CS 513: Knowledge Discovery And Data Mining

Ali Abdullah Ahmad

February 18, 2025

Q1.1 Let:

- $P(J) = 0.2$ be the probability that Jerry goes to the bank on a given day.
- $P(S) = 0.3$ be the probability that Susan goes to the bank on a given day.
- $P(J \cap S) = 0.08$ be the probability that both go to the bank on the same day.

Using the formula for the union of two events:

$$P(J \cup S) = P(J) + P(S) - P(J \cap S)$$

$$P(J \cup S) = 0.2 + 0.3 - 0.08 = 0.42$$

(a) Given that Susan was at the bank last Monday, what's the probability that Jerry was there too?

We need to find $P(J|S)$, which is given by:

$$P(J|S) = \frac{P(J \cap S)}{P(S)}$$

$$P(J|S) = \frac{0.08}{0.3} = \frac{8}{30} = \frac{4}{15} \approx 0.267$$

(b) Given that Susan wasn't at the bank last Friday, what's the probability that Jerry was there?

We need to find $P(J|S')$, where S' denotes that Susan was not at the bank. This is given by:

$$P(J|S') = \frac{P(J \cap S')}{P(S')}$$

First, calculate $P(J \cap S')$:

$$P(J \cap S') = P(J) - P(J \cap S) = 0.2 - 0.08 = 0.12$$

Since $P(S') = 1 - P(S) = 1 - 0.3 = 0.7$, we get:

$$P(J|S') = \frac{0.12}{0.7} = \frac{12}{70} = \frac{6}{35} \approx 0.171$$

(c) Given that at least one of them was at the bank last Wednesday, what is the probability that both of them were there?

We need to find $P(J \cap S | J \cup S)$, which is given by:

$$P(J \cap S | J \cup S) = \frac{P(J \cap S)}{P(J \cup S)}$$

$$P(J \cap S | J \cup S) = \frac{0.08}{0.42} = \frac{8}{42} = \frac{4}{21} \approx 0.190$$

Q1.2 Let:

- $P(H) = 0.8$ be the probability that Harold gets a "B".
- $P(S) = 0.9$ be the probability that Sharon gets a "B".
- $P(H \cup S) = 0.91$ be the probability that at least one of them gets a "B".

Using the formula for the union of two events:

$$P(H \cup S) = P(H) + P(S) - P(H \cap S)$$

$$0.91 = 0.8 + 0.9 - P(H \cap S)$$

$$P(H \cap S) = 1.7 - 0.91 = 0.79$$

(a) Probability that only Harold gets a "B":
We need to find $P(H \cap S')$:

$$P(H \cap S') = P(H) - P(H \cap S)$$

$$P(H \cap S') = 0.8 - 0.79 = 0.01$$

(b) Probability that only Sharon gets a "B":
We need to find $P(S \cap H')$:

$$P(S \cap H') = P(S) - P(H \cap S)$$

$$P(S \cap H') = 0.9 - 0.79 = 0.11$$

(c) Probability that neither Harold nor Sharon gets a "B":
We need to find $P(H' \cap S')$:

$$P(H' \cap S') = 1 - P(H \cup S)$$

$$P(H' \cap S') = 1 - 0.91 = 0.09$$

Thus, the final probabilities are:

- (a) 0.01
- (b) 0.11
- (c) 0.09

Q1.3 **Are the events "Jerry is at the bank" and "Susan is at the bank" independent?**

To determine independence, we check whether:

$$P(J \cap S) = P(J)P(S)$$

Given:

- $P(J) = 0.2$ (Probability that Jerry goes to the bank)
- $P(S) = 0.3$ (Probability that Susan goes to the bank)
- $P(J \cap S) = 0.08$ (Probability that both are at the bank together)

Step 1: Compute $P(J)P(S)$

$$P(J)P(S) = (0.2)(0.3) = 0.06$$

Step 2: Compare with $P(J \cap S)$

$$P(J \cap S) = 0.08 \neq 0.06 = P(J)P(S)$$

Conclusion: Since $P(J \cap S) \neq P(J)P(S)$, the events **Jerry is at the bank** and **Susan is at the bank** are **not independent**.

Q1.4(a) Are the events "the sum is 6" and "the second die shows 5" independent?

