AIH Notes Made by Dare-Marvel

Artificial Intelligence for Healthcare Analytics

1. Importance and Applications of AI and ML in Healthcare
Importance of AI and ML in Healthcare

1. **Enhanced Diagnostics and Imaging**: AI improves accuracy and speed in diagnosing diseases, enabling early detection and reducing human errors.
2. **Predictive Analytics**: AI models predict future health conditions, aiding in preventive care and improving treatment outcomes.
3. **Personalized Medicine**: AI allows for tailored treatments, enhancing therapy effectiveness and reducing adverse effects.
4. **Accelerated Drug Discovery**: AI speeds up the drug discovery process, reducing research time and costs while improving drug development.
5. **Improved Surgical Precision**: AI-driven robots enhance precision in surgeries, reducing complications and recovery times.

6. **24/7 Virtual Health Assistance**: AI-powered virtual assistants provide real-time support to patients, reducing the workload on healthcare professionals.
7. **Optimized Hospital Operations**: AI streamlines administrative tasks, optimizes resource management, and reduces inefficiencies in hospital operations.
8. **Continuous Remote Monitoring**: AI enables real-time patient monitoring, allowing for timely interventions and better chronic disease management.
9. **Informed Clinical Decisions**: AI supports healthcare providers with data-driven insights, improving diagnostic accuracy and treatment decisions.
10. **Enhanced Mental Health Care**: AI tools aid in screening and treating mental health conditions, improving accessibility and early intervention.

Applications of AI and ML in Healthcare

1. **Diagnostics and Imaging**: AI tools analyze medical images (X-rays, MRIs) to detect abnormalities like tumors, fractures, and diseases.
2. **Predictive Analytics**: AI predicts patient risks (e.g., heart attacks, diabetes) and forecasts disease outbreaks for preventive action.
3. **Personalized Medicine**: AI helps customize treatment plans, such as drug recommendations based on genetic and clinical data.
4. **Drug Discovery and Development**: AI identifies drug candidates, predicts interactions, and simulates drug efficacy before clinical trials.
5. **Robotic Surgery**: AI-assisted robotic systems perform minimally invasive surgeries with high precision and reduced human error.
6. **Virtual Health Assistants**: AI-powered chatbots and tools provide symptom checking, preliminary diagnoses, and patient support.

7. **Hospital Operations Management**: AI predicts patient admissions, optimizes bed and staff management, and automates billing and claims.
8. **Remote Patient Monitoring**: AI-integrated wearables monitor vitals and alert healthcare providers for timely interventions.
9. **Clinical Decision Support Systems (CDSS)**: AI-powered systems analyze patient data and provide recommendations for diagnosis and treatment.
10. **Mental Health Care**: AI chatbots and tools screen for mental health disorders and predict potential crises based on behavioral data.

3. Various type of learnings

- Supervised Learning
  - https://www.geeksforgeeks.org/supervised-unsupervised-learning/
  - https://www.superannotate.com/blog/supervised-learning-and-other-machine-learning-tasks

- Unsupervised
  - https://www.geeksforgeeks.org/ml-types-learning-part-2/
  - https://www.javatpoint.com/unsupervised-machine-learning

- Reinforcement Learning
  - https://www.geeksforgeeks.org/what-is-reinforcement-learning/
  -

Differentiates:

1. Supervised Learning vs Unsupervised Learning

No.	Feature	Supervised Learning	Unsupervised Learning
1	**Definition**	Learns from labeled data	Learns from unlabeled data
2	**Goal**	Predict the output based on input	Discover hidden patterns or structure
3	**Input Data**	Labeled (input-output pairs)	Unlabeled (only input data)
4	**Type of Output**	Specific outputs (e.g., classification)	Groups/clusters of similar data
5	**Training Process**	Guided by a "teacher" (labels)	No teacher, self-organizes data
6	**Common Algorithms**	Linear Regression, SVM, Decision Trees	K-Means, PCA, Hierarchical Clustering
7	**Real-World Examples**	Spam detection, image recognition	Customer segmentation, anomaly detection
8	**Complexity**	Generally less complex due to labels	More complex due to no prior labels
9	**Evaluation**	Accuracy measured by comparing predicted labels to true labels	Evaluation often involves manual inspection or domain knowledge
10	**Data Dependency**	Requires a large amount of labeled data	Can work with unlabeled data

2. Supervised Learning vs Reinforcement Learning

No.	Feature	Supervised Learning	Reinforcement Learning
1	**Definition**	Learns from labeled data	Learns by interacting with the environment
2	**Goal**	Predict output based on input	Maximize cumulative reward through actions
3	**Feedback**	Receives direct feedback (labels)	Receives feedback via rewards or penalties
4	**Data Type**	Labeled data	No predefined dataset; learns from experiences
5	**Learning Method**	Learns from examples	Learns by trial and error
6	**Common Algorithms**	Linear Regression, Neural Networks	Q-Learning, Deep Q Networks (DQN)
7	**Real-World Examples**	Email classification, stock price prediction	Game playing (chess, Go), robot control
8	**Environment**	Static dataset	Dynamic environment
9	**Action Dependency**	Predictions do not influence training data	Future actions depend on past actions and rewards
10	**Complexity**	More straightforward, requires labeled data	More complex, involves exploration and exploitation

3. Unsupervised Learning vs Reinforcement Learning

No.	Feature	Unsupervised Learning	Reinforcement Learning
1	**Definition**	Finds patterns in unlabeled data	Learns by interacting with an environment to maximize reward
2	**Goal**	Discover hidden patterns or structure	Learn the best actions to maximize reward
3	**Feedback**	No direct feedback (no labels)	Receives feedback through rewards/penalties
4	**Input Data**	Unlabeled data	No predefined dataset, learns from interaction
5	**Learning Process**	Self-organizes the data	Trial and error learning
6	**Common Algorithms**	Clustering, PCA, K-Means	Q-Learning, SARSA, Deep Q Networks (DQN)
7	**Real-World Examples**	Market segmentation, recommendation systems	Autonomous driving, robotic arm control
8	**Action Dependency**	No concept of actions	Future actions depend on previous actions and rewards
9	**Data Requirement**	Requires large amounts of unlabeled data	Requires interaction with an environment
10	**Adaptability**	Adaptation to new data is limited	Highly adaptive to changing environments

