Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

Checking Independence

Two random variables X and Y are independent if:

$P(X = x, Y = y) = P(X = x) \times P(Y = y)$

(for discrete variables)

$f(x, y) = f(x)_X \times f(y)_Y$

(for continuous variables)

Introduction to Machine Learning

- **Machine Learning (ML)** enables **computers to learn from data.**

- ML makes **predictions or decisions** without being **explicitly programmed.**

- Types of Machine Learning: **Supervised, Unsupervised, Semi-supervised, Reinforcement, Self-supervised, Transfer Learning**.

Types of Machine Learning

1. **Supervised Learning:** Learns from labeled data.
2. **Unsupervised Learning:** Finds patterns in unlabeled data.
3. **Semi-Supervised Learning:** Combines labeled and unlabeled data.
4. **Reinforcement Learning:** Learns by interacting with the environment.
5. **Self-Supervised Learning:** Generates its own labels from data.
6. **Transfer Learning:** Applies knowledge from one task to another.

Supervised Learning

- **Goal:** Learn a mapping from **input** to **output** based on **labeled data**.

- **Common Algorithms: Linear Regression, Decision Trees, SVM, Neural Networks**.

Example: Predicting house prices, email spam detection.

Unsupervised Learning

Goal: Find **hidden patterns** in **unlabeled data**.

Common Algorithms: K-Means Clustering, Hierarchical Clustering, PCA.

Example: Customer segmentation, anomaly detection.

Clustering (e.g., K-Means, Hierarchical Clustering): This involves grouping data points into clusters based on similarity. It's often used in market segmentation, social network analysis, or even image compression.

Semi-Supervised Learning

Goal: Use a **small amount of labeled data** with a **large amount of unlabeled data**. This approach helps improve learning when labeling data is expensive or time-consuming.

Example: Medical Imaging

Labeled medical images, such as **X-rays, MRI scans, or CT scans**, often require expert annotation, which is time-consuming and expensive.

However, a large amount of **unlabeled medical image data may be available.** A semi-supervised learning model can learn to classify images (e.g., diagnosing diseases) using a combination of **expert-labeled images and many unlabeled images**.

Reinforcement Learning

Goal: Reinforcement learning (RL) involves training **an agent** to make a **sequence of decisions by interacting with an environment**, receiving **rewards or penalties** as **feedback**, and learning the best strategies (called policies) over time.

Common Algorithms: Q-Learning, Deep Q-Networks, Policy Gradients.

Example: Game playing (AlphaGo), robotics.

Reinforcement Learning

Recommendation Systems:

Example: Online Recommendations (Netflix, YouTube)

Reinforcement learning is used in recommendation systems (**e.g., Netflix or YouTube**) to suggest content to users.

The system receives **feedback** in the form of **user interactions (e.g., clicks, likes, or time spent watching),** which it uses as **rewards to learn what types of content are most engaging** to the user, improving its future recommendations.

Self-Supervised Learning

Goal: Predict parts of data to **generate labels.**

Self-supervised learning (SSL) is a paradigm of machine learning where the **model learns to predict part** of the **input data using the other parts**, effectively creating its own supervisory signal.

Common Applications: Language models, image inpainting.

Example: Predicting missing parts of images.

Transfer Learning

Goal: Use knowledge from one task to improve another task.

Example: Using a pre-trained model for classifying medical images.

- **Image Classification**
- **Example: Using Pre-trained CNNs for Medical Image Classification** A model trained on a large dataset like **ImageNet** (which contains millions of labeled general images) can be fine-tuned to classify medical images such as X-rays or MRI scans, even though medical images differ significantly from the original dataset.

Finite Stochastic Processes And Tree Diagrams

- A (finite) sequence of experiments in which each experiment has a finite number of outcomes with given probabilities is called a **(finite) stochastic process.**

- A convenient way of describing such a process and computing the probability of any event is by a **tree diagram.**

Example:

We are given three boxes as follows:

Box 1 has **10** light bulbs of which **4** axe defective.

Box 2 has **6** light bulbs of which **1** is defective.

Box 3 has **8** light bulbs of which **3** are defective.

