Knowledge Discovery & Data Mining

HW1 - Probability

CS513 - C

Muhammad Owais Imran
20025554

Homework 1.1:

Jerry and Susan have a joint bank account. Jerry goes to the bank 20% of the days. Susan goes there 30% of the days. Together they are at the bank 8% of the days.

a. Susan was at the bank last Monday. What is the probability that Jerry was there too?
b. Last Friday, Susan was not at the bank. What is the probability that Jerry was there?
c. Last Wednesday at least one of them was at the bank. What is the probability that both were there?

Solution:

$$P(Jerry) = 20\% \sim 0.2$$

$$P(Susan) = 30\% \sim 0.3$$

$$P(Jerry \cap Susan) = 8\% \sim 0.08$$

a. $P(Jerry|Susan) = \frac{P(Jerry \cap Susan)}{P(Susan)} => \frac{0.08}{0.30} = 0.267 \sim \underline{26.7\%}$
b. $P(Jerry|\overline{Susan}) = \frac{P(Jerry \cap \overline{Susan})}{P(\overline{Susan})} = \frac{P(Jerry \cap \overline{Susan})}{1-P(Susan)} = \frac{P(Jerry)-P(Jerry \cap Susan)}{1-P(Susan)} = \frac{0.2-0.08}{1-0.3} = 0.171 \sim \underline{17\%}$
c. $P(Jerry \cup Susan|\, Jerry \cap Susan) = \frac{P(Jerry \cap Susan)}{P(Jerry)+P(Susan)-P(Jerry \cap Susan)} = \frac{0.08}{0.2+0.3-0.08} = 0.190 \sim \underline{19\%}$

Homework 1.2:

Harold and Sharon are studying for a test. Harold chances of getting a “B” are 80%. Sharon chances of getting a “B” are 90%. The probability of at least one of them getting a “B” is 91%.

a. What is the probability that only Harold gets a “B”?
b. What is the probability that only Sharon gets a “B”?
c. What is the probability that both won’t get a “B”?

Solution:

$$P(Harold) = 90\% \sim 0.9$$

$$P(Sharon) = 80\% \sim 0.8$$

$$P(Harold \cup Sharon\ ) = 91\% \sim 0.91$$

$$P(Harold \cap Sharon) = P(Harold) + P(Sharon) - P(Harold \cup Sharon) = 0.79$$

a. $P(only\ Harold) = P(Sharon) - \ P(Harold \cap Sharon) = 0.8 - 0.79 = 0.01 \sim \underline{1\%}$
b. $P(only\ Sharold) = P(Harold) - P(Sharon \cap Sharon) = 0.9 - 0.79 = 0.11 \sim \underline{11\%}$
c. $P\left(\overline{(Harold \cap Sharon)}\right) = 1 - P(Harold \cap Sharon) = 1 - 0.79 = 0.21 \sim \underline{21\%}$

Homework 1.3:

Jerry and Susan have a joint bank account. Jerry goes to the bank 20% of the days. Susan goes there 30% of the days. Together they are at the bank 8% of the days. Are the events "Jerry is at the bank" and "Susan is at the bank" independent?

Solution:

For an event to be independent, the joint probability of both events should be equal to the independent probabilities of events to occur i.e.

$$P(Sharon \cap Harold) = P(Harold) * P(Sharon)$$

But, according to the probabilities given in the question i.e.

$$P(Jerry) = 20\% \sim 0.2$$

$$P(Susan) = 30\% \sim 0.3$$

$$P(Jerry \cap Susan) = 8\% \sim 0.08$$

$$P(Sharon \cap Harold) \neq P(Sharon) * P(Harold)$$
$$0.08 \neq (0.3 * 0.2)$$

Therefore, the events are not independent.

Homework 1.4:

You roll 2 dice.

a. Are the events "the sum is 6" and "the second die shows 5" independent.
b. Are the events "the sum is 7" and "the first die shows 5" independent.

Solution:

a. The outcome table for 2 dice.

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	**6**	7
2	3	4	5	**6**	**7**	8
3	4	5	**6**	7	**8**	9
4	5	**6**	7	8	**9**	10
5	**6**	7	8	9	**10**	11
6	7	8	9	10	**11**	12

$$P(Sum = 6) = \frac{5}{36} = 0.139 \sim \underline{13.9\%}$$
$$P(Second\ Die = 5) = \frac{6}{36} = 0.1666 \sim \underline{16.7\%}$$
$$P(SecondDie = 5 \cap Sum = 6) = \frac{1}{36} = 0.0277 \sim \underline{2.77\%}$$
$$P(SecondDie = 5) * P(Sum = 6) = 0.139 * 0.1666 = 0.0231 \sim \underline{2.31\%}$$

Since $P(SecondDie = 5 \cap Sum = 6) \neq P(SecondDie = 5) * P(Sum = 6)$ therefore, <u>the events are not independent.</u>

b. The outcome table for 2 dice.

	1	2	3	4	5	6
1	2	3	4	5	6	**7**
2	3	4	5	6	**7**	8
3	4	5	6	**7**	8	9
4	5	6	**7**	8	9	10
5	**6**	**7**	**8**	**9**	**10**	**11**
6	**7**	8	9	10	11	12

$$P(Sum = 7) = \frac{6}{36} = 0.1666 \sim \underline{16.7\%}$$
$$P(FirstDie = 5) = \frac{6}{36} = 0.1666 \sim \underline{16.7\%}$$
$$P(FirstDie = 5 \cap Sum = 7) = \frac{1}{36} = 0.0277 \sim \underline{2.77\%}$$
$$P(FirstDie = 5) * P(Sum = 7) = 0.1666 * 0.1666 = 0.0277 \sim \underline{2.77\%}$$

Since $P(FirstDie = 5) * P(Sum = 7) = P(FirstDie = 5 \cap Sum = 7)$, therefore the events are independent.

