SNU Fourth Industrial Revolution Academy

Basic Math for Big Data

Homework 2

Due: July 20, 10:00 AM

Reminders

- T.A.: Chiwan Park (chiwanpark@snu.ac.kr)
- The points of this homework add up to 100.
- This has to be done individually like all the homeworks.
- Please answer clearly; illegible handwriting may get no points.
- Whenever you are making an assumption, please state it clearly.
- If you have a question about assignments, please upload your question in FIRA portal.

Submissions

- You can submit your homework in the class or via email (only PDFs are accepted).
- Do not submit the homework in a photography form.

Question 1 [16 points]

Let S be a set with n elements, and let a and b be distinct elements of S. How many relations R are there such that

(a) $(a, b) \in R$?

2^{n^2-1}

(b) $(a, b) \notin R$?

2^{n^2-1}

(c) no ordered pair in R has a as its first element?

Since $\forall x(a, x) \notin R$, the number of relations is 2^{n^2-n}.

(d) at least one ordered pair in R has a as its first element?

By subtracting the number of relations in (c) from the total number of relations, we obtain $2^{n^2} - 2^{n^2-n}$.

Question 2 [16 points]

How many non-zero entries does the matrix representing a relation R on a set A consisting of first 100 positive integers have if R is

(a) $\{(a, b) \mid a > b\}$?

4950

(b) $\{(a, b) \mid a = b\}$?

100

(c) $\{(a, b) \mid a = 1\}$?

100

(d) $\{(a, b) \mid a$ and b have common divisors except $1\}$?

3913

Question 3 [20 points]

Let R_1 and R_2 be relations on a set $A = \{1,2,3\}$ represented by the following matrices:

$$\boldsymbol{M}_{R_1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } \boldsymbol{M}_{R_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

For each of the following relations, find the matrix that represents the relation.

(a) $R_1 \cup R_2$

$$\boldsymbol{M}_{\boldsymbol{R_1 \cup R_2}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

(b) $R_1 \cap R_2$

$$\boldsymbol{M}_{R_1 \cap R_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

(c) $R_2 \circ R_1$

$$\boldsymbol{M}_{R_2 \circ R_1} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

(d) $R_1 \circ R_1$

$$\boldsymbol{M}_{R_1 \circ R_1} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

(e) $R_1 \oplus R_2$

$$\boldsymbol{M}_{R_1 \oplus R_2} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Question 4 [12 points]

Find the probability of winning a lottery by selecting the correct six integers, where the order in which these integers are selected does not matter, from the positive integers not exceeding

(a) 30.

$$\frac{1}{\binom{30}{6}} = \frac{6!\,24!}{30!} = \frac{1}{593775}$$

(b) 36.

$$\frac{1}{\binom{36}{6}} = \frac{6!\,30!}{36!} = \frac{1}{1947792}$$

(c) 42.

$$\frac{1}{\binom{42}{6}} = \frac{6!\,36!}{42!} = \frac{1}{5245786}$$

(d) 48.

$$\frac{1}{\binom{48}{6}} = \frac{6!\,42!}{48!} = \frac{1}{12271512}$$

Question 5 [16 points]

Suppose that 8% of all bicycle racers use steroids. A bicyclist who uses steroids tests positive for steroids 96% of the time, while a bicyclist who does not use steroids tests positive for steroids 9% of the time. What is the probability that a randomly selected bicyclist who tests positive for steroids actually uses steroids?

Let X be an event that a bicyclist use steroids, and Y be an event that a bicyclist tests positive for steroids. Then, $P(X) = 0.08$, $P(Y|X) = 0.96$, and $P(Y|X^C) = 0.09$. From $P(Y|X) = \frac{P(X \cap Y)}{P(X)}$, we obtain $\mathrm{P(X \cap Y)} = 0.96 \cdot 0.08$. Similarly, $P(X^C \cap Y) = P(Y|X^C) \cdot P(X^C) = 0.09 \cdot 0.92$.

Our desired probability $P(X|Y) = \frac{\mathrm{P(X \cap Y)}}{\mathrm{P(Y)}} = \frac{P(X \cap Y)}{P(X \cap Y) + P(X^C \cap Y)} = \frac{0.96 \cdot 0.08}{0.96 \cdot 0.08 + 0.09 \cdot 0.92} \approx 0.4812$.

Question 6 [20 points]

Suppose that a Bayesian spam filter is trained on a set of 500 spam messages and 200 normal messages. The word "opportunity" appears in 40 spam messages and 25 messages which are not spam. Would an incoming message be rejected as spam if it contains the word "opportunity"? Assume that the threshold for rejecting a message is 0.9.

Let S be an event that an email is a spam, and O be an event that an email contains the word "opportunity". Then, $P(O|S) = \frac{40}{500} = 0.08$ and $P(O|S^C) = \frac{25}{200} = 0.125$. By assuming $P(S) = P(S^C) = 0.5$, we obtain $P(S|O) = \frac{P(O|S)}{P(O|S)+P(O|S^C)} = \frac{0.08}{0.08+0.125} \approx 0.39$.

If we assume that $P(S) = \frac{500}{200+500}$ and $P(S^C) = \frac{200}{200+500}$, $P(S \cap O) = P(O|S) \cdot P(S) = \frac{2}{35}$, and $P(S^C \cap O) = P(O|S^C) \cdot P(S^C) = \frac{1}{28}$. Therefore, $P(S|O) = \frac{P(S \cap O)}{P(S \cap O)+P(S^C \cap O)} = \frac{8}{13} \approx 0.6153$.

Since $P(S|O) < 0.9$, the incoming message containing "opportunity" will NOT be rejected as spam.