To determine independence, we check whether:

$$P(A \cap B) = P(A) \cdot P(B)$$

where:

- A is the event "the sum is 6".
- B is the event "the second die shows 5".

Step 1: Compute $P(A)$
The possible outcomes that give a sum of 6 are:

$$(1,5),(2,4),(3,3),(4,2),(5,1)$$

There are 5 such outcomes out of 36 possible outcomes, so:

$$P(A) = \frac{5}{36}$$

Step 2: Compute $P(B)$
The probability that the second die shows 5 is:

$$P(B) = \frac{1}{6}$$

Step 3: Compute $P(A \cap B)$
The outcome where both events occur is $(1,5)$, so:

$$P(A \cap B) = \frac{1}{36}$$

Step 4: Compare $P(A \cap B)$ with $P(A) \cdot P(B)$

$$P(A) \cdot P(B) = \left(\frac{5}{36}\right)\left(\frac{1}{6}\right) = \frac{5}{216}$$

Since:

$$P(A \cap B) = \frac{1}{36} \neq \frac{5}{216} = P(A) \cdot P(B)$$

the events are not independent

.

(b) Are the events "the sum is 7" and "the first die shows 5" independent?

We check whether:

$$P(C \cap D) = P(C) \cdot P(D)$$

where:

- C is the event "the sum is 7".
- D is the event "the first die shows 5".

Step 1: Compute $P(C)$
The possible outcomes that give a sum of 7 are:

$$(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$$

There are 6 such outcomes out of 36 possible outcomes, so:

$$P(C) = \frac{6}{36} = \frac{1}{6}$$

Step 2: Compute $P(D)$
The probability that the first die shows 5 is:

$$P(D) = \frac{1}{6}$$

Step 3: Compute $P(C \cap D)$
The outcome where both events occur is $(5,2)$, so:

$$P(C \cap D) = \frac{1}{36}$$

Step 4: Compare $P(C \cap D)$ with $P(C) \cdot P(D)$

$$P(C) \cdot P(D) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$$

Since:

$$P(C \cap D) = P(C) \cdot P(D)$$

the events are independent.

Q1.5

1. Probability of finding oil

We are given the following probabilities:

- Probability of choosing TX: $P(TX) = 0.6$
- Probability of choosing AK: $P(AK) = 0.3$ (since $1 - 0.6 - 0.1 = 0.3$)
- Probability of choosing NJ: $P(NJ) = 0.1$
- Probability of finding oil in TX: $P(Oil \mid TX) = 0.3$
- Probability of finding oil in AK: $P(Oil \mid AK) = 0.2$
- Probability of finding oil in NJ: $P(Oil \mid NJ) = 0.1$

The **total probability of finding oil** can be calculated using the law of total probability:

$$P(Oil) = P(TX) \cdot P(Oil \mid TX) + P(AK) \cdot P(Oil \mid AK) + P(NJ) \cdot P(Oil \mid NJ)$$

Substituting the given values:

$$P(Oil) = (0.6 \times 0.3) + (0.3 \times 0.2) + (0.1 \times 0.1)$$

$$P(Oil) = 0.18 + 0.06 + 0.01 = 0.25$$

So, the probability of finding oil is 0.25 **or 25%.**

2. Probability that they drilled in TX given they found oil

This is a conditional probability problem. We need to find $P(TX \mid Oil)$, the probability that they drilled in TX given that they found oil. We can use Bayes' Theorem:

$$P(TX \mid Oil) = \frac{P(Oil \mid TX) \cdot P(TX)}{P(Oil)}$$

From the previous calculations: - $P(Oil \mid TX) = 0.3$ - $P(TX) = 0.6$ - $P(Oil) = 0.25$

Substituting these values into Bayes' Theorem:

$$P(TX \mid Oil) = \frac{(0.3) \cdot (0.6)}{0.25} = \frac{0.18}{0.25} = 0.72$$

So, the probability that they drilled in TX given that they found oil is 0.72 **or 72%.**

Q1.6 Given Data

- Total passengers: $2,201$
- Total survived: 711
- Total not survived: $1,490$
- First-class passengers (total): 325
- First-class passengers (survived): 203
- First-class child survivors: 6
- First-class adult survivors: 197

