Machine Learning (21AI63)

Module 4

Bayesian Learning

- Bayesian learning is an approach to machine learning and statistical inference that is based on Bayes theorem. It provides a framework for updating our beliefs or knowledge about a hypothesis or model as new evidence or data becomes available.

Bayesian learning methods are relevant to study of machine learning for two different reasons.

- First, Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems.
- The second reason is that they provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities.

Features of Bayesian Learning Methods

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct. This provides a more flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. In Bayesian learning, prior knowledge is provided by asserting (i) a prior probability for each candidate hypothesis, and (ii) a probability distribution over observed data for each possible hypothesis.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions.
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

Bayes Theorem

- Bayes theorem, named after Thomas Bayes, is a fundamental concept in probability theory and statistics. It provides a way to update our beliefs or knowledge about an event or hypothesis based on new evidence or data. The theorem is derived from conditional probability.
- Mathematically, Bayes theorem can be stated as follows:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

  where: P(h|D) represents the probability of event h occurring given that event D has occurred. (Posterior probability)
  P(D|h) represents the probability of event D occurring given that event h has occurred.
  P(h) and P(D) represent the probabilities of events h and D occurring, respectively.
- In words, Bayes theorem states that the probability of event h occurring given the occurrence of event D is proportional to the probability of event D occurring given event h has occurred, multiplied by the prior probability of event h, and divided by the prior probability of event D.
- Bayes' theorem allows us to update our beliefs by incorporating new evidence. We start with a prior probability, which represents our initial belief about the occurrence of an event. As we observe new

data or evidence, we update our belief using Bayes theorem to obtain the posterior probability, which represents our updated belief after considering the new evidence.

- In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis h ∈ H given the observed data D. Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis.
- Bayes theorem to calculate the posterior probability of each candidate hypothesis is h_{MAP}.

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{argmax}\ P(h|D) \\ &= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{argmax}\ P(D|h)P(h) \end{aligned}$$

P(D) can be dropped, because it is a constant independent of h.

Example of Bayes Theorem

- Consider a medical diagnosis problem in which there are two alternative hypotheses: (i) that the patient has particular form of cancer, and (ii) that the patient does not.
- The available data is from a particular laboratory test with two possible outcomes: + (positive) and - (negative).
- We have prior knowledge that over the entire population of people only .008 have this disease.
- The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result.
- The above situation can be summarized by the following probabilities:

$$\begin{array}{ll} P(cancer) = .008 & P(\neg cancer) = 0.992 \\ P(\oplus|cancer) = .98 & P(\ominus|cancer) = .02 \\ P(\oplus|\neg cancer) = .03 & P(\ominus|\neg cancer) = .97 \end{array}$$

- Suppose a new patient is observed for whom the lab test returns a positive (+) result. Should we diagnose the patient as having cancer or not?

$$\begin{gathered} P(\oplus|cancer)P(cancer) = (.98).008 = .0078 \\ P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298 \\ \Rightarrow h_{MAP} = \neg cancer \end{gathered}$$

Result: Cancer not present

- The exact posterior probabilities can also be determined by normalizing the above quantities so that they sum to 1.

$$P(cancer|\oplus) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\neg cancer|\oplus) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

- Suppose a new patient is observed for whom the lab test returns a negative (-) result. Should we diagnose the patient as having cancer or not?
Result: Cancer not present

Bayes Theorem and Concept Learning

- In Bayesian learning, concept learning refers to the process of learning concepts or categories from data by applying Bayesian principles. It involves using Bayes' Theorem and probability theory to infer the underlying concept that generated the observed data.
- Bayesian learning is a principled and probabilistic approach that allows us to incorporate prior knowledge and update beliefs as new evidence becomes available.
- Bayes' Theorem enables us to perform probabilistic inference in concept learning. It allows us to calculate the posterior probabilities of different concepts given the observed data. These posterior probabilities represent the updated beliefs about the likelihood of each concept after considering the evidence.
- In concept learning, we often have some prior knowledge or beliefs about the concepts before observing any data. Bayes' Theorem incorporates this prior knowledge into the learning process as the "prior probabilities." These prior probabilities act as a starting point for the learning process and can be based on domain knowledge, previous experience, or even assumptions about the distribution of concepts.
- In Bayes' Theorem, the likelihood function represents the probability of observing the given data given a specific concept. In concept learning, the likelihood function captures how likely the observed data is under each concept. The likelihood function is influenced by the model assumptions and how well the model can explain the data for each concept.

Brute-Force Bayes Concept Learning

- The "Brute-Force Map Learning" algorithm is a simple and straightforward approach to solving the problem of Maximum A Posteriori (MAP) estimation in a probabilistic setting. MAP learning aims to find the most probable hypothesis or model parameters given the observed data. The brute-force approach involves exhaustively evaluating all possible hypotheses and selecting the one with the highest posterior probability.
- Given the training data, the Bayes theorem determines the posterior probability of each hypothesis. It calculates the likelihood of each conceivable hypothesis before determining which is the most likely.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability.

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

- To calculate, we need to know the values of P(h) and P(D/h). To choose these to be consistent with the following assumptions:
  - There is no noise in the training data D (i.e., $d_i = c(x_i)$).
  - The hypothesis space H contains the goal notion c.
  - We have no reason to conclude that one hypothesis is more likely than another based on prior evidence.
- Since we're assuming the training data to be noise-free, the chances of observing classification d_i given h are 1 if d_i = h(xi) and 0 if di != (x_i). Therefore,

$$P(D|h) = \begin{cases} 1 \text{ if } d_i = h(x_i) \text{ for all } d_i \text{ in } D \\ 0 \text{ otherwise} \end{cases}$$

- Given no previous information of which hypothesis is more likely, it is fair to give each hypothesis h in H the same prior probability. We should require that these prior probabilities amount to 1 because we presume the target notion is contained in H.
- We can derive P(D) from the theorem of total probability and the fact that the hypothesis is mutually exclusive.

$$\text{(i.e., } (\forall i \neq j)(P(h_i \wedge h_j) = 0))$$

$$\begin{aligned} P(D) &= \sum_{h_i \in H} P(D|h_i)P(h_i) \\ &= \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{|H|} + \sum_{h_i \notin VS_{H,\ D}} 0 \cdot \frac{1}{|H|} \\ &= \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{|H|} \\ &= \frac{|VS_{H,D}|}{|H|} \end{aligned}$$

- To summarize, Bayes theorem implies that the posterior probability p(h/D) under the assumed P(h) and P(D/h) is

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$

- where | VSH,D | is the number of hypotheses from H consistent with D.

