the first time the chain enters state 3. Define

$$h_i = \mathbb{E}[T_3 \mid X_1 = i], \quad i = 1, 2, 3$$

$h_3 = 0$ since we want to find the expected value of hitting state 3 for the first time. Conditioning on the first step, we obtain the following system of equations. From state 1, the chain moves to states 1 and 2 with equal probability, therefore we get:

$$h_1 = 1 + 0.5h_1 + 0.5h_2$$

From state 2, the chain moves to state 1 with probability 0.5 and to state 3 with probability 0.5, which gives:

$$h_2 = 1 + 0.5h_1 + 0.5h_3$$

Since $h_3 = 0$, this simplifies to :

$$h_2 = 1 + 0.5h_1$$

Substituting this expression for h_2 into the equation for h_1, we obtain:

$$h_1 = 1 + 0.5h_1 + 0.5(1 + 0.5h_1) = 1.5 + 0.75h_1 \iff 0.25h_1 = 1.5 \iff h_1 = 6$$

Therefore, the expected time until the chain first enters state 3, given that it starts in state 1, is 6 steps. Therefore, the expected time until the chain first enters state 3, given that it starts in state 1, is 6 steps.

e)

Since $P_{11} = 0.5 > 0$, the chain can return to state 1 in one step, and thus state 1 has period 1. Because the Markov chain is irreducible, all states have the same period. Hence, every state in the chain has period 1.

2)

a)

With prediction like this we have four different cases for the output. We can get a true positive (TP), a false positive (FP), a true negative (TN) or a false negative (FN), given testing data we can compute the empirical versions of these.

$$\begin{aligned}
\hat{TP} &= \sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1) \\
\hat{TN} &= \sum_{i=1}^{n} I(g(X_i) = 0, Y_i = 0) \\
\hat{FP} &= \sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 0) \\
\hat{FN} &= \sum_{i=1}^{n} I(g(X_i) = 0, Y_i = 1)
\end{aligned} \tag{1}$$

Here, I is the identity function being 1 if the inside of the parenthesis holds true and 0 otherwise. So, the sums count when the classifier gives a 1 and Y is 1, a true positive, when the classifier gives a 0 and Y is 0, a true negative, when the classifier gives a 1 but Y is not 1, a false positive, and when the classifier gives a 0 but Y is not 0, a false positive.

Then, we can use the formulas for precision and recall with the above to get the empirical versions.

$$\begin{aligned}
\hat{Precision} = \frac{\hat{TP}}{\hat{TP} + \hat{FP}} &= \frac{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1)}{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1) + \sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 0)} = \\
&= \frac{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1)}{\sum_{i=1}^{n} I(g(X_i) = 1)} \\
\hat{Recall} = \frac{\hat{TP}}{\hat{TP} + \hat{FN}} &= \frac{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1)}{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1) + \sum_{i=1}^{n} I(g(X_i) = 0, Y_i = 1)} = \\
&= \frac{\sum_{i=1}^{n} I(g(X_i) = 1, Y_i = 1)}{\sum_{i=1}^{n} I(Y_i = 1)}
\end{aligned} \tag{2}$$

So, precision and recall are the true positives divided by the classified positives and actual positives respectively.

b)

If $g(X) = 1$, that is we have predicted a battery to have deteriorated, then we need to run a test to confirm if this is true. The cost of running the test is c, if in fact the battery was not deteriorated, $Y = 0$. If instead, $Y = 1$, we do have deterioration of the battery, and the test is not run, that is $g(X) = 0$, then the cost is d.

We can summarize the cost by the precision and the recall. We have that the expected cost is c times the probability that $Y = 0$, given that $g(X) = 1$, which is 1 minus the precision, plus d times the probability that $g(X) = 0$ given that $Y = 1$, which is 1 minus the recall.

So, we let C be a random variable representing the cost of the decision $g(X)$ and then we have the expected cost $E(C)$:

$$\begin{aligned}E(C) = c * P(Y = 0|g(X) = 1) + d * P(g(X) = 0|Y = 1) = \\ = c * (1 - P(Y = 1|g(X) = 1)) + d * (1 - P(g(X) = 1|Y = 1)) = \\ = c * (1 - Precision) + d * (1 - Recall) = \\ = c + d - c * Precision - d^* Recall\end{aligned} \tag{3}$$