How to build a decision Tree for Boolean Function

(a) $A \wedge \neg B$

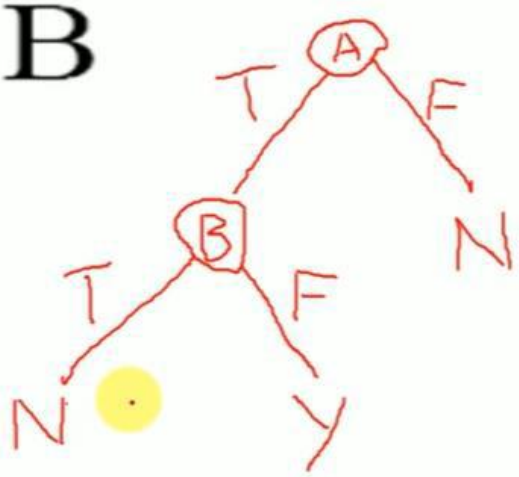
(b) $A \vee [B \wedge C]$

(c) A XOR B

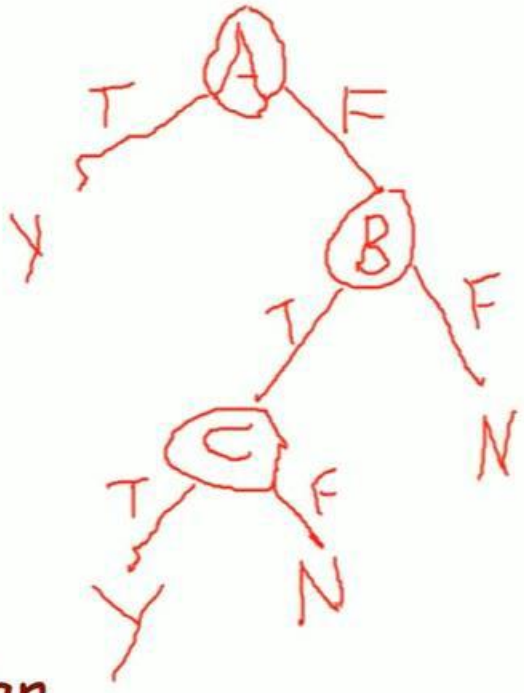
(d) $[A \wedge B] \vee [C \wedge D]$

- Every Variable in Boolean function such as A, B, C etc. has two possibilities that is True and False
- Every Boolean function is either True or False
- If the Boolean function is true we write YES (Y)
- If the Boolean function is False we write NO (N)

(a) $A \wedge \neg B$

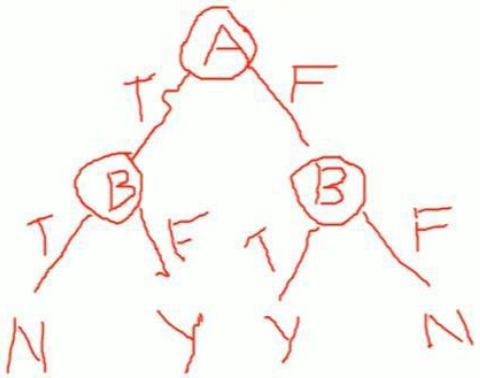


(b) $A \vee [B \wedge C]$



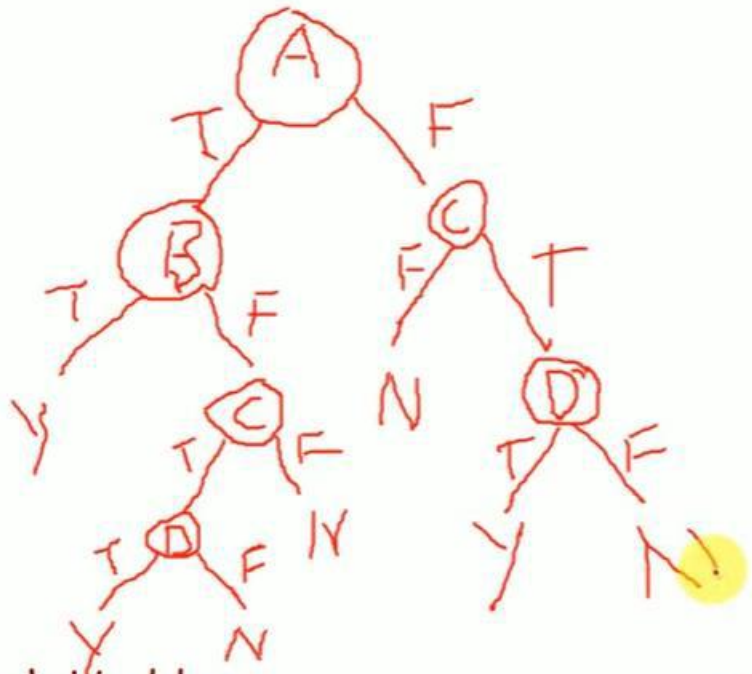
Mahesh Huddar

(c) A XOR B $\equiv (A \wedge \neg B) \vee (\neg A \wedge B)$



A	B	Q
0	0	0
0	1	1
1	0	1
1	1	0

(d) $[A \wedge B] \vee [C \wedge D]$



Mahesh Huddar

vtu

How to find the Entropy and Information Gain in Decision Tree

ENTROPY AND INFORMATION GAIN

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

ENTROPY AND INFORMATION GAIN

ENTROPY MEASURES HOMOGENEITY OF EXAMPLES

- Entropy measures the *impurity* of a collection of examples. It depends from the distribution of the random variable p.
  - S is a collection of training examples
  - p_+ the proportion of positive examples in S
  - p_- the proportion of negative examples in S

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Examples

$Entropy([14+, 0-]) = -14/14\ log_2(14/14) - 0\ log_2(0) = 0$

$Entropy([9+, 5-]) = -9/14\ log_2(9/14) - 5/14\ log_2(5/14) = 0.94$

$Entropy([7+, 7-]) = -7/14\ log_2(7/14) - 7/14\ log_2(7/14) = 1/2 + 1/2 = 1$

INFORMATION GAIN MEASURES THE EXPECTED REDUCTION IN ENTROPY

- Given entropy as a measure of the impurity in a collection of training examples, the ***information gain,*** is simply the expected reduction in entropy caused by partitioning the examples according to an attribute.
- More precisely, the information gain, ***Gain(S, A)*** of ***an*** attribute **A,** relative to a collection of examples ***S,*** is defined as,

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- where ***Values(A)*** is the set of all possible values for attribute A, and S_v, is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S | A(s) = v\}$)

- Given

$$p_1 = 0.1,\ p_2 = 0.2,\ p_3 = 0.3\ and\ p_4 = 0.4$$

- Find the Entropy ?