Maximum likelihood hypotheses for predicting probabilities

- Maximum likelihood hypothesis used to estimate the parameters of a statistical model by finding the parameter values that maximize the likelihood of the observed data.
- Consider the setting in which we wish to learn a nondeterministic (probabilistic) function f : X → {0, 1}, which has two discrete output values.
- We want a function approximator whose output is the probability that f(x) = 1. In other words, learn the target function f ` : X → [0, 1] such that f ` (x) = P(f(x) = 1)
- To find a maximum likelihood hypothesis for f' in this setting, optimize the criteria:
    - First obtain an expression for P(D|h)
    - Assume the training data D is of the form D = {(x1, d1) . . . (x_m, d_m)}, where d_i is the observed 0 or 1 value for f (x_i).
    - Both x_i and d_i as random variables, and assuming that each training example is drawn independently, we can write P(D|h) as

$$P(D \mid h) = \prod_{i=1}^{m} P(x_i, d_i \mid h)$$

    - The probability P(di|h, xi)

$$P(d_i|h, x_i) = \begin{cases} h(x_i) & \text{if } d_i = 1 \\ (1 - h(x_i)) & \text{if } d_i = 0 \end{cases}$$

    - We write an expression for the maximum likelihood hypothesis

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i)$$

    - The last term is a constant independent of h, so it can be dropped

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i}$$

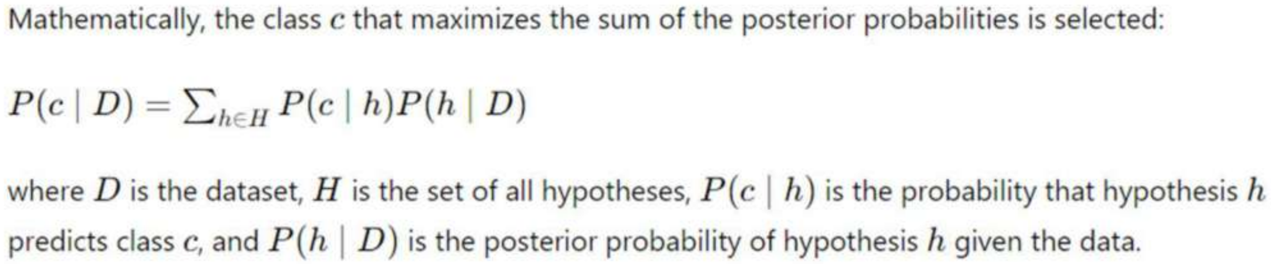
- It easier to work with the log of the likelihood, yielding

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i))$$

- The above equation describes the quantity that must be maximized in order to obtain the maximum likelihood hypothesis in our current problem setting.

Bayes Optimal Classifier

- The Bayes Optimal Classifier is a theoretical model that represents the best possible classification performance given a particular dataset and a model of uncertainty.
- It makes predictions based on the posterior probability distribution over all possible hypotheses.
- This classifier chooses the class with the highest probability, taking into account all the possible hypotheses weighted by their posterior probabilities.



Mathematically, the class c that maximizes the sum of the posterior probabilities is selected:

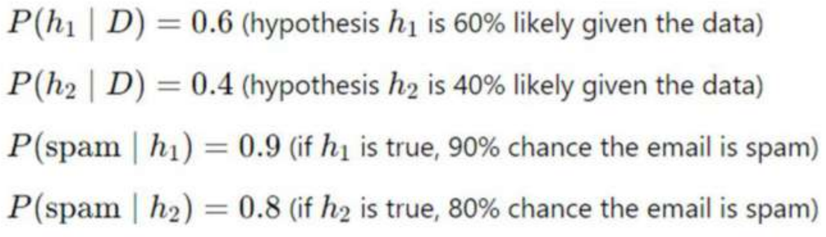
$$P(c \mid D) = \sum_{h \in H} P(c \mid h) P(h \mid D)$$

where D is the dataset, H is the set of all hypotheses, $P(c \mid h)$ is the probability that hypothesis h predicts class c, and $P(h \mid D)$ is the posterior probability of hypothesis h given the data.

Example

- Consider a case where we want to classify an email as "spam" or "not spam" using a dataset of emails. Suppose we have two hypotheses:
  - h1: An email is spam if it contains the word "offer".
  - h2: An email is spam if it contains the word "free".

Let's assume the following probabilities based on our training data:



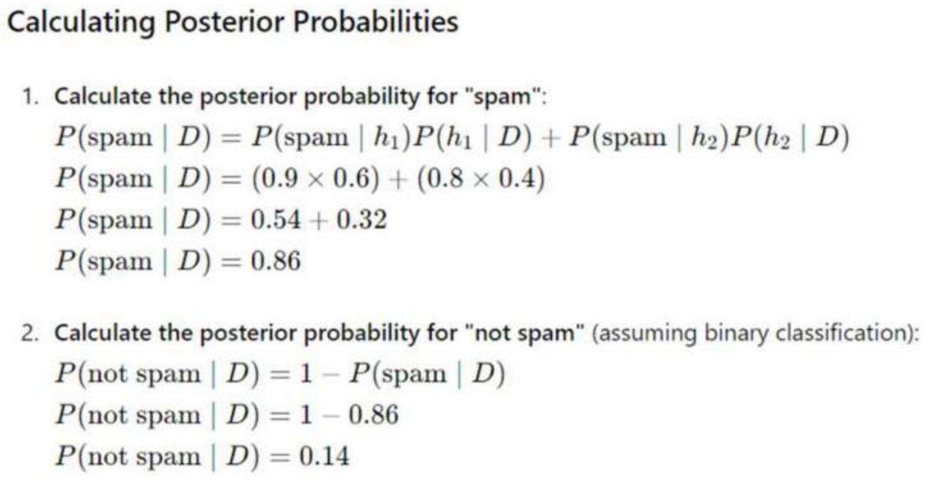
$P(h_1 \mid D) = 0.6$ (hypothesis h_1 is 60% likely given the data)

$P(h_2 \mid D) = 0.4$ (hypothesis h_2 is 40% likely given the data)

$P(\text{spam} \mid h_1) = 0.9$ (if h_1 is true, 90% chance the email is spam)

$P(\text{spam} \mid h_2) = 0.8$ (if h_2 is true, 80% chance the email is spam)

We want to classify a new email that contains both "offer" and "free".