Since the formulation of the problem is a bit unclear to me I will also include the answer if instead probability that we have cost c is $P(Y = 0, g(X) = 1)$ and the probability that we have cost d is $P(Y = 1, g(X) = 0)$, that is we are talking about the joint probabilities and not the conditional ones. Here we would have:

$$\begin{aligned}Precision = P(Y = 1|g(X) = 1) \Longrightarrow 1 - Precision = 1 - P(Y = 1|g(X) = 1) = \\ = P(Y = 0|g(X) = 1) = \frac{P(Y = 0, g(X) = 1)}{P(g(X) = 1)} \\ Recall = P(g(X) = 1|Y = 1) \Longrightarrow 1 - Recall = 1 - P(g(X) = 1|Y = 1) = \\ = P(g(X) = 0|Y = 1) = \frac{P(Y = 1, g(X) = 0)}{P(Y = 1)}\end{aligned} \tag{4}$$

So we get instead for random variable C representing cost of $g(X)$:

$$\begin{aligned} E(C) = c * P(Y=0, g(X)=1) + d * P(Y=1, g(X)=0) = \\ = c * P(g(X)=1) * (1 - P(Y=1|g(X)=1)) + d * P(Y=1) * (1 - P(g(X)=1|Y=1)) = \\ = c * P(g(X)=1) * (1 - Precision) + d * P(Y=1) * (1 - Recall) \end{aligned} \tag{5}$$

c)

Since we have that precision is the proportion of classified positives that are true, we could say it is a binomial proportion conditional on the amount of classified positives. Similarly for recall which is the proportion of positives that are classified correctly, we could say it is a binomial proportion conditional on the amount of positives in the training data.

That is we have, for $\sum_{i=1}^{n} I(g(X_i)=1) = m_P$, $\hat{TP} \sim Bin(m_P, Precision)$ and for $\sum_{i=1}^{n} I(Y_i = 1) = m_R$, $\hat{TP} \sim Bin(m_R, Recall)$. Now, using this we can create $100 * (1-\alpha)\%$ Wald confidence intervals using the empirical precision and recall as estimates of the true vales.

$$\begin{aligned} \hat{Precision} \pm z_{1-\alpha/2} * \sqrt{\frac{\hat{Precision}(1-\hat{Precision})}{m_P}} \\ \hat{Recall} \pm z_{1-\alpha/2} * \sqrt{\frac{\hat{Recall}(1-\hat{Recall})}{m_R}} \end{aligned} \tag{6}$$

Note, these two confidence interval only work well for large enough m_P and m_R, since we are using normal approximation by the central limit theorem. Also note, z is the quantile of the standard normal distribution.

To get a confidence interval for the expected cost which is a bit more complex in dependence than the precision and recall we could use the bootstrap method.

We would for each X_i compute $\hat{g} = g(X_i)$ and store the pairs $(\hat{g}, Y_i)$. Then, for a sample of size n as we have, we would draw (with replacement) n samples from the stored pairs and get a new data set to use. Now using this new data set we could compute TP, FP, FN to compute Precision and Recall and then using that compute expected cost $E(C)$. Repeat these steps B amount of times possibly 1000 or maybe 5000 for more accuracy. Then simply take quantiles from the B different stored $E(C)$:s to get a confidence interval. Clearly, this method would also work for getting confidence intervals for precision and recall as well.

3)

For $X, Y \sim N(0, \mathbf{1}_d)$, their dot-product is calculated through the following definition:

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

Assuming independence, we can define $Z = X \cdot Y$ and find its parameters:

$$E\left[Z\right] = E\left[\sum_{i}^{d} X_i Y_i\right] = \sum_{i}^{d} E\left[X_i Y_i\right] = 0$$

$$\operatorname{Var}\left(Z\right) = \sum_{i}^{d} \operatorname{Var}\left(X_i Y_i\right) = \underbrace{\sum_{i}^{d} E\left[X_i^2 Y_i^2\right]}_{=d(1\cdot 1)} - \underbrace{\sum_{i}^{d} \left(E\left[X_i Y_i\right]\right)^2}_{=0} = d$$

In order to show that they are nearly orthogonal, we can use Chebychevs inequality in the following manner:

$$P(|Z| > \varepsilon d) \leq \frac{E\left[Z^2\right]}{(\varepsilon d)^2} = \frac{d}{\varepsilon^2 d^2} = \frac{1}{\varepsilon^2 d} \overset{\text{as } d \to \infty}{\to 0}$$

Therefore, $|Z| \leq \varepsilon d$ almost surely. Since Z was the dot-product, we have shown here that the dot-product is nearly orthogonal and bounded the probability of it being larger than ε.

