

PROBABILITY FOR DATA SCIENCE

An Internship Project Report

Submitted in partial fulfillment of the requirement for the award of the
Degree of Bachelor of Science in Mathematics

submitted by

S.THEJESSHREE

Register no. 18BM7476

Under the guidance of

(Ms.)T.Saradhadevi M.Sc., M.Phil.,

Assistant Professor

Department of Mathematics (Self Finance)



Sri G.V.G Visalakshi College for Women (Autonomous)

(Affiliated to Bharathiyar University, Coimbatore)

Accredited at 'A+' Grade by NAAC (CGPA 3.27)

An ISO 9001: 2015 Certified Institution

Udumalpet-642 128

March-2021



greatlearning
Power Ahead

CERTIFICATE OF COMPLETION

Presented to

Thejesshree.S

For successfully completing a free online course
Probability for Data Science

Provided by
Great Learning Academy
(On March 2021)



To verify this certificate visit verify.greatlearning.in/EOVQJFGG

CERTIFICATE

This is to certify that the Internship project work entitled “**PROBABILITY FOR DATA SCIENCE**” is a bonafied record work done by **S.THEJESSHREE (18BM7476)** submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Science in Mathematics at Sri G.V.G Visalakshi college for Women (Autonomous), Udumalpet during the academic year 2018-2021.

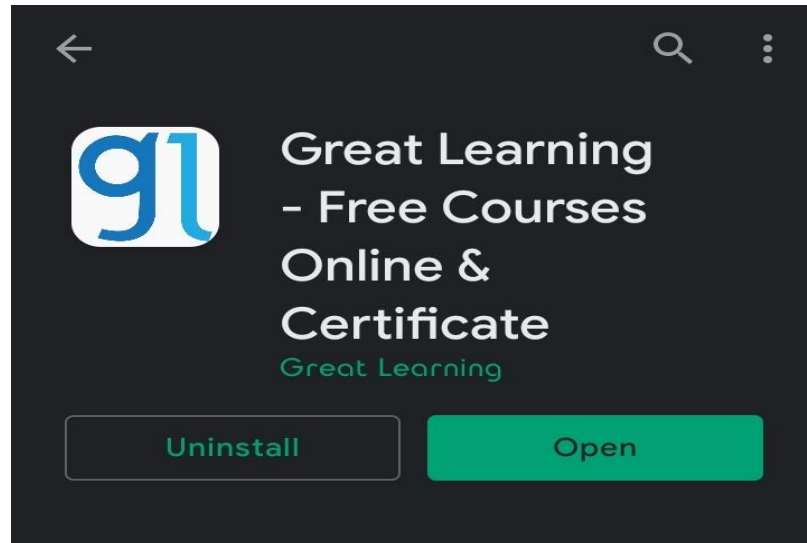
Signature of the staff-in-charge

Signature of the HOD

Introduction:

The internship training program was organized in Great Learning Application launched by Mohan lakhamraju in 2013. This is the No. 1 ranked online classroom courses on Artificial Intelligence, Machine Learning, Data Science Engineering and deep learning for college students professionally. The trainer Dr. P.K.Viswanathan, Professor of probability from Great Lakes institute of Management taught us about “Probability of Data Sciences”.

Digital Tool:

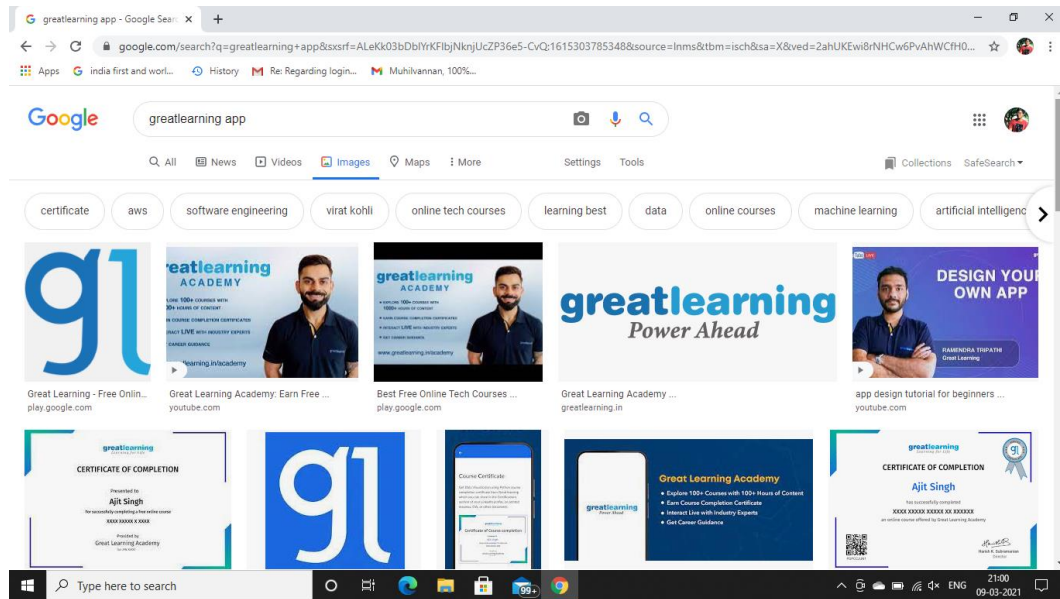


Dr. P.K.Viswanathan, Professor of probability gave training in concepts of “Probability of Data Sciences”

1. Meaning and concepts of Probability
2. Types of Probability
3. A Priori Classical Probability
4. Rules for Computing Probability
5. Marginal Probability
6. Bayes’ Theorem

1. App Search:

- ❖ Great learning platform is India's leading professional learning platform, with mission to make professionals proficient and future-ready.



2. Uses of probability in data science:

- ❖ Helps in making informed decisions about likelihood of events, based on a pattern of collected data.
- ❖ Helps to predict trends from data to use probability distribution of data.
- ❖ It provides the possibilities of occurrence of various possible outcomes that can occur in an experiment.
- ❖ Probability is the base of statistics & statistics is the base of data science. It involves the analysis of sample data, so there is always some degree of uncertainty present in it. Hence, we use probability concepts like Mean, Variance, Expected values and Prediction intervals.
- ❖ It is indispensable for analyzing data affected by chance.
- ❖ Since, probability is easy to understand, comprehensive and practical, data prediction can be made quick.

The screenshot displays a web browser window with the URL `greatlearning.in/academy/learn-for-free/courses/probability-for-data-science?career_path_id=2`. The page content includes:

- Data Science** category header.
- Probability for Data Science** course title.
- Rating: **4.35 (152 Ratings)**.
- Skill level: **Beginner**.
- Course cost: **Free**.
- Course content summary: **1.0 Hrs of video content** and **1 Quiz**.
- A prominent **RESUME LEARNING** button.
- A **Share with friends** link.
- About this course** section: "Probability is a branch of mathematics which teaches us to deal with occurrence of an event after certain repeated trials. The value here is expressed from zero to one. Probability provides basic foundations for most of the Machine Learning Algorithms. This course will give you the basic knowledge of Probability and will make you familiar with the concept of Marginal probability and Bayes theorem."
- Skills covered** section.