Calculating Posterior Probabilities

1. **Calculate the posterior probability for "spam":**
   $P(\text{spam} \mid D) = P(\text{spam} \mid h_1)P(h_1 \mid D) + P(\text{spam} \mid h_2)P(h_2 \mid D)$
   $P(\text{spam} \mid D) = (0.9 \times 0.6) + (0.8 \times 0.4)$
   $P(\text{spam} \mid D) = 0.54 + 0.32$
   $P(\text{spam} \mid D) = 0.86$

2. **Calculate the posterior probability for "not spam"** (assuming binary classification):
   $P(\text{not spam} \mid D) = 1 - P(\text{spam} \mid D)$
   $P(\text{not spam} \mid D) = 1 - 0.86$
   $P(\text{not spam} \mid D) = 0.14$

- Since P(spam|D)=0.86 is greater than P(not spam|D)=0.14, the Bayes Optimal Classifier would classify this email as "spam".
- The Bayes Optimal Classifier provides a way to make the most accurate predictions possible by considering all possible hypotheses and their posterior probabilities.

Gibbs Algorithm

- Although the Bayes optimal classifier obtains the best performance that can be achieved from the given training data, it can be quite costly to apply. The expense is due to the fact that it computes the posterior probability for every hypothesis in H and then combines the predictions of each hypothesis to classify each new instance.
- An alternative, less optimal method is the Gibbs algorithm, defined as follows:
  - Choose a hypothesis h from H at random, according to the posterior probability distribution over H.
  - Use h to predict the classification of the next instance x.
- Given a new instance to classify, the Gibbs algorithm simply applies a hypothesis drawn at random according to the current posterior probability distribution.
- Under certain conditions the expected misclassification error for Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier.

Minimum Description Length Principle

- In Bayesian learning, the Minimum Description Length (MDL) principle is closely related to the concepts of model selection and inference.
- Both MDL and Bayesian learning aim to find the best model for a given set of data, but they do so from different perspectives.
- The MDL principle can be seen as a practical implementation of Bayesian model selection. It focuses on the trade-off between model complexity and data fit by minimizing the total description length, which corresponds to maximizing the posterior probability of a model.
- The Minimum Description Length principle recommends choosing the hypothesis that minimizes the sum of the two description lengths.

$$h_{MAP} = \underset{h \in H}{argmin}\ L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

  Where, C_H and $C_{D|h}$ are the optimal encodings
- Minimum Description Length principle:

$$h_{MDL} = \underset{h \in H}{\operatorname{argmin}}\ L_{C_1}(h) + L_{C_2}(D \mid h)$$

  Where, codes C_1 and C_2 to represent the hypothesis and the data given the hypothesis

Naive Bayes Classifier

- Naive Bayes classifier is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- The Bayesian approach to classifying the new instance is to assign the most probable target value, V_{MAP}, given the attribute values (a_1, a_2, . . .a_m) that describe the instance:

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}}\ P(v_j|a_1, a_2 \ldots a_n)$$

- Use Bayes theorem to rewrite this expression as

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

- The Naive Bayes classifier is based on the assumption that the attribute values are conditionally independent given the target value. Means, the assumption is that given the target value of the instance, the probability of observing the conjunction (a_1, a_2, . . .a_m), is just the product of the probabilities for the individual attributes:

  $$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

  Substituting this into equation (1),

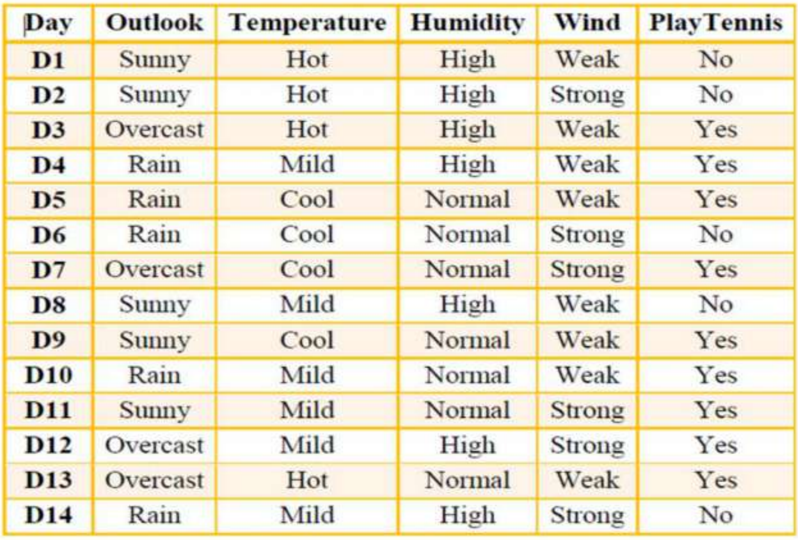
  $$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

  Where, V_{NB} denotes the target value output by the Naive Bayes classifier.

- Use this equation to find the posterior probability for each and every hypothesis. The hypothesis which gives the maximum value is considered as the solution.
- It is mainly used in text classification that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naive Bayes Algorithm are spam filtration, sentimental analysis, and classifying articles.

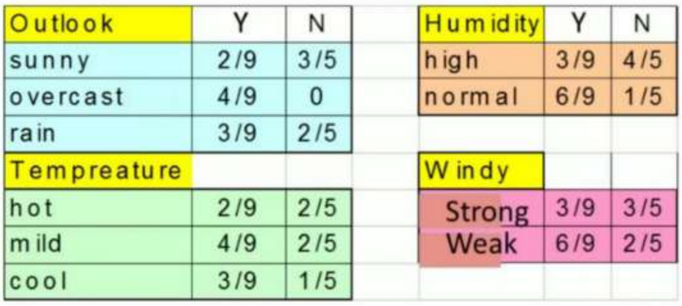
An Illustrative Example:

- Let us apply the Naive Bayes classifier to a concept learning problem i.e., classifying days according to whether someone will play tennis.
- The below table provides a set of 14 training examples of the target concept PlayTennis, where each day is described by the attributes Outlook, Temperature, Humidity, and Wind.
- Use the naive Bayes classifier and the training data from this table to classify the following test instance: < Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong >
- Our task is to predict the target value (yes or no) of the target concept PlayTennis for this new instance.