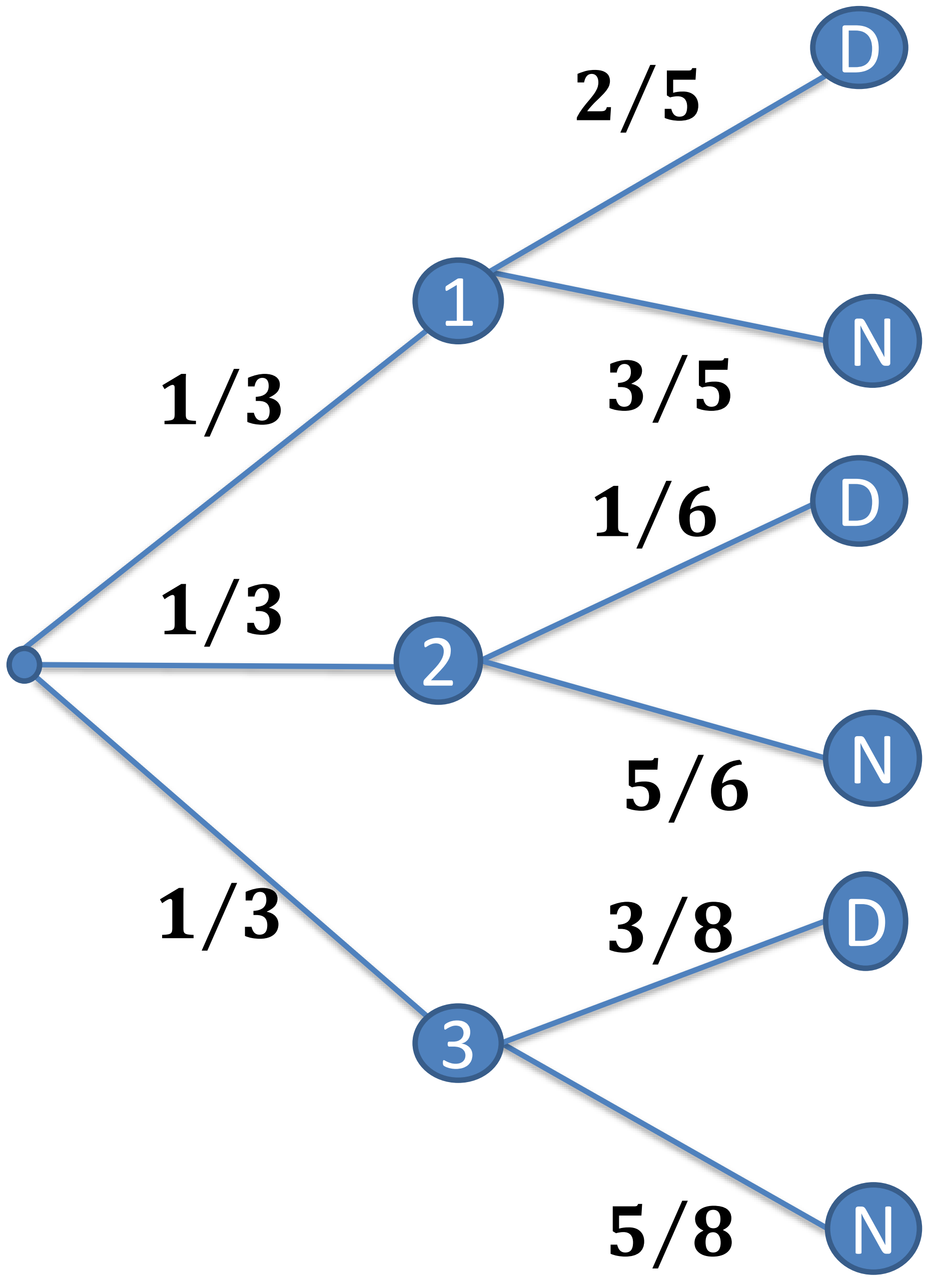
We select a box at random and then draw a bulb at random. What is the probability **p** that the **bulb** is **defective**?