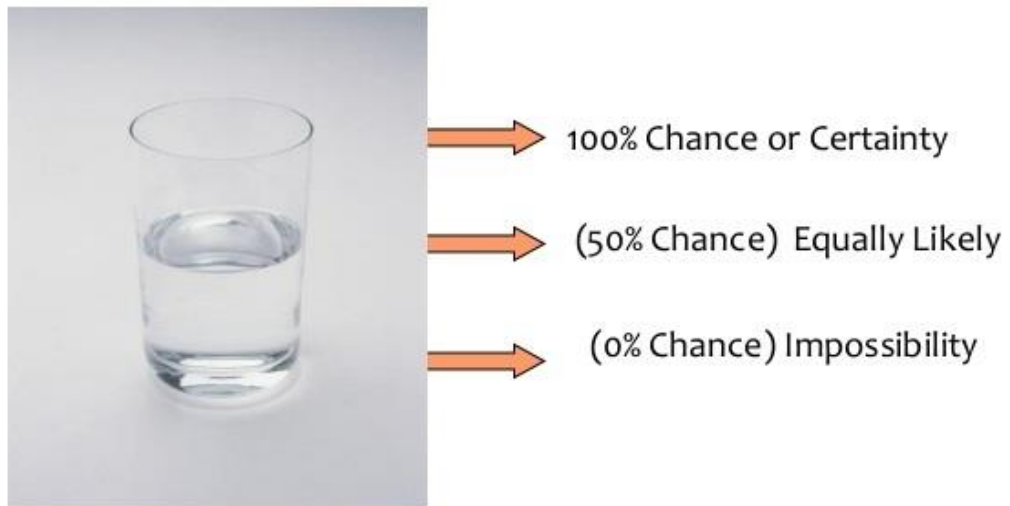
The Windows taskbar at the bottom shows the search bar, task view, and system tray with the time `21:26` and date `09-03-2021`.

3. Concepts of Probability:

- ❖ Probability refers to chance or likelihood of a particular event-taking place.
- ❖ Probability of an event A [$P(A)$] is defined as the ratio of number of ways that are favorable to the occurrence of A[m] to the total number of possible outcomes of the experiment[n] i.e. $P(A) = m : n$. $P(A) \geq 0$ and $P(A) \leq 1$ where, $P(A)$ is a pure number.
- ❖ Sample Space: The collection of all possible outcomes to a random experiment is called the sample space.
- ❖ Event: An event is an outcome or defined collection of outcomes of a random experiment (i.e.) any subset of a sample space. For example, if sample space, $S = \{56, 78, 98, 68, 45, 93\}$ then $E = \{78\}$ is an event.
- ❖ Experiment: An experiment or trial is any procedure that can be infinitely repeated and has a well-defined set of possible outcomes. The experiment is said to be random if it has more than one possible. For example, on flipping a coin twice, the possible outcomes of the experiment are $\{(H,T), (T,H), (T,T), (H,H)\}$

4. Three extreme values of probability:

The range with which probability of an event occurs refers the extreme values of probability.



The above diagram explains that,

- ❖ At the bottom of the glass, there is no place for the water to be collected, so there is 0% chance that is, there exists an impossibility of probability.
- ❖ At the middle of the glass, we have place from bottom to middle, so there is 50% chance for the glass to be filled with water that is, Equal probability.
- ❖ At the top of the glass, we have place from bottom to top, so there is 100% chance for the glass to be filled with water that is, Certain probability.

Types of probability

5. Priori classical probability:

- ❖ A priori classical probability = Desired outcomes/The total number of outcomes.
- ❖ A priori classical probability refers to the likelihood of an event occurring when there is a finite amount of outcomes and each is equally likely to occur.
- ❖ The outcomes in a priori classical probability are not influenced by the prior outcome.
- ❖ For example: The odds of rolling a 2 on a fair die are one out of 6 or $1/6$. That's one possible outcome (there's only one way to roll a 2) divided by the number of possible outcomes (1,2,3, 4,5,6).

6. Empirical Probability:

- ❖ It is also known as practical or experimental probability and it illustrates the likelihood of an event occurring based on historical data.
- ❖ Empirical probability can be calculated by dividing the number of times the event occurred to the total number of observations.
- ❖ If we observe 's' successes for 'n' trials then, the probability of success is s/n .
- ❖ For example: In a buffet, 95 out of 100 people choose to order coffee over tea. What is the empirical probability of someone ordering tea?
 - Empirical probability = $5/100=5\%$
 - The empirical probability of someone ordering tea is 5%

7. Subjective probability:

- ❖ It is derived from an individual's personal judgment or own experience about whether a specific outcome is likely to happen.
- ❖ It differs from person to person and contains a high degree of personal bias.
- ❖ It contains no formal calculations and reflects only the subject's opinion and past experience rather than on data or computation.