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- The probabilities of the different target values can easily be estimated based on their frequencies over the 14 training examples
  - P(P1ayTennis = yes) = 9/14 = 0.64
  - P(P1ayTennis = no) = 5/14 = 0.36
- Compute the conditional probabilities:



Outlook	Y	N		Humidity	Y	N
sunny	2/9	3/5		high	3/9	4/5
overcast	4/9	0		normal	6/9	1/5
rain	3/9	2/5				
Tempreature				Windy		
hot	2/9	2/5		Strong	3/9	3/5
mild	4/9	2/5		Weak	6/9	2/5
cool	3/9	1/5				

- Calculate V_{NB}

$$\mathbf{V_{NB} = \underset{v_j \in V}{argmax}\ P(v_j) \prod_i P(a_i|v_j)}$$

$$\mathbf{V_{NB} = \underset{v_j \in \{yes\ ,no\}}{argmax}\ P(v_j)}\ P(\text{Outlook=sunny}|v_j\ )\ P(\text{Temperature=cool}|v_j\ )\ P(\text{Humidity=high}|v_j\ )\ P(\text{Wind=strong}|v_j\ )$$

$$P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes)\ = .0053$$

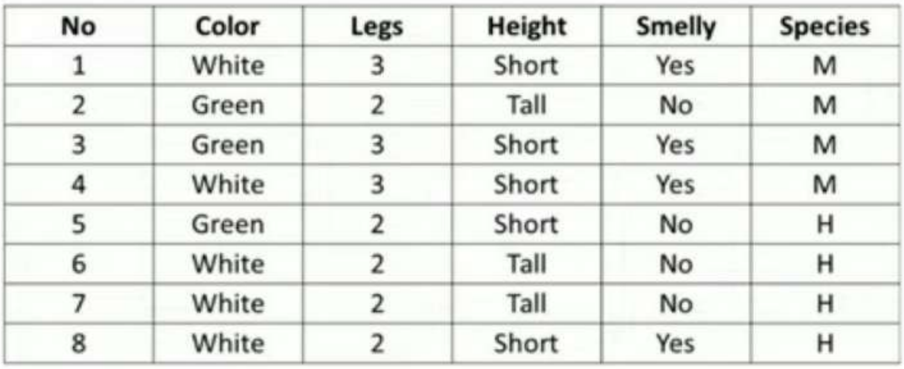
$$P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no)\ \ = .0206$$

- Thus, the naive Bayes classifier assigns the target value PlayTennis to "No" to this new instance, based on the probability estimates learned from the training data.

$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = \mathbf{0.205} \qquad v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = \mathbf{0.795}$$

Example 2:

- The below table provides a set of 8 training examples of the target concept Species, where each day is described by the attributes Colour, Legs, Height, and Smelly.



No	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

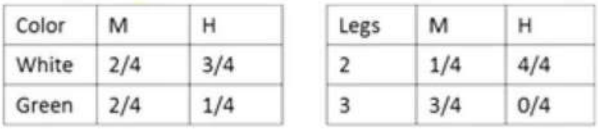
- Use the Naive Bayes classifier and the training data from this table to classify the following test instance:

(Color=Green, legs=2, Height=Tall, and Smelly=No)

- Our task is to predict the target value (M or H) of the target concept Species for this new instance.

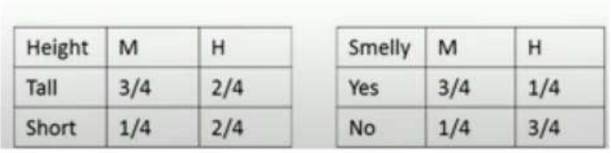
$$P(M) = \frac{4}{8} = 0.5 \quad P(H) = \frac{4}{8} = 0.5$$

- Compute the conditional probabilities:



Color	M	H
White	2/4	3/4
Green	2/4	1/4

Legs	M	H
2	1/4	4/4
3	3/4	0/4



Height	M	H
Tall	3/4	2/4
Short	1/4	2/4

Smelly	M	H
Yes	3/4	1/4
No	1/4	3/4

- Calculate V_{NB}

$$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}}\ P(v_j) \prod_i P(a_i | v_j)$$

$$p(M|New\ Instance) = p(M) * p(Color = Green|M) * p(Legs = 2|M) * p(Height = tall|M) * p(Smelly = no\ |M)$$

$$p(M|New\ Instance) = 0.5 * \frac{2}{4} * \frac{1}{4} * \frac{3}{4} * \frac{1}{4} = 0.0117$$

$$p(H|New\ Instance) = p(H) * p(Color = Green|H) * p(Legs = 2|H) * p(Height = tall|H) * p(Smelly = no\ |H)$$

$$p(H|New\ Instance) = 0.5 * \frac{1}{4} * \frac{4}{4} * \frac{2}{4} * \frac{3}{4} = 0.047$$

$p(H|New\ Instance) > p(M|New\ Instance)$

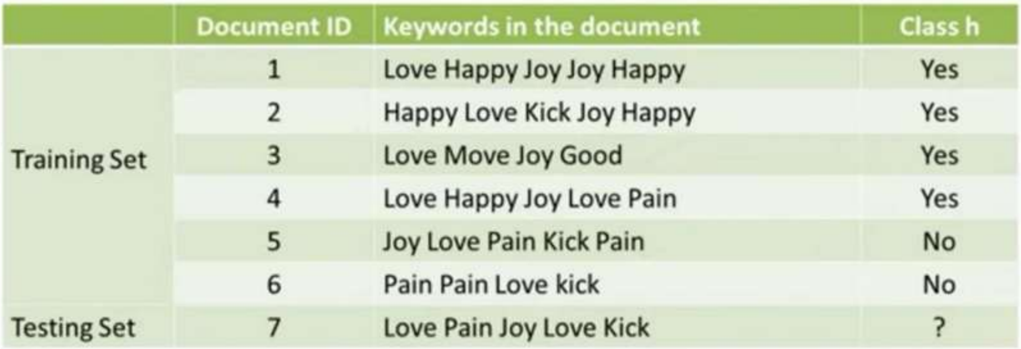
Hence the new instance belongs to Speices H

- Thus, the Naive Bayes classifier assigns the target value Species to “H” to this new instance, based on the probability estimates learned from the training data.

Naive Bayes Classifier for Text Classification

- Naive Bayes is often applied in various text classification tasks such as sentiment analysis, spam filtering, topic classification, and document categorization.
- In this example, the goal is to classify the test data into class h or not h (⌐h). Dataset for the text classification with training and test data is given below.
- The probability of the document ‘d’ being in class c is computed as follows:

$$p(c|d) \propto p(c) \prod_{1 \le k \le n_d} p(t_k|c)$$



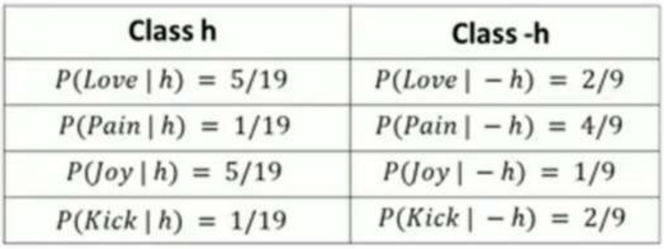
	Document ID	Keywords in the document	Class h
Training Set	1	Love Happy Joy Joy Happy	Yes
	2	Happy Love Kick Joy Happy	Yes
	3	Love Move Joy Good	Yes
	4	Love Happy Joy Love Pain	Yes
	5	Joy Love Pain Kick Pain	No
	6	Pain Pain Love kick	No
Testing Set	7	Love Pain Joy Love Kick	?