Homework 1.5

An oil company is considering drilling in either TX, AK or NJ. The company may operate in only one state. There is a 60% chance the company will choose TX and a 10% chance – NJ.
There is a 30% chance of finding oil in TX, 20% - in AK, and 10% - in NJ.

1. What is the probability of finding oil?
2. The company decided to drill and found oil. What is the probability that they drilled in TX?

Solution:

$$P(Choosing\ TX) = 60\%\ or\ 0.6$$
$$P(Choosing\ AK) = 30\%\ or\ 0.3$$
$$P(Choosing\ NJ) = 10\%\ or\ 0.1$$

1. $P(Oil) = P(Oil \cap TX) + P(Oil \cap AK) + P(Oil \cap NJ)$
   $\therefore P(Oil \cap State) = P(Oil|State) * P(Choosing\ State)$
   $P(Oil) = (0.6 * 0.3) + (0.3 * 0.2) + (0.1 * 0.1)$
   $P(Oil) = 0.18 + 0.06 + 0.01$
   $P(Oil) = 0.25 \sim \underline{25\%}$

2. $P(TX|Oil) = \frac{P(TX \cap Oil)}{P(Oil)}$
   $P(TX|Oil) = \dfrac{0.3 * 0.6}{0.25}$
   $P(TX|Oil) = 0.72 \sim \underline{72\%}$

<u>Homework 1.6:</u>

The following slide shows the survival status of individual passengers on the Titanic. Use this information to answer the following questions:

Solutions:

1. What is the probability that a passenger did not survive?

$$\boldsymbol{P(NotSurvivedPassengers) = 1 - \frac{711}{2201} = 1 - 0.323 = 0.677 \sim \underline{67.7\%}}$$

2. What is the probability that a passenger was staying in the first class?

$$\boldsymbol{P(FirstClassPassenger) = \frac{325}{2201} = 0.1476 \sim \underline{14.76\%}}$$

3. Given that a passenger survived, what is the probability that the passenger was staying in the first class?

$$\boldsymbol{P(FirstClass|Survived) = \frac{P(FirstClass \cap Survived)}{P(Survived)}}$$
$$\boldsymbol{P(FirstClass|Survived) = \frac{203}{711} = 0.2855 \sim \underline{28.55\%}}$$

4. Are survival and staying in the first class independent?

$$\boldsymbol{P(FirstClass) = 14.76\% \text{ or } 0.1476}$$
$$\boldsymbol{P(Survived) = 32.3\% \text{ or } 0.323}$$
$$\boldsymbol{P(FirstClass) * P(Survived) = 0.1476 * 0.323 = 0.0476}$$
$$\boldsymbol{P(FirstClass \cap Survived) = \frac{203}{711} = 0.2855 \sim \underline{28.55\%}}$$
$$\boldsymbol{Events\ are\ not\ independent}$$

5. Given that a passenger survived, what is the probability that the passenger was staying in the first class and the passenger was a child?

$$\boldsymbol{P(FirstClass \cap Child\ |Survived) = \frac{P(FirstClass \cap Child \cap Survived)}{P(Survived)}}$$
$$\boldsymbol{P(FirstClass \cap Child\ |Survived) = \frac{6}{711} = 0.008 \sim \underline{0.8\%}}$$

6. Given that a passenger survived, what is the probability that the passenger was an adult?

$$\boldsymbol{P(Adult|Survived) = \frac{P(Adult \cap Survived)}{P(Survived)} = \frac{442}{711} = 0.6216 \sim \underline{62.16\%}}$$

7. Given that a passenger survived, are age and staying in the first class independent?

$$P(Age\ |Survived) = \frac{P(Age \cap Survived)}{P(Survived)} = \frac{711}{711} = 1$$

Given passenger survived, probability of age staying in first class = 40.68%
Probability of age and staying in first class = 40.68%
Since product of <u>Given passenger survived, probability of age staying in first class</u> and <u>probability of age and staying in first class</u> is equal than, events are independent.

Homework 1.7:

A developer claims that her app can distinguish AI-generated documents from human-generated ones. To assess its performance, we have submitted 1000 AI-generated and 1000 human-generated documents to the app.

• The app misclassified 70 human-generated documents as AI-generated
• and 30 AI generated documents as human- generated.

Build the confusion matrix for the above app and calculate the following: Accuracy, precision, recall and F1.

Solution:

Confusion matrix:

	Actual AI Generated	**Actual Human Generated**	**Total**
Predicted as AI Generated	**970**	**70**	**1040**
Predicted as Human Generated	**30**	**930**	**960**
	1000	**1000**	**2000**

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{970 + 930}{970 + 70 + 30 + 930} = \frac{1900}{2000} = 0.95 \sim 95\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{970}{970 + 70} = \frac{970}{1040} = 0.9326 \sim 93.26\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{970}{970 + 30} = \frac{970}{1000} = 0.97 \sim 97\%$$

$$F1 = \frac{2 * Precision * Recall}{2 + Precision + Recall} = \frac{2 * 0.9326 * 0.97}{0.9326 + 0.97} = \frac{1.8092}{1.9026} = 0.9509 \sim 95.09\%$$