- ❖ For example: The weather department is predicting that it will rain in the next two hours based on wind pattern, weather situation and their software analysis. But, you may have the same predictions of rain in the net two hours based on your experience with weather or rain.

8. Mutually Exclusive Events:

- ❖ In probability theory, two events are said to be mutually exclusive if they cannot occur at the same time or simultaneously. In other words, mutually exclusive events are called disjoint events
- ❖ If two events are considered disjoint events, then the probability of both events occurring at the same time will be zero.
- ❖ If A and B are the two events, then the probability of disjoint of event A and B is written by:

$$\text{Probability of Disjoint (or) Mutually Exclusive Event} = P (A \text{ and } B) = 0$$

- ❖ The probability of either event occurring is the sum of probabilities of each event occurring that is if A and B are said to be mutually exclusive events, the probability of an event A occurring or the probability of event B occurring is given as $P(A) + P(B)$, i.e.,

$P (A \text{ or } B) = P(A) + P(B)$, it is called as specific addition rule and is valid only when two events are mutually exclusive.

- ❖ If the events A and B are not mutually exclusive, the probability of getting A or B is given as:

$$P (A \text{ or } B) = P(A) + P(B) - P (A \text{ and } B)$$

Some of the examples of the mutually exclusive events are:

- ❖ When tossing a coin, the event of getting head and tail are mutually exclusive because the probability of getting head and tail simultaneously is 0.
- ❖ In a six-sided die, the events “2” and “5” are mutually exclusive because, we cannot get both the events 2 and 5 at the same time when we threw one die.

- ❖ In a deck of 52 cards, drawing a red card and drawing a club are mutually exclusive events because all the clubs are black.

9. Independent events:

- ❖ Independent events are those events whose occurrence is not dependent on any other event that is the outcome of one event does not affect the outcome of other event.
- ❖ If A and B are independent, then the probability of both occurring is,

$$P(A \text{ and } B) = P(A) \times P(B)$$

- ❖ For example: If we flip a coin in the air and get the outcome a head, again if we flip the coin, this time we get the outcome as Tail. In both cases, the occurrence of both events is independent of each other.

Additive rule for probability:

The addition rule for probabilities describes two formulas, one for the probability for either of two mutually exclusive events happening and the other for the probability of two non- mutually exclusive events happening.

10. Addition Rule for mutually exclusive events:

- ❖ Rule: If X and Y are two mutually exclusive events, then the probability of 'X union Y' is the sum of the probability of X and the probability of Y and represented as,
$$P(X \cup Y) = P(X) + P(Y)$$

- ❖ Derivation: Let E be a random experiment and N(X) be the number of frequency of the event X in E. Since X and Y are two mutually exclusive events then;

$$N(X \cup Y) = N(X) + N(Y)$$

Dividing both the sides by N, we get,

$$N(X \cup Y)/N = N(X)/N + N(Y)/N;$$

Now taking limit N to ∞ , we get

$$\text{Probability of } P(X \cup Y) = P(X) + P(Y)$$

- ❖ For example: A single 6-sided die is rolled. What is the probability of rolling 2 or 5?

Here the events are mutually exclusive, by using additive rule for mutually exclusive events,

$$P(A \text{ or } B) = P(A) + P(B)$$

Given, A= 2, B =5

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

$$P(5) = \frac{1}{6}$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{2}{6}$$

$$= \frac{1}{3}$$

- ❖ For example: A spinner has 4 equal sectors colored yellow, blue, green, and red. What is the probability of landing on red or blue after spinning this spinner?