- Where $p(t_k|c)$ is the conditional probability of term t_k occurring in document of class c.
- The prior probabilities are:

$$P(h) = \frac{4}{6} = \frac{2}{3}$$
$$and$$
$$p(-h) = \frac{2}{6} = \frac{1}{3}$$

Testing Example:

Love Pain Joy Love Kick = ?



Class h	Class -h
$P(Love \mid h) = 5/19$	$P(Love \mid -h) = 2/9$
$P(Pain \mid h) = 1/19$	$P(Pain \mid -h) = 4/9$
$P(Joy \mid h) = 5/19$	$P(Joy \mid -h) = 1/9$
$P(Kick \mid h) = 1/19$	$P(Kick \mid -h) = 2/9$

$$p(c|d) \propto p(c) \prod_{1 \le k \le n_d} p(t_k|c)$$

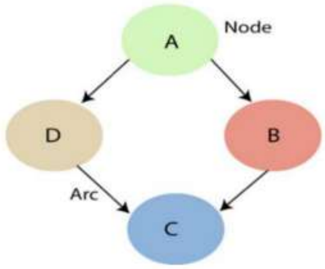
$$\begin{aligned} P(h|d_7) &= P(h) * (P(Love|h) * (P(Love|h) * P(Pain|h) * P(Joy|h) * P(Kick|h) \\ &= (2/3) * (5/19) * (5/19) * (1/19) * (5/19) * (1/19) = 0.0000336 \end{aligned}$$

$$\begin{aligned} p(-h|d7) &= p(-h) * P(Love| - h) * P(Love| - h) * P(Pain| - h) * P(Joy| - h) * P(Kick| - h) \\ &= (1/3) * (2/9) * (2/9) * (4/9) * (1/9) * (2/9) = \mathbf{0.00018} \end{aligned}$$

- Therefore, class label for test data is ‘No’.

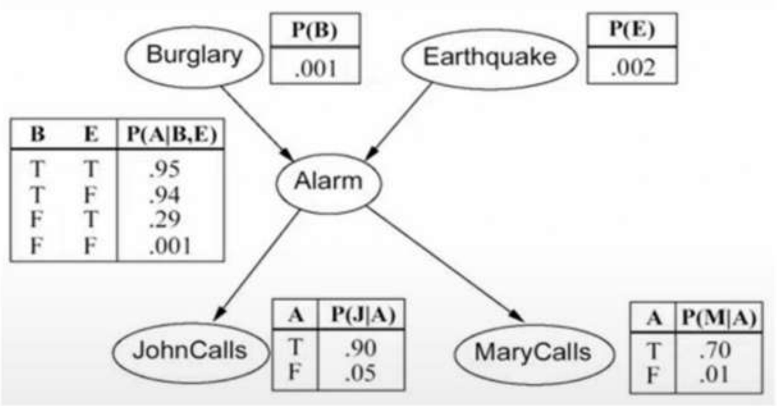
Bayesian Belief Network

- Bayesian Belief Network is a graphical representation of different probabilistic relationships among random variables in a particular set. It is a classifier with no dependency on attributes i.e it is condition independent.
- A Bayesian belief network describes the probability distribution governing a set of variables by specifying a set of conditional independence assumptions along with a set of conditional probabilities.
- In contrast to the naive Bayes classifier, which assumes that all the variables are conditionally independent given the value of the target variable, Bayesian belief networks allow stating conditional independence assumptions that apply to subsets of the variables.
- Thus, Bayesian belief networks provide an intermediate approach that is less constraining than the global assumption of conditional independence made by the naive Bayes classifier, but more tractable than avoiding conditional independence assumptions altogether.
- Bayesian networks are probabilistic, because these networks are built from a probability distribution, and also use probability theory for prediction and anomaly detection.
- Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network.
- It can also be used in various tasks including prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction, and decision making under uncertainty.
- The main advantages of Bayesian belief networks are their ability to handle uncertainty, model complex dependencies among variables, and provide a transparent representation of the probabilistic relationships.
- A Bayesian network graph is made up of nodes and Arcs. Each node corresponds to the random variables, and a variable can be continuous or discrete.
- Arc or directed arrows represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.
- These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other.



Example 1:

- You have installed a new burglar alarm at your home to detect burglary. The alarm reliably responds at detecting a burglary but also responds for minor earthquakes. You have two neighbours John and Mary, who have taken a responsibility to inform you when they hear the alarm. John always calls when he hears the alarm, but sometimes he gets confused with the phone ringing and calls at that time too. On the other hand, Mary likes to listen to high music, so sometimes she misses to hear the alarm.
- Problem: Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and John and Mary, both called you.
- List of all events occurring in this network:
    - Burglary (B)
    - Earthquake(E)
    - Alarm(A)
    - John Calls(J)
    - Mary calls(M)



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

- The probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and John and Mary both called you is:

$$
\begin{aligned}
P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) &= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b, \neg e)\, P(\neg b)\, P(\neg e) \\
&= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 \\
&= 0.00062
\end{aligned}
$$

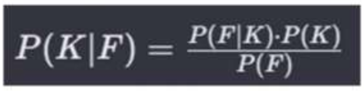
Problem:

From a standard deck of playing cards, a single card is drawn. The probability that the card is king is 4/52, then calculate posterior probability P(King|Face), which means the drawn face card is a king card.

Solution:

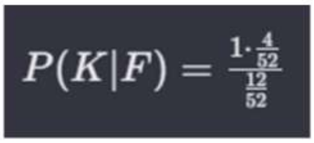
- Let:
    - K: The drawn card is a king.

    - F: The drawn card is a face card (king, queen, or jack).
- The probability of drawing a king, P(K), is given as 4/52, because there are four kings in a standard deck, and there are 52 cards in total.
- The probability of drawing a face card, P(F), is the probability of drawing a king, queen, or jack. There are 4 of each in a standard deck, so P(F)=12/52.
- Now, Bayes' Theorem is given by:



$$P(K|F) = \frac{P(F|K) \cdot P(K)}{P(F)}$$

    - P(F|K) is the probability of drawing a face card given that it is a king. Since every king is also a face card, P(F|K)=1.
    - P(K) is the probability of drawing a king.
    - P(F) is the probability of drawing a face card.
- Now, substitute these values into the formula:

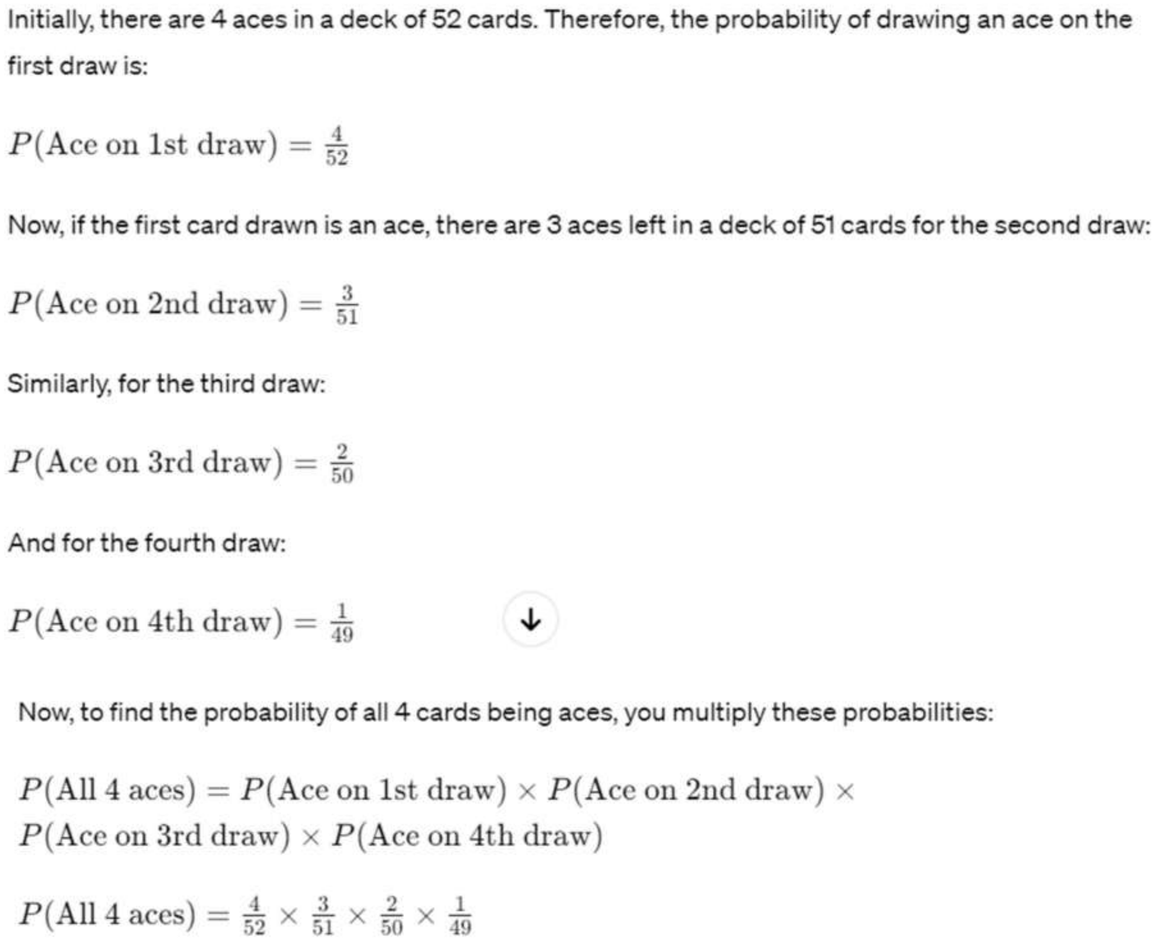


$$P(K|F) = \frac{1 \cdot \frac{4}{52}}{\frac{12}{52}}$$

- So, using Bayes' Theorem, the posterior probability that the drawn face card is a king is 1/3.

Problem:

In a deck of 52 playing cards, 4 cards are drawn without replacement. What is the probability that all 4 cards are ace cards, given that the first card drawn is an ace?



Initially, there are 4 aces in a deck of 52 cards. Therefore, the probability of drawing an ace on the first draw is:

$$P(\text{Ace on 1st draw}) = \frac{4}{52}$$

Now, if the first card drawn is an ace, there are 3 aces left in a deck of 51 cards for the second draw:

$$P(\text{Ace on 2nd draw}) = \frac{3}{51}$$

Similarly, for the third draw:

$$P(\text{Ace on 3rd draw}) = \frac{2}{50}$$

And for the fourth draw:

$$P(\text{Ace on 4th draw}) = \frac{1}{49}$$

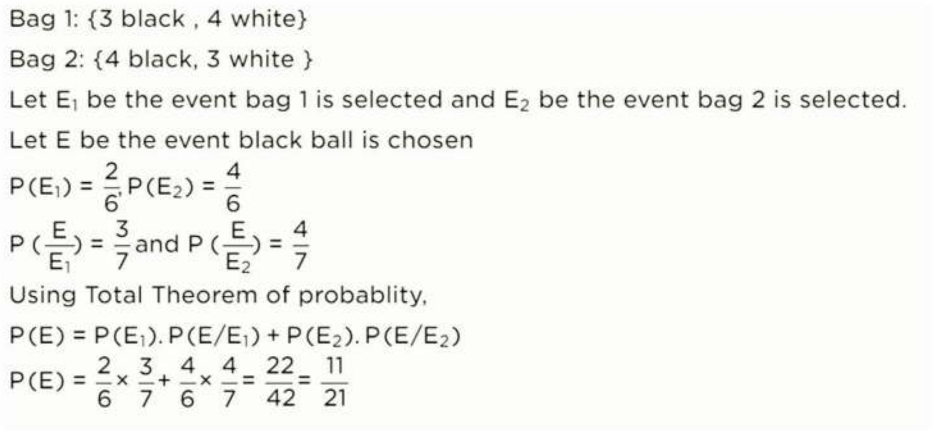
Now, to find the probability of all 4 cards being aces, you multiply these probabilities:

$$P(\text{All 4 aces}) = P(\text{Ace on 1st draw}) \times P(\text{Ace on 2nd draw}) \times P(\text{Ace on 3rd draw}) \times P(\text{Ace on 4th draw})$$

$$P(\text{All 4 aces}) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}$$

Problem:

There are two bags, one of which contains 3 black and 4 white balls while the other contains 4 black and 3 white balls. A die is thrown. If it shows up 1 or 3, a ball is taken from the 1st bag; but it shows up any other number, a ball is chosen from the second bag. Find the probability of choosing a black ball.



Bag 1: {3 black , 4 white}

Bag 2: {4 black, 3 white }

Let E_1 be the event bag 1 is selected and E_2 be the event bag 2 is selected.