Solution:

Here we perform a sequence of two experiments:

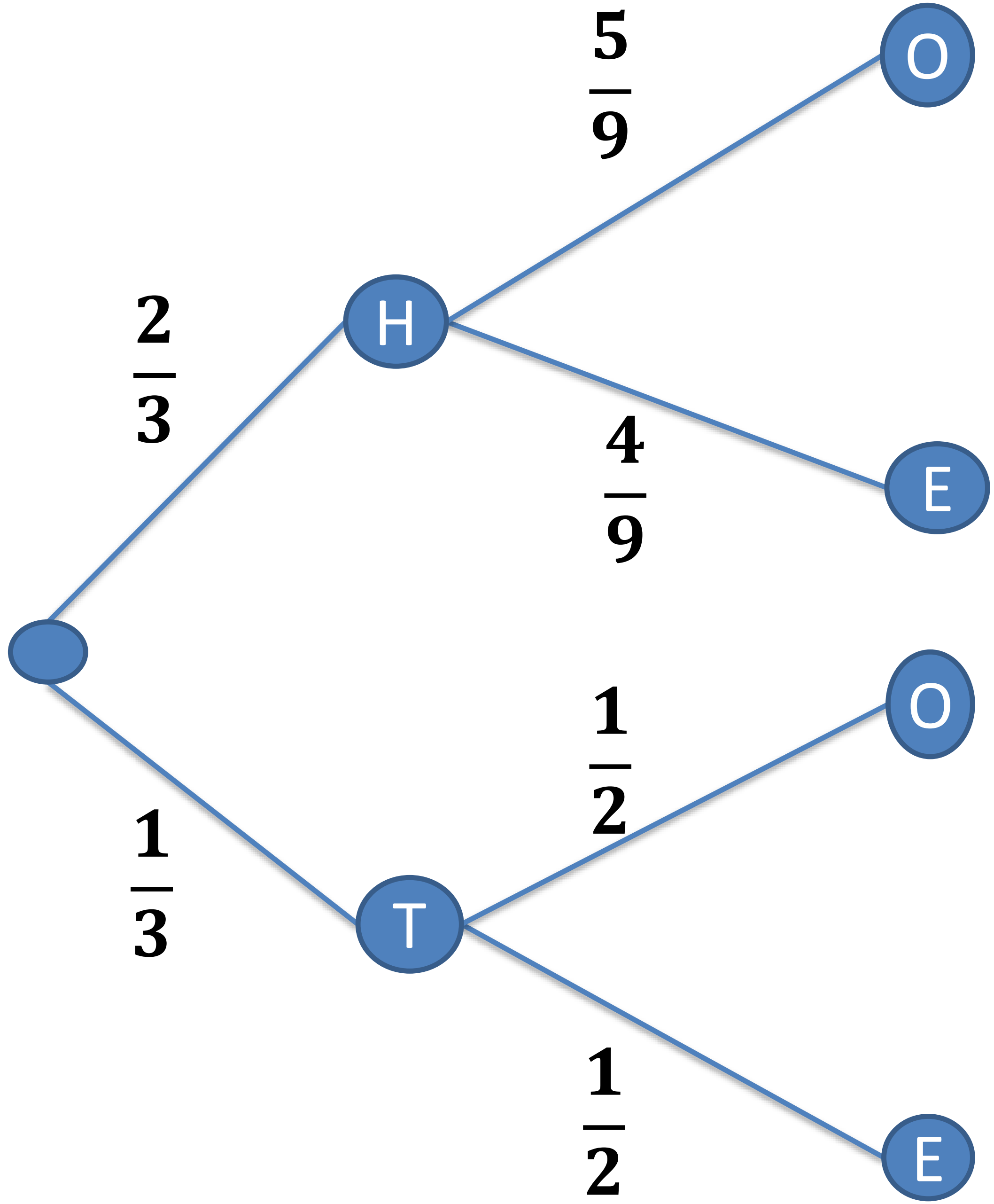
(i) select one of the three boxes;

(ii) select a bulb which is either defective ***(D)*** or nondefective ***(N).***


D
2/5
1
N
1/3
3/5
D
1/6
1/3
2
N
5/6
1/3
D
3/8
3
N
5/8

Example : A coin, weighted so that **P(H) =** $\frac{2}{3}$ and **P(T) =** $\frac{1}{3}$, is tossed. If **heads** appears, then a number is selected at random from the numbers **1 through 9**; if **tails** appears, then a number is selected at random from the numbers **1 through 6.**

Find the probability p that an **even number** is selected.