$p_1 = 0.1,\ p_2 = 0.2,\ p_3 = 0.3\ and\ p_4 = 0.4$

- $Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i)$
- $Entropy = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2) - p_3 * \log_2(p_3) - p_4 * \log_2(p_4)$
- $Entropy = -0.1 * \log_2(0.1) - 0.2 * \log_2(0.2) - 0.3 * \log_2(0.3) - 0.4 * \log_2(0.4)$
- $Entropy = -0.1 * (-3.322) - 0.2 * (-2.322) - 0.3 * (-1.736) - 0.4 * (-1.322)$
- $Entropy = 0.3322 + 0.4644 + 0.5208 + 0.5288$
- $Entropy = 1.8462$

$$\log_2(0.1) = \frac{\log(0.1)}{\log(2)} = \frac{-1}{0.3010} = -3.322$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

$$Values(Wind) = Weak, Strong$$

$$S = [9+, 5-]$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - (8/14)Entropy(S_{Weak}) - (6/14)Entropy(S_{Strong})$$

$$= 0.940 - (8/14)0.811 - (6/14)1.00$$

$$= 0.048$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Outlook

$$Values(Outlook) = Sunny, Overcast, Rain$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-] \qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Overcast} \leftarrow [4+, 0-] \qquad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$$

$$S_{Rain} \leftarrow [3+, 2-] \qquad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook) = Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy(S_{Overcast}) - \frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971 = 0.2464$$

Mahesh Huddar

vtupulse.com

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Temp

$Values\ (Temp) = Hot, Mild, Cool$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{Hot} \leftarrow [2+, 2-]$ $\quad Entropy(S_{Hot}) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} = 1.0$

$S_{Mild} \leftarrow [4+, 2-]$ $\quad Entropy(S_{Mild}) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0.9183$

$S_{Cool} \leftarrow [3+, 1-]$ $\quad Entropy(S_{Cool}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.8113$

$$Gain\ (S, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp) = Entropy(S) - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild}) - \frac{4}{14} Entropy(S_{Cool})$$

$$Gain(S, Temp) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.028$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: H How to find Entropy | Information Gain | Gain in terms of ...

$Values\ (Humidity) = High, Normal$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{High} \leftarrow [3+, 4-]$ $\quad Entropy(S_{High}) = -\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7} = 0.9852$

$S_{Normal} \leftarrow [6+, 1-]$ $\quad Entropy(S_{Normal}) = -\frac{6}{7} log_2 \frac{6}{7} - \frac{1}{7} log_2 \frac{1}{7} = 0.5916$

$$Gain\ (S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Humidity) = Entropy(S) - \frac{7}{14} Entropy(S_{High}) - \frac{7}{14} Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14} 0.9852 - \frac{7}{14} 0.5916 = 0.1516$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Wind

$Values\ (Wind) = Strong, Weak$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$

$S_{Strong} \leftarrow [3+, 3-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [6+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.8113$

$$Gain\ (S, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{Strong}) - \frac{8}{14} Entropy(S_{Weak})$$

$$Gain(S, Wind) = 0.94 - \frac{6}{14}\ 1.0 - \frac{8}{14}\ 0.8113 = 0.0478$$

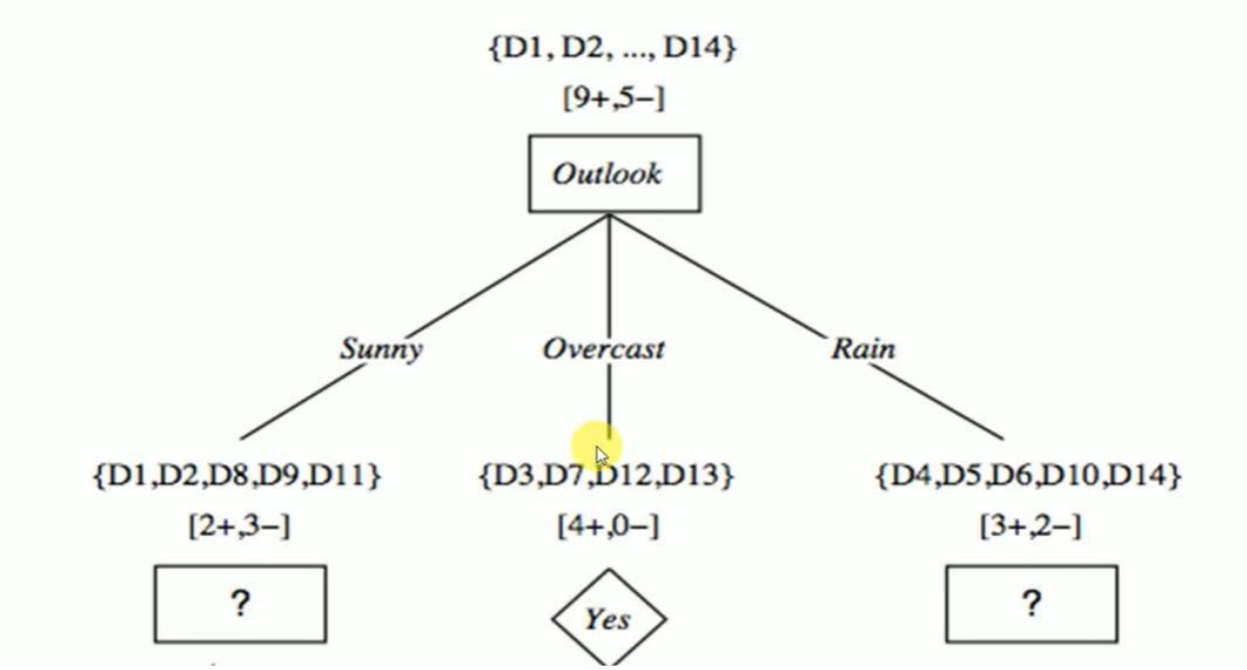
Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	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