Here the events are mutually exclusive, by using additive rule for mutually exclusive events,

$$P(A \text{ or } B) = P(A) + P(B)$$

Given, A= red, B =blue

$$P(\text{red}) = \frac{1}{4}$$

$$P(\text{blue}) = \frac{1}{4}$$

$$P(\text{red or blue}) = P(\text{red}) + P(\text{blue})$$

$$= \frac{1}{4} + \frac{1}{4}$$

$$= \frac{2}{4}$$

$$= \frac{1}{2}$$

11. Additive Rule for Mutually Non-Exclusive Events:

- ❖ Rule: If X and Y are two mutually Non- Exclusive Events, then the probability of ‘X union Y’ is the difference between the sum of the probability of X and the probability of Y and the probability of ‘X intersection Y’ and represented as,

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

- ❖ Derivation: The events X - XY, XY and Y - XY are pair-wise mutually exclusive events then,

$$X = (X - XY) + XY$$

$$Y = XY + (Y - XY)$$

Now, $P(X) = P(X - XY) + P(XY)$ or, $P(X - XY) = P(X) - P(XY)$

Similarly, $P(Y - XY) = P(Y) - P(XY)$

Again, $P(X + Y) = P(X - XY) + P(XY) + P(Y - XY)$

$$\Rightarrow P(X + Y) = P(X) - P(XY) + P(XY) + P(Y) - P(XY)$$

$$\Rightarrow P(X + Y) = P(X) + P(Y) - P(XY)$$

$$\Rightarrow P(X + Y) = P(X) + P(Y) - P(X) P(Y)$$

Therefore, $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$

- ❖ For example: If a single card is chosen at random from a standard deck of 52 playing cards. What is the probability of choosing a king or a club?

Here, the events are non-mutually exclusive. The addition causes the king of clubs to be counted twice, so its probability must be subtracted. By using additive rule for non-mutually exclusive events,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$\begin{aligned} P(\text{king or club}) &= P(\text{king}) + P(\text{club}) - P(\text{king of clubs}) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} \\ &= \frac{16}{52} \\ &= \frac{4}{13} \end{aligned}$$

- ❖ For example: In a math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade. If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?

Here, the events are non-mutually exclusive that is the addition causes the girls with A grade to be counted twice, so its probability must be subtracted. By using additive rule for non-mutually exclusive events,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(\text{girl or A}) = P(\text{girl}) + P(A) - P(\text{girl and A})$$

$$\begin{aligned} &= \frac{13}{30} + \frac{9}{30} - \frac{5}{30} \\ &= \frac{17}{30} \end{aligned}$$

Multiplicative rules for probability:

12. Multiplication Rule for dependent events

- ❖ Rule: When two events A and B are dependent, the probability of intersection of A and B equals the product of the probability of A and the probability of B given that A has happened.

$$P(A \cap B) = P(A) \times P(A|B)$$

Here, $P(A|B)$ is the conditional probability of event A given that B has happened.

- ❖ Derivation: We know that the conditional probability of event A given that B has occurred is denoted by $P(A|B)$ and is given by:

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

Where, $P(B) \neq 0$

$$P(A \cap B) = P(B) \times P(A|B) \dots \dots (1)$$

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

Where, $P(A) \neq 0$.

$$P(B \cap A) = P(A) \times P(B|A)$$

Since, $P(A \cap B) = P(B \cap A)$

$$P(A \cap B) = P(A) \times P(B|A) \dots \dots (2)$$

From (1) and (2), we get:

$$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A)$$

where, $P(A) \neq 0, P(B) \neq 0$.

- ❖ For example: An urn contains 20 red and 10 blue balls. Two balls are drawn from a bag one after the other without replacement. What is the probability that both the balls drawn are red?

Solution: Let A and B denote the events that first and second ball drawn are red balls. To find: $P(A \cap B)$ or $P(AB)$.

$$P(A) = P(\text{red balls in first draw}) = 20/30$$

Now, only 19 red balls and 10 blue balls are left in the bag. Probability of drawing a red ball in second draw too is an example of conditional probability where drawing of second ball depends on the drawing of first ball.