Probability Calculations

1. Probability that a passenger did not survive

$$P(NotSurvived) = \frac{TotalNotSurvived}{TotalPassengers} = \frac{1,490}{2,201} \approx 0.677$$

2. Probability that a passenger was staying in first class

$$P(FirstClass) = \frac{TotalFirstClass}{TotalPassengers} = \frac{325}{2,201} \approx 0.148$$

3. Given that a passenger survived, probability that they were in first class

$$P(FirstClass \mid Survived) = \frac{FirstClassSurvived}{TotalSurvived} = \frac{203}{711} \approx 0.285$$

4. Are survival and staying in first class independent?

If survival and first-class status were independent, then:

$$P(FirstClass \cap Survived) = P(FirstClass) \times P(Survived)$$

From earlier,

$$P(FirstClass) \approx 0.148, \quad P(Survived) \approx 0.323$$

$$P(FirstClass \cap Survived)_{expected} = 0.148 \times 0.323 = 0.048$$

$$P(FirstClass \cap Survived)_{actual} = \frac{203}{2,201} \approx 0.092$$

Since $0.092 \neq 0.048$, survival and first-class status are **not independent**.

5. Given that a passenger survived, probability that they were a first-class child

$$P(FirstClassChild \mid Survived) = \frac{FirstClassChildSurvived}{TotalSurvived} = \frac{6}{711} \approx 0.0084$$

6. Given that a passenger survived, probability that they were an adult

$$P(Adult \mid Survived) = \frac{TotalSurvivedAdults}{TotalSurvived} = \frac{654}{711} \approx 0.920$$

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

If age and first-class status were independent given survival, then:

$$P(FirstClass \cap Adult \mid Survived) = P(FirstClass \mid Survived) \times P(Adult \mid Survived)$$

From previous calculations:

$$P(FirstClass \mid Survived) \approx 0.285, \quad P(Adult \mid Survived) \approx 0.920$$

$$P(FirstClass \cap Adult \mid Survived)_{expected} = 0.285 \times 0.920 = 0.262$$

$$P(FirstClassAdult \mid Survived)_{actual} = \frac{197}{711} \approx 0.277$$

Since $0.277 \neq 0.262$, age and first-class status are **not independent** given survival.

Q1.7 Confusion Matrix

The classification results are as follows:

- False Positives (FP): 70 human-generated documents misclassified as AI-generated.
- False Negatives (FN): 30 AI-generated documents misclassified as human-generated.
- Total AI-generated documents (Actual Positives): 1000.
- Total human-generated documents (Actual Negatives): 1000.
- True Positives (TP): Correctly classified AI-generated documents: $1000 - 30 = 970$.
- True Negatives (TN): Correctly classified human-generated documents: $1000 - 70 = 930$.

Thus, the confusion matrix is:

	$PredictedAI$	$PredictedHuman$
$ActualAI$	$TP = 970$	$FN = 30$
$ActualHuman$	$FP = 70$	$TN = 930$

Performance Metrics Calculation

1. Accuracy

Accuracy is the proportion of correctly classified documents:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{970 + 930}{970 + 930 + 70 + 30} = \frac{1900}{2000} = 0.95 \quad (95\%)$$

2. Precision (for AI-generated classification)

Precision measures how many of the predicted AI-generated documents were actually AI-generated:

$$Precision = \frac{TP}{TP + FP}$$

$$= \frac{970}{970 + 70} = \frac{970}{1040} \approx 0.933$$

3. Recall (for AI-generated classification)

Recall (or sensitivity) measures how many actual AI-generated documents were correctly classified:

$$Recall = \frac{TP}{TP + FN}$$

$$= \frac{970}{970 + 30} = \frac{970}{1000} = 0.97$$

4. F1 Score

The F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= 2 \times \frac{0.933 \times 0.97}{0.933 + 0.97}$$

$$= 2 \times \frac{0.905}{1.903} \approx 0.951$$

Final Results

- **Accuracy** = 95%
- **Precision** = 93.3%
- **Recall** = 97%
- **F1 Score** = 95.1%

These metrics indicate that the app performs well, with high accuracy and recall, though there is a slight trade-off in precision due to false positives.