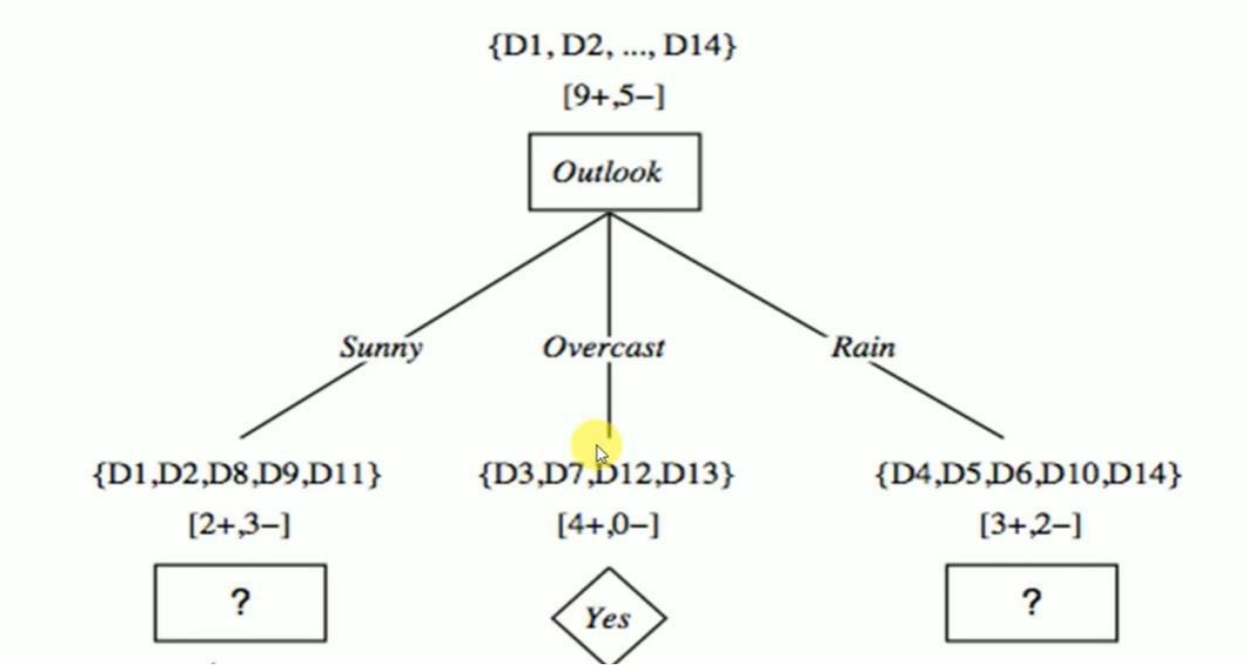
$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Humidity) = 0.1516$$

$$Gain(S, Wind) = 0.0478$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
{D1,D2,D8,D9,D11}
[2+,3−]
?
{D3,D7,D12,D13}
[4+,0−]
Yes
{D4,D5,D6,D10,D14}
[3+,2−]
?


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
{D1,D2,D8,D9,D11}
[2+,3−]
?
{D3,D7,D12,D13}
[4+,0−]
Yes
{D4,D5,D6,D10,D14}
[3+,2−]
?

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Attribute: Temp

$Values\ (Temp) = Hot, Mild, Cool$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 2-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Mild}) = 1.0$

$S_{Cool} \leftarrow [1+, 0-]$ $\quad Entropy(S_{Cool}) = 0.0$

$$Gain\ (S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp) = Entropy(S) - \frac{2}{5}Entropy(S_{Hot}) - \frac{2}{5}Entropy(S_{Mild}) - \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S_{sunny}, Temp) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = 0.570$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Humidity

$Values\ (Humidity) = High, Normal$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{high} \leftarrow [0+, 3-]$ $\quad Entropy(S_{High}) = 0.0$

$S_{Normal} \leftarrow [2+, 0-]$ $\quad Entropy(S_{Normal}) = 0.0$

$$Gain\ (S_{Sunny}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Humidity) = Entropy(S) - \frac{3}{5}Entropy(S_{High}) - \frac{2}{5}Entropy(S_{Normal})$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Wind

$Values\ (Wind) = Strong, Weak$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Strong} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [1+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$

$$Gain\ (S_{Sunny}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

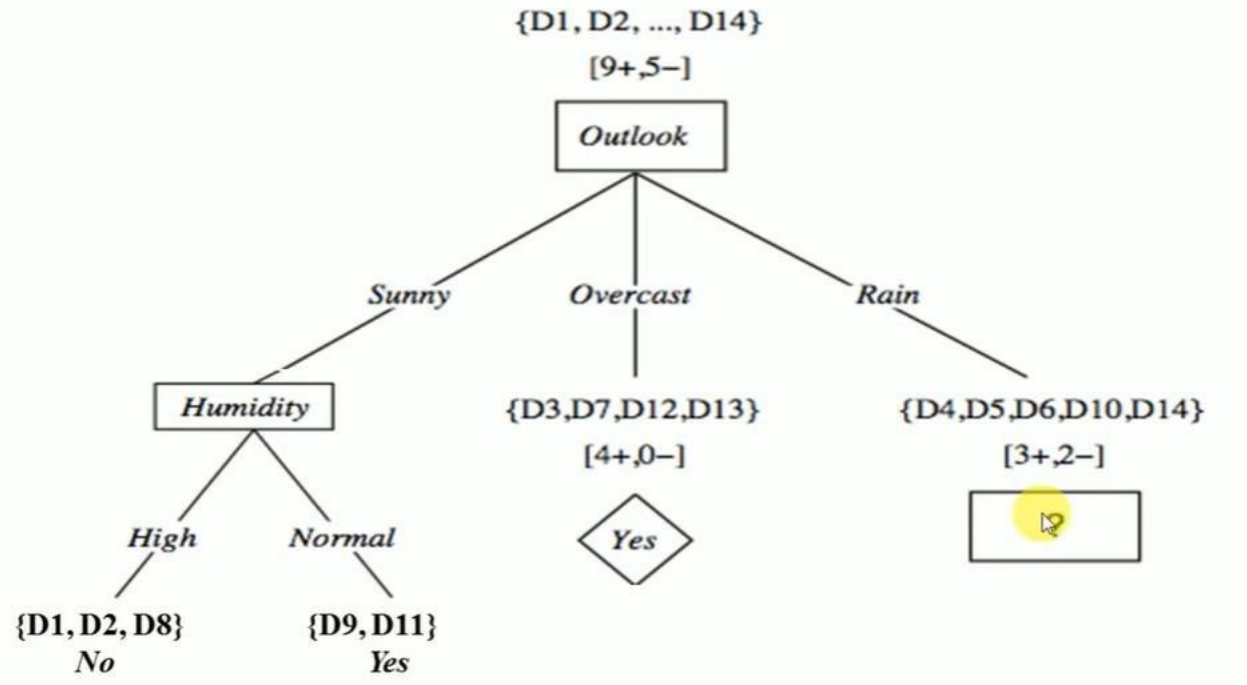
$$Gain(S_{sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

$$Gain(S_{sunny}, Temp) = 0.570$$

$$Gain(S_{sunny}, Humidity) = 0.97$$

$$Gain(S_{sunny}, Wind) = 0.0192$$



Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Attribute: Temp

$$Values\ (Temp) = Hot, Mild, Cool$$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 0-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Mild}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$S_{Cool} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Cool}) = 1.0$

$$Gain\ (S_{Rain}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp) = Entropy(S) - \frac{0}{5}Entropy(S_{Hot}) - \frac{3}{5}Entropy(S_{Mild}) - \frac{2}{5}Entropy(S_{Cool})$$

$$Gain(S_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}0.918 - \frac{2}{5}1.0 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Humidity

$$Values\ (Humidity) = High, Normal$$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{High} \leftarrow [1+, 1-]$ $\quad Entropy(S_{High}) = 1.0$