Hence Conditional probability of B on A will be, $P(B|A) = 19/29$

By multiplication rule of probability,

$$P(A \cap B) = P(A) \times P(B|A)$$

$$P(A \cap B) = 20/30 \times 19/29 = 3887$$

13. Multiplication Rule for independent events:

- ❖ Rule: When the two events A and B are independent, the probability of the simultaneous occurrence of A and B (probability of intersection of A and B) equals the probability of A and the probability of B.

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

- ❖ Derivation: By using multiplication rules in probability, such as;

$$P(A \cap B) = P(A) \times P(B|A) ; \text{ if } P(A) \neq 0$$

$$P(A \cap B) = P(B) \times P(A|B) ; \text{ if } P(B) \neq 0$$

Now, from multiplication rule we know;

$$P(A \cap B) = P(A) \times P(B|A)$$

Since A and B are independent, therefore;

$$P(B|A) = P(B)$$

Therefore, again we get;

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

- ❖ This rule can be extended to more than two independent events.
- ❖ For example: If you have a cowboy hat, a top hat, and an Indonesian hat called a songkok and also have four shirts: white, black, green, and pink. If you want to

choose one hat and one shirt at random then what will be the probability that you choose the songkok and the black shirt?

Solution:

The two events are independent events. So the choice of hat has no effect on the choice of shirt.

There are three different hats, so the probability of choosing the songkok is $1/3$.

There are four different shirts, so the probability of choosing the black shirt is $1/4$.

So, by the Multiplication Rule for independent events,

$$P(\text{songkok and black shirt}) = 1/3 \cdot 1/4 = 1/12$$

14. Marginal Probability

- ❖ In probability theory, the marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset.
- ❖ It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables.
- ❖ Marginal probability is the probability of an event irrespective of the outcome of another variable.
- ❖ A marginal distribution of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.
- ❖ The term “Marginal” is used to indicate that the probabilities are calculated using a contingency table (i.e. Joint probability table).
- ❖ Contingency table consists of rows and columns of two attribute at different levels with frequencies or numbers in each of the cells. It is a matrix of frequencies assigned to rows and columns
- ❖ The marginal probability of one random variable in the presence of additional random variables is referred to as the marginal probability distribution.
- ❖ Marginal variables are those variables in the subset of variables being retained. These concepts are "marginal" because they can be found by summing values in a table along rows or columns, and writing the sum in the margins of the table.
- ❖ The distribution of the marginal variables (the marginal distribution) is obtained by marginalizing that is, focusing on the sums in the margin over the distribution of the variables being discarded, and the discarded variables are said to have been marginalized out.
- ❖ In many applications, an analysis may start with a given collection of random variables, then first extend the set by defining new ones (such as the sum of the original random variables) and finally reduce the number by placing interest in the marginal distribution of a subset (such as the sum).
- ❖ Different types of analyses can be done by treating a different subset of variables as the marginal variables.

- ❖ For example: Of the cars on a used car lot, 70% have air conditioning(AC) and 40% have a CD player(CD), 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC? (i.e.) we have to find $P(\text{CD}/\text{AC})$.

Attributes	CD	No CD	Marginal total
AC	20	50	70
No AC	20	10	30
Marginal total	40	60	100

From the table, it is easy to see that there are 70 cars with AC out of which 20 have CD. Hence, $P(\text{CD}/\text{AC}) = 20/70 = 2/7$.