$\frac{2}{3}$
H
$\frac{5}{9}$
O
$\frac{4}{9}$
E
$\frac{1}{3}$
T
$\frac{1}{2}$
O
$\frac{1}{2}$
E

Probability of even using 1 through 9:

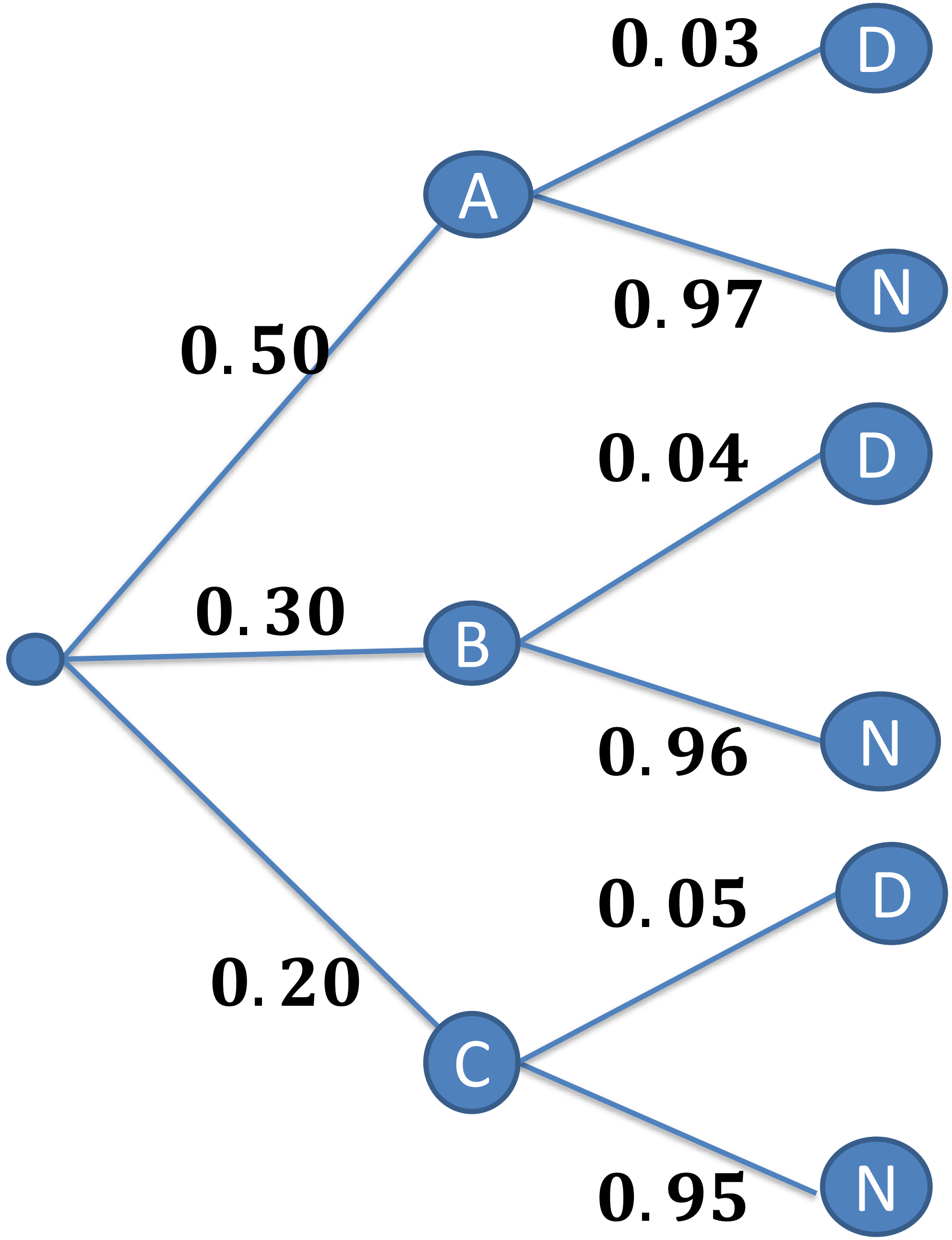
$$P(E) = \frac{4}{9}$$

Probability of even using 1 through 6:

$$P(E) = \frac{3}{6} = \frac{1}{2}$$

$$\mathbf{p} = \left(\frac{2}{3}\right)\left(\frac{4}{9}\right)+\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \mathbf{\frac{25}{54}}$$