$S_{Normal} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Normal}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$$Gain\ (S_{Rain}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Humidity) = Entropy(S) - \frac{2}{5}Entropy(S_{High}) - \frac{3}{5}Entropy(S_{Normal})$$

$$Gain(S_{Rain}, Humidity) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Wind

$Values\ (wind) = Strong, Weak$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

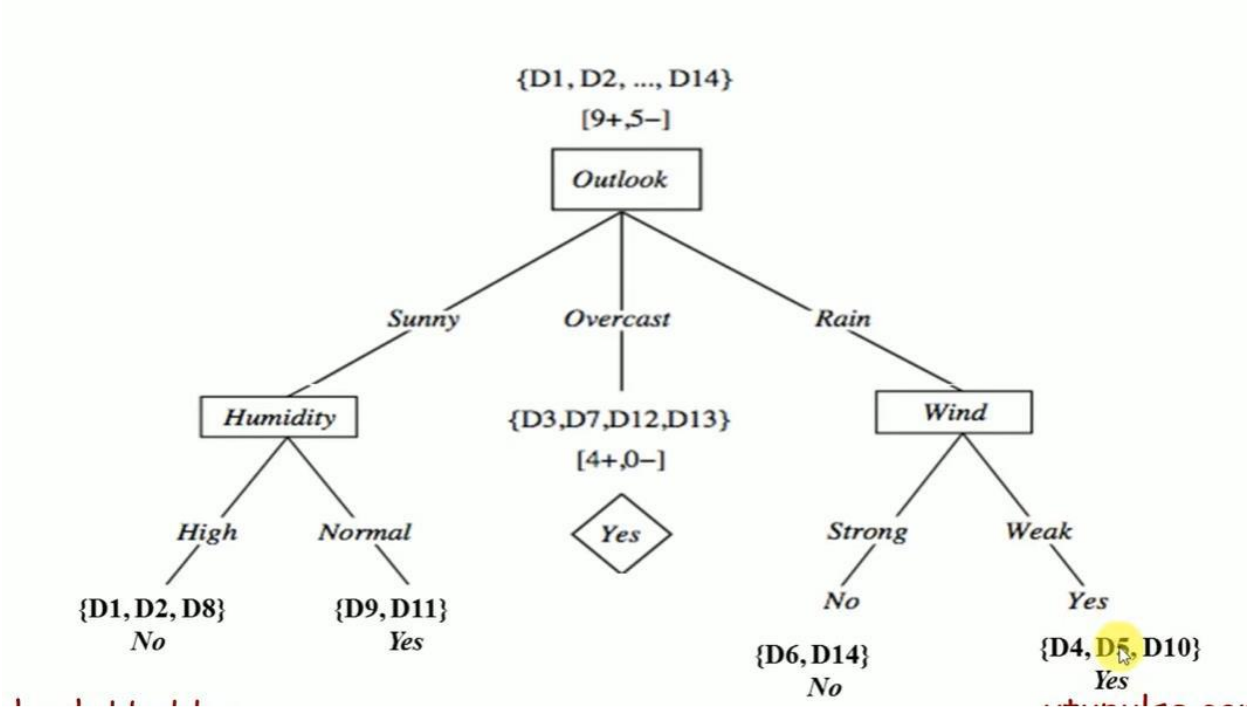
$S_{Strong} \leftarrow [0+, 2-]$ $\quad Entropy(S_{Strong}) = 0.0$

$S_{Weak} \leftarrow [3+, 0-]$ $\quad Entropy(S_{weak}) = 0.0$

$$Gain\ (S_{Rain}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

$$Gain(S_{Rain}, Wind) = 0.97 - \frac{2}{5}\,0.0 - \frac{3}{5}0.0 = 0.97$$



Decision Tree Algorithm – ID3 Solved Example

1. What is the entropy of this collection of training examples with respect to the target function classification?
2. What is the information gain of *a1* and *a2* relative to these training examples?
3. Draw decision tree for the given dataset.

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

Example - 2
Decision Tree Algorithm – ID3
Solved Example

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

Attribute: a1

$Values\,(a1) = T, F$

$S = [3+, 3-]$ $\quad Entropy(S) = 1.0$

$S_T = [2+, 1-]$ $\quad Entropy(S_T) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$S_F \leftarrow [1+, 2-]$ $\quad Entropy(S_F) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$

$$Gain\,(S, a1) = Entropy(S) - \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a1) = Entropy(S) - \frac{3}{6}Entropy(S_T) - \frac{3}{6}Entropy(S_F)$$

$$Gain(S, a1) = 1.0 - \frac{3}{6} * 0.9183 - \frac{3}{6} * 0.9183 = 0.0817$$

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

Attribute: a2

$Values\,(a2) = T, F$

$S = [3+, 3-]$ $\quad Entropy(S) = 1.0$

$S_T = [2+, 2-]$ $\quad Entropy(S_T) = 1.0$

$S_F \leftarrow [1+, 1-]$ $\quad Entropy(S_F) = 1.0$

Example - 2
Decision Tree Algorithm – ID3
Solved Example

$$Gain\,(S, a2) = Entropy(S) - \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a2) = Entropy(S) - \frac{4}{6} Entropy(S_T) - \frac{2}{6} Entropy(S_F)$$

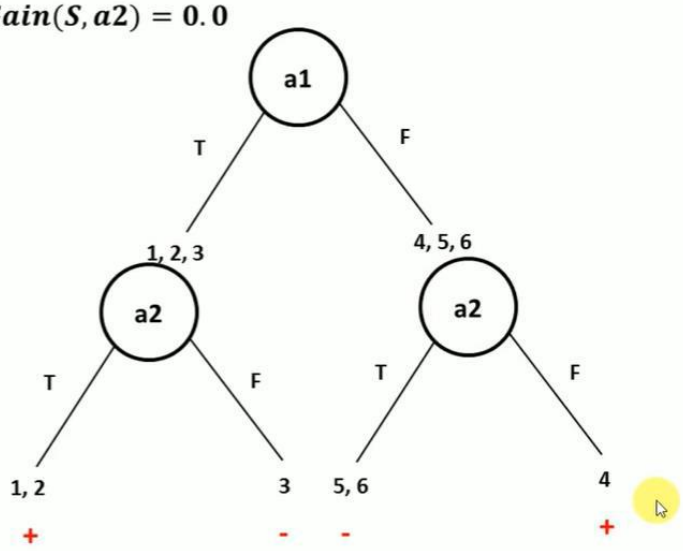
$$Gain(S, a2) = 1.0 - \frac{4}{6} * 1.0 - \frac{2}{6} * 1.0 = 0.0$$

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

Example - 2
Decision Tree Algorithm – ID3
Solved Example

$Gain(S, a1) = 0.0817$ $-$ ***Maximum Gain***

$Gain(S, a2) = 0.0$



KNN Classifier Solved Example - 1

Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

Sepal Length	Sepal Width	Species
5.2	3.1	?