15. Bayes' theorem:

- ❖ Bayes' theorem was developed and named by British mathematician Thomas Bayes after 18th-century. It is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics.
- ❖ Bayes' theorem: $P(A/B) = \frac{P(B/A).P(A)}{P(B)}$
- ❖ Here, P(A) and P(B) are the probabilities of events A and B respectively. P(B/A) is the conditional probability.
- ❖ It is a mathematical formula for determining conditional probability of the given event and used to find the reverse probabilities if we know the conditional probability of an event.
- ❖ Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring.
- ❖ For example: Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weather man has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecast rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?
- ❖ Here $P(A_1) = 5/365 = 0.0136985$ (Since, it rains 5 days a year) and $P(B/A_1) = 0.9$ (When it rains, the weatherman predicts rain 90% of the time)
- ❖ $P(B_2) = 360/365 = 0.9$ (It does not rain 360 days a year) and $P(B/A_2) = 0.1$ (When it does not rain, the weatherman predicts rain 10% of the time)
- ❖ Now, Probability of raining,

$$P(A_1/B) = \frac{P(A_1) P(B/ A_1)}{P(A_1) P(B/ A_1) + P(A_2) P(B/ A_2)}$$

$$P(A_1/B) = (0.014)(0.9) / [0.014(0.9) + 0.986(0.1)]$$

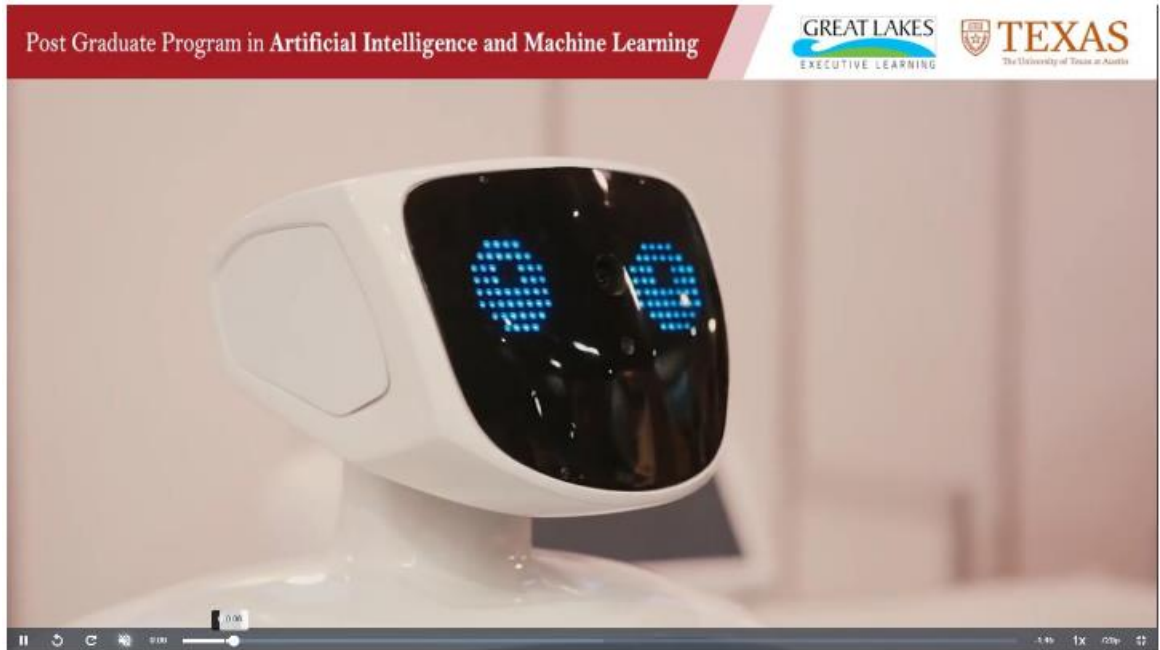
$$P(A_1/B) = 0.111$$

16. Applications of Bayes' theorem

- ❖ It provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.
- ❖ It shows the Relation between one conditional probability and its inverse.
- ❖ It is the most important concept in Data Science and is most widely used in Machine Learning as a classifier that makes use of Naive Bayes' Classifier.
- ❖ It is also used in discriminant functions, decision surfaces and Bayesian parameter estimation
- ❖ It allows you to update predicted probabilities of an event by incorporating new information.