Let E be the event black ball is chosen

$P(E_1) = \frac{2}{6}, P(E_2) = \frac{4}{6}$

$P(\frac{E}{E_1}) = \frac{3}{7}$ and $P(\frac{E}{E_2}) = \frac{4}{7}$

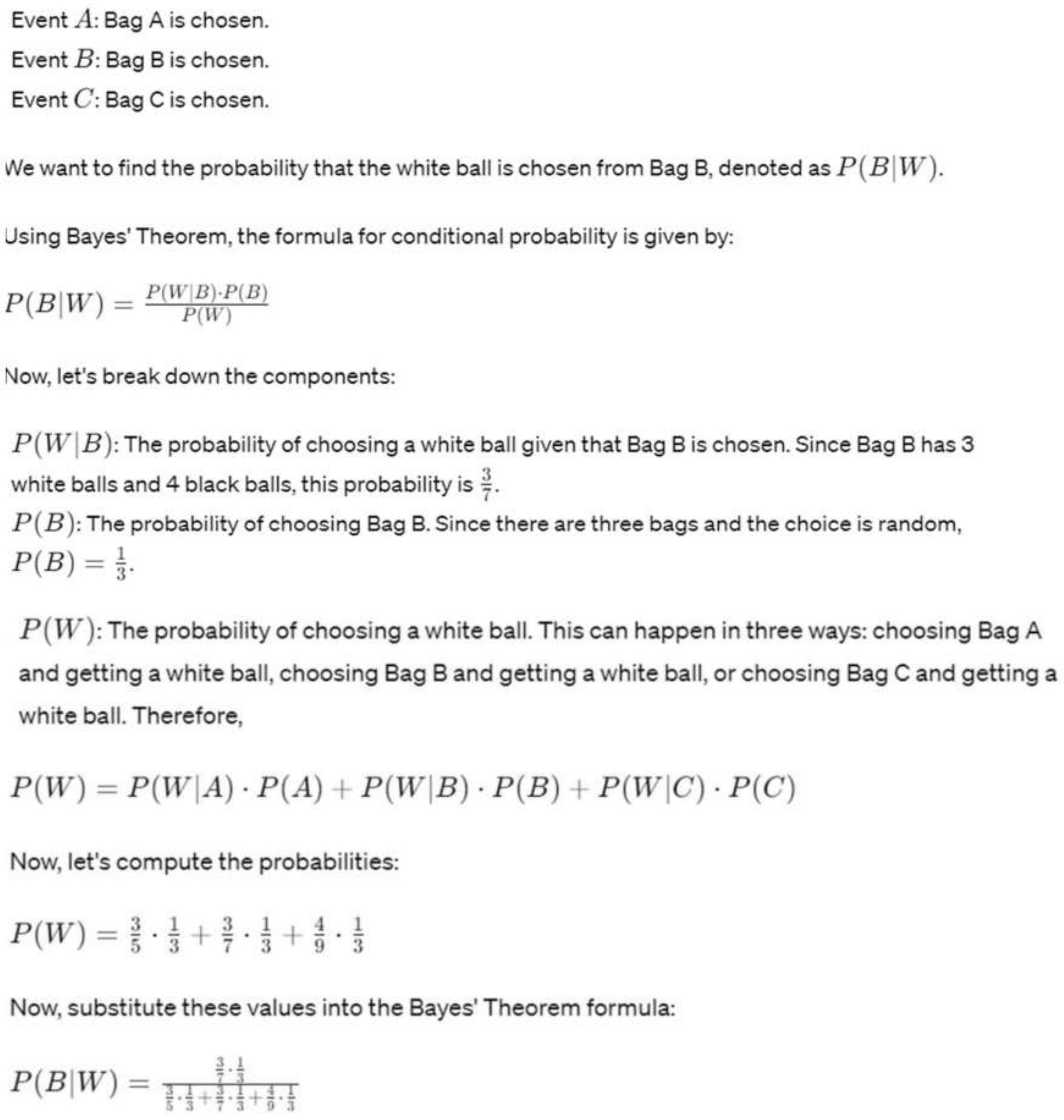
Using Total Theorem of probablity,

$P(E) = P(E_1).P(E/E_1) + P(E_2).P(E/E_2)$

$P(E) = \frac{2}{6} \times \frac{3}{7} + \frac{4}{6} \times \frac{4}{7} = \frac{22}{42} = \frac{11}{21}$

Problem:

Bag A contains 3 white ball and 2 black ball, bag B contains 3 white ball and 4 black ball and bag C contains 4 white ball and 5 black balls. If a white ball is chosen find the probability that it is chosen from bag B.



Event A: Bag A is chosen.

Event B: Bag B is chosen.

Event C: Bag C is chosen.

We want to find the probability that the white ball is chosen from Bag B, denoted as $P(B|W)$.

Using Bayes' Theorem, the formula for conditional probability is given by:

$$P(B|W) = \frac{P(W|B) \cdot P(B)}{P(W)}$$

Now, let's break down the components:

$P(W|B)$: The probability of choosing a white ball given that Bag B is chosen. Since Bag B has 3 white balls and 4 black balls, this probability is $\frac{3}{7}$.

$P(B)$: The probability of choosing Bag B. Since there are three bags and the choice is random, $P(B) = \frac{1}{3}$.

$P(W)$: The probability of choosing a white ball. This can happen in three ways: choosing Bag A and getting a white ball, choosing Bag B and getting a white ball, or choosing Bag C and getting a white ball. Therefore,

$$P(W) = P(W|A) \cdot P(A) + P(W|B) \cdot P(B) + P(W|C) \cdot P(C)$$

Now, let's compute the probabilities:

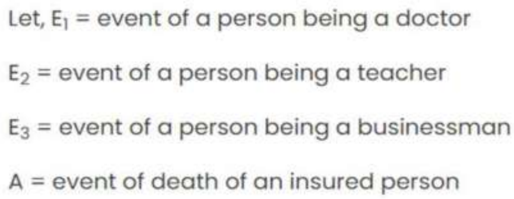
$$P(W) = \frac{3}{5} \cdot \frac{1}{3} + \frac{3}{7} \cdot \frac{1}{3} + \frac{4}{9} \cdot \frac{1}{3}$$

Now, substitute these values into the Bayes' Theorem formula:

$$P(B|W) = \frac{\frac{3}{7} \cdot \frac{1}{3}}{\frac{3}{5} \cdot \frac{1}{3} + \frac{3}{7} \cdot \frac{1}{3} + \frac{4}{9} \cdot \frac{1}{3}}$$

Problem:

An insurance company has insured 4000 doctors, 8000 teachers and 12000 businessmen. The chances of a doctor, teacher and businessman dying before the age of 58 is 0.01, 0.03 and 0.05, respectively. If one of the insured people dies before 58, find the probability that he is a doctor.