Step 1: Find Distance

$$\text{Distance (Sepal Length, Sepal Width)} = \sqrt{(x-a)^2 + (y-b)^2}$$

$$\text{Distance (Sepal Length, Sepal Width)} = \sqrt{(5.2-5.3)^2 + (3.1-3.7)^2}$$

$$\text{Distance (Sepal Length, Sepal Width)} = 0.608$$

Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608

KNN Classifier Solved Example - 1

Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 2: Find Rank

KNN Classifier Solved Example - 1

Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 3: Find the Nearest Neighbor

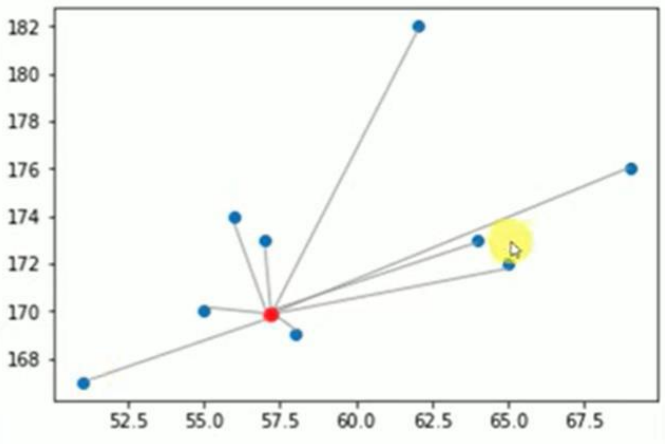
If k = 1 – Setosa

If k = 2 – Setosa

If k = 5 – Setosa

K Nearest Neighbor Algorithm - Solved Example - 2

Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?



Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

THE DISTANCE FORMULA

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$d_1 = \sqrt{(170-162)^2 + (57-51)^2}$

$d_2 = \sqrt{(170-180)^2 + (57-62)^2}$

Height (CM)	Weight (KG)	Class	Distance	Rank
169	58	Normal	1.4	1
170	55	Normal	2	2
173	57	Normal	3	3
174	56	Underweight	4.1	4
167	51	Underweight	6.7	5
173	64	Normal	7.6	6
172	65	Normal	8.2	7
182	62	Normal	13	8
176	69	Normal	13.4	9
170	57	?		

- If K=1, Normal
- If K=2, Normal
- If K=3, Normal
- If K=4, Normal
- If K=5, Normal

K-Nearest Neighbors Algorithm

	Pepper	Ginger	Chilly	Liked
A	True	True	True	False
B	True	False	False	True
C	False	True	True	False
D	False	True	False	True
E	True	False	False	True

Solved Example 3

K-Nearest Neighbors Algorithm Solved Example - 3

- The "Restaurant A" sells burger with optional flavors: Pepper, Ginger and Chilly.
- Every day this week you have tried a burger (A to E) and kept a record of which you liked.
- Using Hamming distance, show how the 3NN classifier with majority voting would classify

{ pepper: false, ginger: true, chilly : true}

K-Nearest Neighbors Algorithm Solved Example - 3

	Pepper	Ginger	Chilly	Liked
A	True	True	True	False
B	True	False	False	True
C	False	True	True	False
D	False	True	False	True
E	True	False	False	True

New Example - Q: pepper: false, ginger: true, chilly : true

K-Nearest Neighbors Algorithm Solved Example - 3

- But How to calculate the distance for attributes with nominal or categorical values.
- Here we can use Hamming distance to find the distance between the categorical values.
- Let x1 and x2 are the attribute values of two instances.
- Then, in hamming distance, if the categorical values are same or matching that is x1 is same as x2 then distance is 0, otherwise 1.
- For example,
- If value of **x1 is blue** and **x2 is also blue** then the distance between x1 and x2 is **0**.
- If value of **x1 is blue** and **x2 is red** then the distance between x1 and x2 is **1**.

K-Nearest Neighbors Algorithm Solved Exampl

	Pepper	Ginger	Chilly	Liked	Distance
A	True	True	True	False	1 + 0 + 0 = 1
B	True	False	False	True	1 + 1 + 1 = 3
C	False	True	True	False	0 + 0 + 0 = 0
D	False	True	False	True	0 + 0 + 1 = 1
E	True	False	False	True	1 + 1 + 1 = 3

New Example - Q: pepper: false, ginger: true, chilly : true

Use Hamming Distance and

K-Nearest Neighbors Algorithm Solved Example - 3

	Pepper	Ginger	Chilly	Liked	Distance	3NN
A	True	True	True	False	1 + 0 + 0 = 1	2
B	True	False	False	True	1 + 1 + 1 = 3	
C	False	True	True	False	0 + 0 + 0 = 0	1
D	False	True	False	True	0 + 0 + 1 = 1	2
E	True	False	False	True	1 + 1 + 1 = 3	

New Example - Q: pepper: false, ginger: true, chilly : true

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P'	**N'**	**All**

- **Classifier Accuracy,** or recognition rate: percentage of test set tuples that are correctly classified

  Accuracy = (TP + TN)/All

- **Error rate:** *1 – accuracy*, or

  Error rate = (FP + FN)/All

- **Class Imbalance Problem**:
  - One class may be *rare*, e.g. fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class
  - **Sensitivity**: True Positive recognition rate
    - **Sensitivity = TP/P**
  - **Specificity**: True Negative recognition rate
    - **Specificity = TN/N**

Classifier Evaluation Metrics: Precision and Recall, and F-measures

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive

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

- **Recall:** completeness – what % of positive tuples did the classifier label as positive?
- Perfect score is 1.0

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

- Inverse relationship between precision & recall
- ***F* measure (*F_1* or *F*-score)**: harmonic mean of precision and recall,