Example: Three machines ***A, B*** and ***C*** produce respectively **50%**, **30%** and **20%** of the total number of items of a factory. The percentages of **defective output** of these machines are **3%**, **4%** and **5%.** If an item is selected at random, find the probability that the item is **defective**.


0.50
0.30
0.20
A
B
C
0.03
0.97
0.04
0.96
0.05
0.95
D
N
D
N
D
N

Solution:

Let **D** be the event that the item is **defective**

$\because$ *P(B)* = $P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \ldots + P(A_n)P(B|A_n)$

$\therefore$ ***P(D)*** = $\mathbf{P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C)}$

$= (0.50)(0.03) + (0.30)(0.04) + (0.20)(0.05)$

$=$ **0.037** **(or 3.7%)**

Example: Consider the factory in the preceding example. Suppose an item is selected at random and is found to be **defective**. Find the probability that the item was produced by **machine A**; that is, find **$P(A|D)$.**

Solution:

By Bayes' theorem,

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_j)}$$

$$P(A|D) = \frac{P(A)P(D|A)}{P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C)}$$

$$P(A|D) = \frac{(0.50)(0.03)}{(0.50)(0.03) + (0.30)(0.04) + (0.20)(0.05)}$$

$$= \frac{15}{37} \text{ (or 0.4054)}$$

What is Bayes' Law?

Formula: **$P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$**

Bayes' Law is a formula for determining **conditional probabilities**.

Components:

$P(A|B)$: Posterior probability

$P(B|A)$: Likelihood

$P(A)$: Prior probability

$P(B)$: Marginal likelihood

Joint Probability Distribution

Consider the following **random variables** used in a **machine learning** context:

- X: **Weather feature** (**X = 0: Rainy**, **X = 1: Sunny**)

- Y: **Target label** (**Y = 0: No Tennis**, **Y = 1: Plays Tennis**)

You are given the following joint probability table:

	Y = No Tennis (0)	Y = Plays Tennis (1)
X = Rainy (0)	**0.3**	**0.2**
X = Sunny (1)	**0.1**	**0.4**

Joint Probability Distribution

Questions:

1. (a) What is the probability that it is **"Sunny"** and **"Plays Tennis"**?

2. (b) What is the marginal probability of **"Playing Tennis"**?

3. (c) Given that it is **"Rainy,"** what is the probability of **"No Tennis"**?

4. (d) Are the events **"Weather"** and **"Playing Tennis"** independent? Justify your answer.

X = 1: Sunny, Y = 1: Plays Tennis

1(a) Solution: The probability that it is **"Sunny"** and **"Plays Tennis"** is:

$P(X = 1, Y = 1) = 0.4$

Marginal Probability

Y = 1: Plays Tennis

1 (b)Solution The **marginal probability** of **"Playing Tennis"** is calculated as:

Marginal Probability of Y:

$\mathbf{P(Y = 0)} = P(X = 0, \mathbf{Y = 0}) + P(X = 1, \mathbf{Y = 0})$

$= 0.1 + 0.3 = 0.4$

$\mathbf{P(Y = 1)} = P(X = 0, \mathbf{Y = 1}) + P(X = 1, \mathbf{Y = 1})$

$= 0.2 + 0.4 = 0.6$

Marginal Probability of "Plays Tennis"

y = 0	0	1	Total
h(y)	0.40	0.60	1

Y = 0: No Tennis, X = 0: Rainy,

1 (c) Conditional Probability of **"No Tennis"** given **"Rainy“**

The conditional probability is:

$$\mathbf{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

$$P(Y = 0 \mid X = 0) = \frac{P(X = 0, Y = 0)}{P(X = 0)}$$

$$P(Y = 0 \mid X = 0) = \frac{0.3}{0.5} = 0.6$$

Joint Probability Distribution

Consider the following **random variables** used in a **machine learning** context:

- X: **Weather feature** (**X = 0: Rainy**, **X = 1: Sunny**)

- Y: **Target label** (**Y = 0: No Tennis**, **Y = 1: Plays Tennis**)

You are given the following joint probability table:

	Y = No Tennis (0)	Y = Plays Tennis (1)
X = Rainy (0)	**0.3**	**0.2**
X = Sunny (1)	**0.1**	**0.4**

1 (d)Solution: Independence of **"Weather"** and **"Playing Tennis"**

Events **X (Weather)** and **Y (Playing Tennis)** are independent if:

$$P(X, Y) = P(X) \times P(Y)$$

For $P(X = 0, Y = 0)$:

$P(X = 0, Y = 0) = 0.3$

$P(X = 0) \times P(Y = 0) = 0.5 \times 0.4$

$= 0.2$

$\Rightarrow P(X, Y) \neq P(X) \times P(Y)$

Since **$0.3 \neq 0.2$**, so the events are **not independent**.

Classifying Emails: Spam Detection

Given the presence of words **"offer"** and **"win"** :

P(spam) = 0.4, **P(not spam)** = 0.6

P(offer | spam) = 0.7, **P(offer | not spam)** = 0.2

P(win | spam) = 0.8, **P(win | not spam)** = 0.1

1. What is the **probability of spam given "offer"** is present?
2. What is the **probability of spam given** both **"offer"** and **"win"** are present?

Given Data

P(spam) = 0.4, **P(not spam)** = 0.6

P(offer | spam) = 0.7, **P(offer | not spam)** = 0.2

P(win | spam) = 0.8, **P(win | not spam)** = 0.1

Probability of spam given 'offer' is present

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes' Theorem:

$$P(\text{spam} \mid \text{offer}) = \frac{P(\text{offer} \mid \text{spam}) \times P(\text{spam})}{P(\text{offer})}$$

$P(\text{offer}) = P(\text{offer} \mid \text{spam}) \times P(\text{spam}) + P(\text{offer} \mid \text{not spam}) \times P(\text{not spam})$

$P(\text{offer}) = 0.7 \times 0.4 + 0.2 \times 0.6 = 0.4$

$$P(\text{spam} \mid \text{offer}) = \frac{0.7 \times 0.4}{0.4} = 0.70$$

By Bayes' theorem,

$$\mathbf{P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i\,)P(B|A_j)}}$$

Using Bayes' Theorem:

P(**spam** | **offer** and **win**)

$$= \frac{P(\textbf{offer} \text{ and } \textbf{win} \mid \textbf{spam}) \times P(\textbf{spam})}{P(\textbf{offer} \text{ and } \textbf{win})} \quad \text{----------(1)}$$

Computation of P(**offer** and **win** | **spam**)

P(**offer** and **win** | **spam**) = P(**offer** | **spam**) $\times$ P(**win** | **spam**)

Since the two events are conditionally independent given spam, we multiply their individual probabilities:

$$= 0.7 \times 0.8 = 0.56$$

Since the two events are conditionally independent given spam, we multiply their individual probabilities:

P(**offer** and **win** | **not spam**) = P(**offer** | **not spam**) $\times$ P(**win** | **not spam**)

$$= 0.2 \times 0.1 = 0.02$$

Computation of P(**offer** and **win**) = ?

P(**offer** and **win**) = P(**offer** and **win** | **spam**) × **P(spam)** + P(**offer** and **win** | not **spam**) × **P(not spam)** $= 0.56 \times 0.40 + 0.02 \times 0.06$

$= 0.236$

Put values in equation (1), we get

P(**spam** | **offer** and **win**) $= \dfrac{0.56 \times 0.4}{0.236}$

$= 0.949$

Interpretation of the Result:

$$P(\textbf{spam} \mid \textbf{offer} \text{ and } \textbf{win}) = \frac{0.56 \times 0.4}{0.236}$$

$$= 0.949$$

It means that if an email contains both the words **"offer"** and **"win,"** there is approximately a **94.9%** chance that the **email is spam**.

This high probability indicates that these words are **strong indicators of spam**, as they are commonly used in spam emails to attract attention.