4)

(a)

Let $U_i = u_i u_i^\top$ for all i. For any i, all columns of U_i are multiples of u_i. Hence, the range of U_i is span($\{u_i\}$). The rank of U_i is the dimension of its range, which obviously is 1. For any vector $v \in \mathbb{R}^n$ we have that $U_i v = (u_i \cdot v) u_i$. This implies that the null-space of U_i is all vectors in $\mathbb{R}^n$ that are orthogonal to u_i.

(b)

As mentioned in (a), the columns of each U_i are multiples of u_i. Hence, in $U = \sum_{i=1}^{r} U_i$ the columns are all linear combinations of $u_1, u_2, \ldots, u_r$. Since all the u_i are linearly independent, the range of U is $\operatorname{span}(\{u_1, u_2, \ldots, u_r\})$ and its rank is r.

(c)

i.

This is not the case. The right singular vectors are orthonormal which the u_i not necessarily are. Consider the following example where $n = 3$ and

$$u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}.$$

Then,

$$U = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Performing SVD on U we get that

$$U = \begin{bmatrix} -0.9238 & -0.3826 & 0 \\ -0.3826 & 0.9238 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1.7071 & 0 & 0 \\ 0 & 0.2928 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.9238 & -0.3826 & 0 \\ -0.3826 & 0.9238 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and we see that the right singular vectors are not the same as the u_i.

ii.

Suppose that the vectors $u_1, u_2, \ldots, u_r$ are all orthogonal. If $r < n$, let $u_{r+1}, \ldots, u_n$ be unit length vectors such that the set $\{u_1, u_2, \ldots, u_n\}$ is an orthonormal basis of $\mathbb{R}^n$. Let the $n \times n$ matrix A be the matrix with the u_i as columns vectors and let $B = A^\top$. Note that both matrices A and B are unitary since their columns and rows respectively form orthonormal bases. Let Σ be a diagonal matrix with the first r diagonal entries being 1 and the rest 0. Now,

$$\begin{aligned} U &= \sum_{i=1}^{r} u_i u_i^\top \\ &= \sum_{i=1}^{n} I_{\{i \le r\}} u_i u_i^\top \\ &= A \Sigma B^\top . \end{aligned}$$

I.e., a SVD for U is $U = A\Sigma B^\top$. Since the singular values of a matrix is the same for each SVD, the singular values for U are r 1s and $n - r$ 0s.

5)

Let $X \sim \text{Uniform}(B_1)$, where

$$B_1 = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\},$$

and define

$$Y = \|X\|_2.$$

a)

For $0 \le r \le 1$,

$$F_Y(r) = \mathbb{P}(Y \le r) = \mathbb{P}(\|X\|_2 \le r) = \frac{|B_r|}{|B_1|}.$$

Since $|B_r| = r^d |B_1|$, here $|\cdot|$ means volume like in the lecture notes, we obtain

$$F_Y(r) = r^d.$$

Thus,

$$F_Y(r) = \begin{cases} 0, & r < 0, \\ r^d, & 0 \le r \le 1, \\ 1, & r \ge 1. \end{cases}$$

The density of Y is

$$f_Y(r) = d r^{d-1}, \qquad 0 < r < 1.$$

b)

Define

$$Z = \ln(1/Y) = -\ln Y.$$

Since $Y \in (0,1]$, we have $Z \in [0,\infty)$.

For $z \geq 0$,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) = \mathbb{P}(-\ln Y \leq z) \\ &= \mathbb{P}(Y \geq e^{-z}) = 1 - \mathbb{P}(Y \leq e^{-z}) \\ &= 1 - (e^{-z})^d = 1 - e^{-dz}. \end{aligned}$$

Hence,

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

with density

$$f_Z(z) = de^{-dz}, \qquad z \geq 0.$$

c)

First, using the distribution of Y:

$$\begin{aligned} \mathbb{E}[\ln(1/Y)] = \mathbb{E}[-\ln Y] = \int_0^1 (-\ln r)\, dr^{d-1}\, dr = -d \int_0^1 (\ln r)\, r^{d-1}\ dr = \\ = -d\left(\left[\frac{r^d \ln r}{d}\right]_0^1 - \int_0^1 \frac{r^{d-1}}{d} dr\right) = -d\left(0 - 0 - \frac{1}{d}\left[\frac{r^d}{d}\right]_0^1\right) = \\ = -d * \frac{1}{d^2}(-1+0) = \frac{1}{d} \end{aligned}$$

So,

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

Now, using the distribution of $\ln(1/Y)$:
Since $Z = \ln(1/Y) \sim \text{Exponential}(d)$,

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

So, both methods gave the same resulting expectation.

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