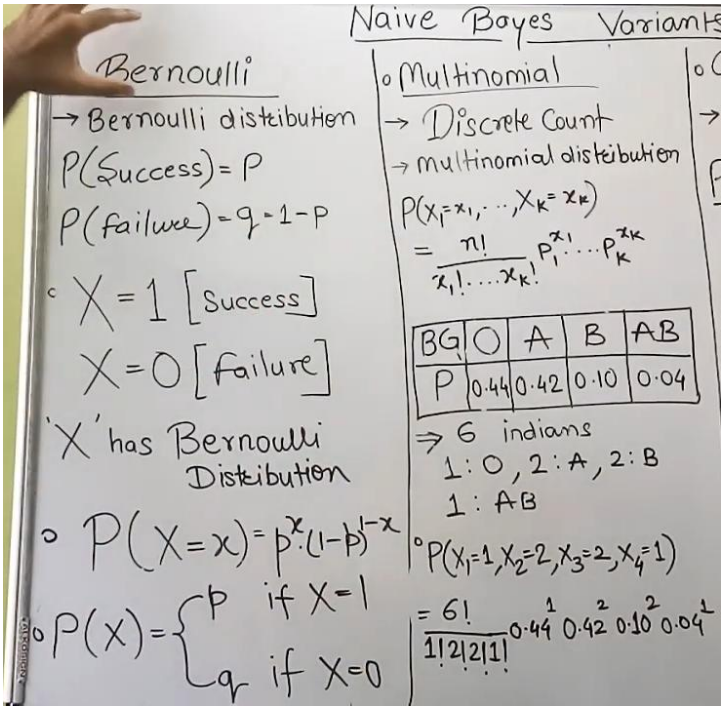
$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- ***F_β*:** weighted measure of precision and recall
  - assigns ß times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

51

Naïve Bayes Variants



Gaussian Naïve Bayes

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

BAYES THEOREM

- Bayes theorem is the cornerstone of Bayesian learning methods because it provides a way to calculate the posterior probability **P(h|D), from**
- **the prior** probability **P(h),**
- **Probability over the data set P(D) and**
- **Current probability P(D(h)**

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

Maximum A Posteriori (MAP) Hypothesis

- The learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis **h ϵ H** given the observed data **D (or at least one of the maximally probable if there** are several).
- Any such maximally probable hypothesis is called a **maximum a posteriori (MAP) hypothesis.**
- **We can determine the MAP hypotheses by using** Bayes theorem to calculate the posterior probability of each candidate hypothesis.

Maximum A Posteriori (MAP) Hypothesis

- More precisely, we will say that h_{MAP} **is a MAP hypothesis provided**

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

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

1. For each hypothesis h in H, calculate the posterior probability

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

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

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

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

- This algorithm may require significant computation, because it applies Bayes theorem to each hypothesis in H to calculate P(h | D) .
  - While this is impractical for large hypothesis spaces,
  - The algorithm is still of interest because it provides a standard against which we may judge the performance of other concept learning algorithms.

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

- Brute Force MAP learning algorithm must specify values for P(h) and P(D|h).
- P(h) and P(D|h) can be chosen to be consistent with the assumptions:
  1. The training data D is noise free.
  2. The target concept c is contained in the hypothesis space H
  3. We have no a priori reason to believe that any hypothesis is more probable than any other.

Brute-Force Bayes Concept Learning

- Given these assumptions, what values should we specify for P(h)?
- Given no prior knowledge that one hypothesis is more likely than another, it is reasonable to assign the same prior probability to every hypothesis h in H.
- Furthermore, because we assume the target concept is contained in H we should require that these prior probabilities sum to 1.
- Together these constraints imply that we should choose

$$P(h) = \frac{1}{|H|} \quad \text{for all } h \text{ in } H$$

Brute-Force Bayes Concept Learning

- What choice shall we make for P(D|h)?
- P(D|h) is the probability of observing the target values D = <d1 . . .dm> for the fixed set of instances <X1 . . . Xm>.
- Since we assume noise-free training data, the probability of observing classification di given h is just 1 if di = h(xi) and 0 if di != h(xi).
- Therefore,

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

Brute-Force Bayes Concept Learning

- **Let us consider the** first step of this algorithm, which uses Bayes theorem to compute the posterior probability **P(h|D) of each hypothesis h given the observed training data D.**

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

- First consider the case where h is inconsistent with the training data D.
- We know that P(D|h) to be 0 when h is inconsistent with D, we have,

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0 \text{ if } h \text{ is inconsistent with } D$$

Brute-Force Bayes Concept Learning

Now consider the case where **h is consistent** with D.

We know that P(Dlh) to be 1 when h is consistent with D, we have

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

$$P(h|D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)}$$

$$= \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}}$$

$$= \frac{1}{|VS_{H,D}|} \text{ if } h \text{ is consistent with } D$$

Brute-Force Bayes Concept Learning

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

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

Does patient have cancer or not?

- Suppose we now observe a new patient for whom the lab test returns a **positive** result.
- Should we diagnose the patient as having cancer or not?

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

$$P(cancer|+) = P(+|cancer) * P(cancer) = 0.98 * 0.008 = 0.0078$$

$$P(\neg cancer|+) = P(+|\neg cancer) * P(\neg cancer) = 0.03 * 0.992 = 0.0298$$

Does patient have cancer or not?

- Suppose we now observe a new patient for whom the lab test returns a **negative** result.
- Should we diagnose the patient as having cancer or not?

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

$$P(cancer|-) = P(-|cancer) * P(cancer) = 0.02 * 0.008 = 0.00016$$

$$P(\neg cancer|-) = P(-|\neg cancer) * P(\neg cancer) = 0.97 * 0.992 = 0.96224$$

$$h_{MAP} = \neg cancer$$

Bayes Theorem - Defective Bulb by Factory A?

- Three factories A, B, C of an electric bulb manufacturing company produce respectively 35%, 35% and 30% of the total output.
- Approximately 1.5%, 1% and 2% of the bulbs produced by these factories are known to be defective.
- If a randomly selected bulb manufactured by the company was found to be defective, what is the probability that the bulb was manufactures in factory A?

- Let A, B, C denote the events that a randomly selected bulb was manufactured in factory A, B, C respectively.
- Let D denote the event that a bulb is defective.
- We have the following data:

$$P(A) = 0.35,\ P(B) = 0.35,\ P(C) = 0.30$$

$$P(D|A) = 0.015,\ P(D|B) = 0.010,\ P(D|C) = 0.020$$

Bayes Theorem - Defective Bulb by Factory A?