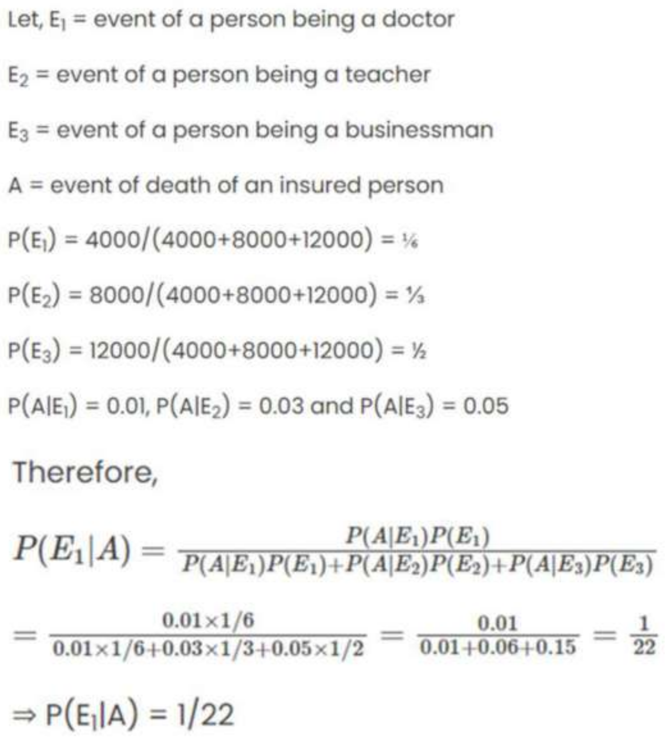


Let, E_1 = event of a person being a doctor

E_2 = event of a person being a teacher

E_3 = event of a person being a businessman

A = event of death of an insured person



Let, E_1 = event of a person being a doctor

E_2 = event of a person being a teacher

E_3 = event of a person being a businessman

A = event of death of an insured person

$P(E_1)$ = 4000/(4000+8000+12000) = ⅙

$P(E_2)$ = 8000/(4000+8000+12000) = ⅓

$P(E_3)$ = 12000/(4000+8000+12000) = ½

$P(A|E_1)$ = 0.01, $P(A|E_2)$ = 0.03 and $P(A|E_3)$ = 0.05

Therefore,

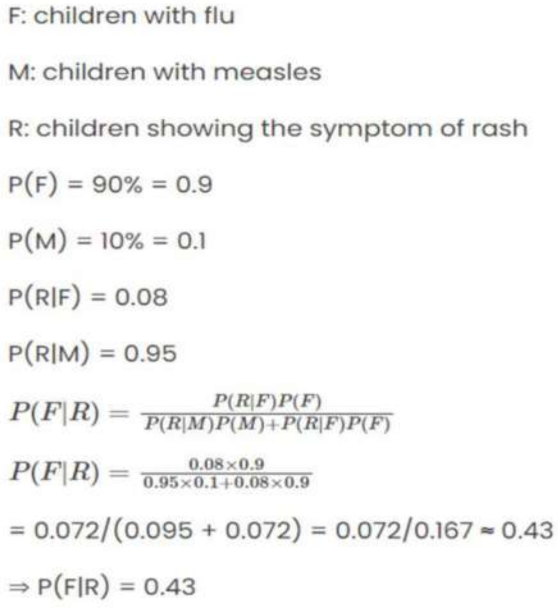
$$P(E_1|A) = \frac{P(A|E_1)P(E_1)}{P(A|E_1)P(E_1)+P(A|E_2)P(E_2)+P(A|E_3)P(E_3)}$$

$$= \frac{0.01\times1/6}{0.01\times1/6+0.03\times1/3+0.05\times1/2} = \frac{0.01}{0.01+0.06+0.15} = \frac{1}{22}$$

⇒ $P(E_1|A)$ = 1/22

Problem:

In a neighbourhood, 90% children were falling sick due to flu and 10% due to measles. The probability of observing rashes for measles is 0.95 and for flu is 0.08. If a child develops rashes, find the child’s probability of having flu.



F: children with flu

M: children with measles

R: children showing the symptom of rash

P(F) = 90% = 0.9

P(M) = 10% = 0.1

P(R|F) = 0.08

P(R|M) = 0.95

$$P(F|R) = \frac{P(R|F)P(F)}{P(R|M)P(M)+P(R|F)P(F)}$$

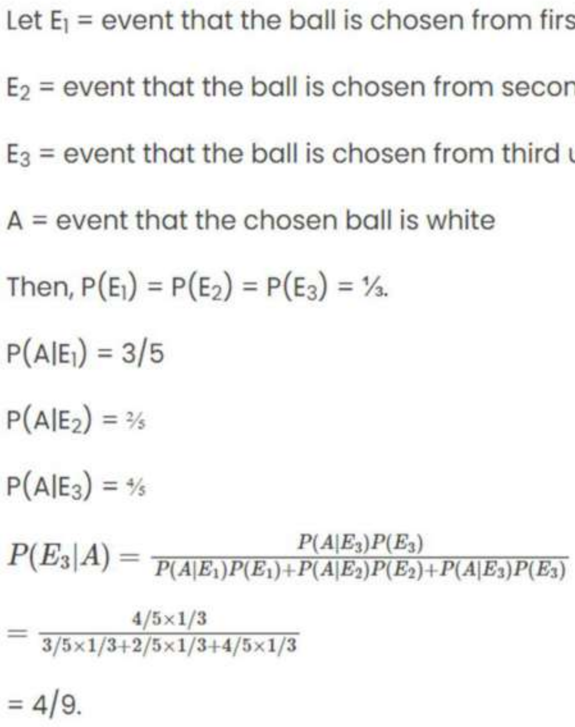
$$P(F|R) = \frac{0.08\times0.9}{0.95\times0.1+0.08\times0.9}$$

= 0.072/(0.095 + 0.072) = 0.072/0.167 ≈ 0.43

⇒ P(F|R) = 0.43

Problem:

Three urns are there containing white and black balls; first urn has 3 white and 2 black balls; second urn has 2 white and 3 black balls and third urn has 4 white and 1 black balls. Without any biasing one urn is chosen from that one ball is chosen randomly which was white. What is probability that it came from the third urn?



Let E_1 = event that the ball is chosen from first urn

E_2 = event that the ball is chosen from second urn

E_3 = event that the ball is chosen from third urn

A = event that the chosen ball is white

Then, $P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$.

$P(A|E_1) = 3/5$

$P(A|E_2) = \frac{2}{5}$

$P(A|E_3) = \frac{4}{5}$

$$P(E_3|A) = \frac{P(A|E_3)P(E_3)}{P(A|E_1)P(E_1)+P(A|E_2)P(E_2)+P(A|E_3)P(E_3)}$$

$$= \frac{4/5\times1/3}{3/5\times1/3+2/5\times1/3+4/5\times1/3}$$

$= 4/9$.