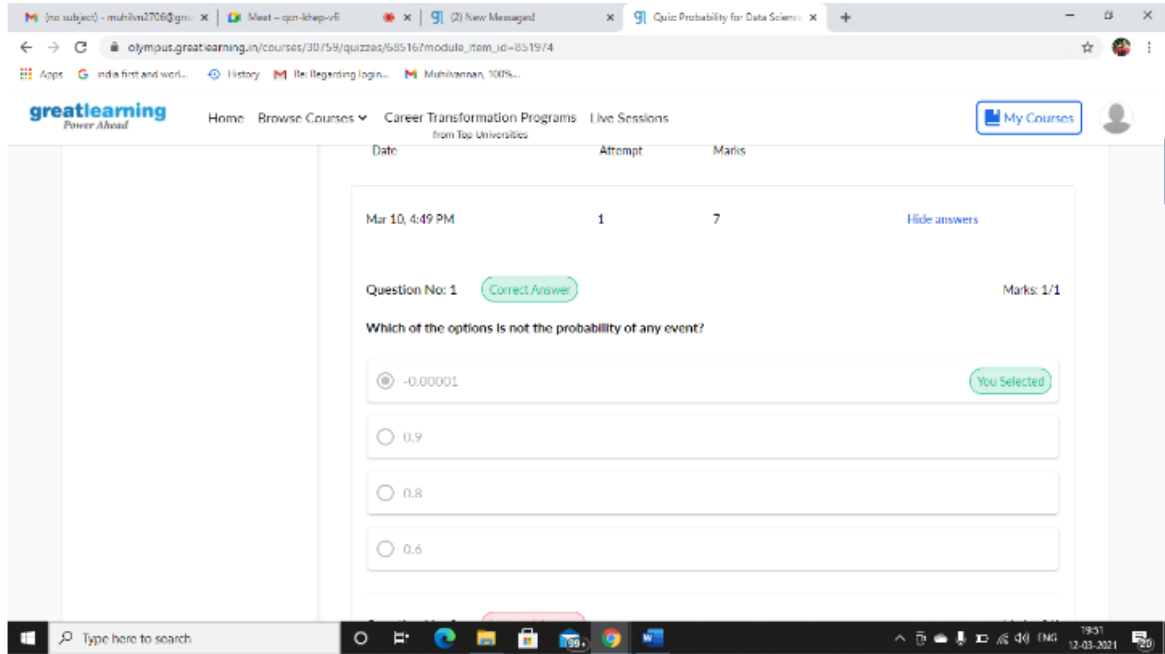
PG programs overview:

Kumar Muthuraman, Faculty director, Centre for Analytics and Transformative Technologies, PGP-AIML introduced that post graduate programs in Artificial Intelligence and Machine Learning were offered by University of Texas in Austin in collaboration with Great learning.



QUIZ

Checkpoint questions were provided to check our understandability of the concepts and to increase the real life application of probability. The quiz answers were checked and the correct answers were provided for our reference.



The screenshot shows a web browser window displaying a quiz on the Great Learning platform. The browser's address bar shows the URL: `olympus.greatlearning.in/courses/30/59/quizzes/585167/module_item_id=851974`. The page header includes the Great Learning logo and navigation links for Home, Browse Courses, Career Transformation Programs, and Live Sessions. A table at the top of the quiz content shows the following data:

Date	Attempt	Marks
Mar 10, 4:49 PM	1	7

Below the table, the quiz details are shown: "Question No: 1" with a "Correct Answer" label and "Marks: 1/1". The question text is: "Which of the options is not the probability of any event?". The options are: -0.00001, 0.9, 0.8, and 0.6. A "You Selected" label is next to the first option. The browser's taskbar at the bottom shows the date and time as 12:03:2021.

Conclusion:

The Internship is very useful and it gave me an excellent experience. I have learned concepts of probability to measure uncertainty and perform associated analyses that are essential in making effective business decisions.