- Generalization of Bayes Theorem

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

- Let the sample space be divided into disjoint events $B_1, B_2, \ldots, B_n$ and A be any event.
- Then we have

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

- We are required to find P(A|D).
- By the generalization of the Bayes' theorem, we have:

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

$$= \frac{0.015 \times 0.35}{0.015 \times 0.35 + 0.010 \times 0.35 + 0.020 \times 0.30}$$

$$= 0.356.$$

NAIVE BAYES CLASSIFIER – Example -1

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

$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$

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

NAIVE BAYES CLASSIFIER
Example - 1

$P(PlayTennis = yes) = 9/14 = .64$

$P(PlayTennis = no) = 5/14 = .36$

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5

Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5

Windy		
Strong	3/9	3/5
Weak	6/9	2/5

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\text{argmax}} \; P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\text{argmax}} \; P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

$$v_{NB}(yes) = P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$v_{NB}(no) = P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

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

NAIVE BAYES CLASSIFIER Example – 2

- Estimate conditional probabilities of each attributes {color, legs, height, smelly} for the species classes: {M, H} using the data given in the table.
- Using these probabilities estimate the probability values for the new instance – (Color=Green, legs=2, Height=Tall, and Smelly=No).

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

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

New Instance

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

NAIVE BAYES CLASSIFIER EXAMPLE - 2

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

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

NAIVE BAYES CLASSIFIER - EXAMPLE - 2

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

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

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

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

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

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

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

Hence the new instance belongs to Speices H

NAIVE BAYES CLASSIFIER – Example – 3

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

NAIVE BAYES CLASSIFIER
EXAMPLE - 3

$$p(Yes) = \frac{5}{10} = 0.5$$

$$p(No) = \frac{5}{10} = 0.5$$

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

$$New\ Instance = (Red, SUV, Domestic)$$

$$P(Yes|New\ Instance) = p(Yes) * P(Color = Red|Yes) * P(Type = SUV|Yes) * P(Origin = Domestic|Yes)$$

$$P(Yes|New\ Instance) = \frac{5}{10} * \frac{3}{5} * \frac{1}{5} * \frac{2}{5} = \frac{3}{125} = 0.024$$

$$P(No|New\ Instance) = p(No) * P(Color = Red|No) * P(Type = SUV|No) * P(Origin = Domestic|No)$$

$$P(No|New\ Instance) = \frac{5}{10} * \frac{2}{5} * \frac{3}{5} * \frac{3}{5} = \frac{9}{125} = 0.072$$

$$P(No|New\ Instance) > P(Yes|New\ Instance)$$

NAIVE BAYES CLASSIFIER EXAMPLE – 4

Consider a football game between two rival teams, say team A and team B. Suppose team A wins 65% of the time and team B wins the remaining matches. Among the games won by team A, only 35% of them comes from playing at team B's football field. On the other hand, 75% of the victories for team B are obtained while playing at home.

1. If team B is to host the next match between the two teams, what is the probability that it will emerge as the winner?
2. If team B is to host the next match between the two teams, who will emerge as the winner?

Solution:

Y – Winning football match
X – Hosting football match

Probability that team A wins is $P(Y_A) = 0.65$.

Probability that team B wins is $P(Y_B) = 1 - P(Y_A) = 0.35$

Probability that team B hosted the match it had won is $P(X_B|Y_B) = 0.75$.

Probability that team B hosted the match won by team A is $P(X_B|Y_A) = 0.35$.

1. If team B is to host the next match between the two teams, what is the probability that it will emerge as the winner?

Solution:

$$P(Y_B|X_B) = \frac{P(X_B|Y_B)\times P(Y_B)}{P(X_B)}$$

$$= \frac{P(X_B|Y_B)\times P(Y_B)}{P(X_B|Y_A)P(Y_A) + P(X_B|Y_B)P(Y_B)}$$

$$= \frac{0.75 \times 0.35}{(0.35 \times 0.65 + 0.75 \times 0.35)}$$

$$= 0.5357$$

Probability that team A wins is $P(Y_A) = 0.65$.

Probability that team B wins is $P(Y_B) = 1 - P(Y_A) = 0.35$

Probability that team B hosted the match it had won is $P(X_B|Y_B) = 0.75$.

Probability that team B hosted the match won by team A is $P(X_B|Y_A) = 0.35$.

2. If team B is to host the next match between the two teams, who will emerge as the winner?

Solution:

$$P(Y_A|X_B) = \frac{P(X_B|Y_A)\times P(Y_A)}{P(X_B)}$$

$$= \frac{P(X_B|Y_A)\times P(Y_A)}{P(X_B|Y_A)P(Y_A) + P(X_B|Y_B)P(Y_B)}$$

$$= \frac{0.35 \times 0.65}{(0.35 \times 0.65 + 0.75 \times 0.35)}$$

$$= 0.4642$$

Probability that team A wins is $P(Y_A) = 0.65$.

Probability that team B wins is $P(Y_B) = 1 - P(Y_A) = 0.35$

Probability that team B hosted the match it had won is $P(X_B|Y_B) = 0.75$.

Probability that team B hosted the match won by team A is $P(X_B|Y_A) = 0.35$.

$$P(Y_B|X_B) = 0.5357$$

Naïve Bayes Classifier – Solved Example 5

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

How to Compute

- **Conditional Probabilities**
- **Predict the Label for the new instance**

Consider the dataset given in the table and

1. Estimate the conditional probabilities of P(A|+), P(B|+), P(C|+), P(A|-), P(B|-) and P(C|-)
2. Use the conditional probability estimates and predict the class label for the test sample P(A=0, B=1, C=0)

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

A	0	1
+	2/5	3/5
-	3/5	2/5

B	0	1
+	4/5	1/5
-	3/5	2/5

$P(A=0|+)=\frac{2}{5}$

C	0	1
+	1/5	4/5
-	0/5	5/5

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

A	0	1
+	2/5	3/5
-	3/5	2/5

B	0	1
+	4/5	1/5
-	3/5	2/5

C	0	1
+	1/5	4/5
-	0/5	5/5

$$New => P(A=0, B=1, C=0)$$

$$P(+|New) = \frac{P(+) * P(A=0|+) * P(B=1|+) * P(C=0|+)}{P(A=0, B=1, C=0)}$$

$$= 0.5 * \frac{2}{5} * \frac{1}{5} * \frac{1}{5} = 0.008$$

$$P(-|New) = \frac{P(-) * P(A=0|-) * P(B=1|-) * P(C=0|-)}{P(A=0, B=1, C=0)}$$

$$= 0.5 * \frac{3}{5} * \frac{2}{5} * \frac{0}{5} = 0.0$$

$$P(+|New) > P(-|New)$$

- **Solved Example Text Analytics or Text Classification using Naïve Bayes Classifier by Mahesh Huddar**

Naïve Bayes Model – Text Classification Example

- Dataset for the text classification with training and test data is given below.
- The goal is to classify the test data into the right class as h or —h (read as not h).

	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	?

Naïve Bayes Model – Text Classification Example

- The probability of the document 'd' being in class 'c' is computed as follows,

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

- Where, $p(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c.

Naïve Bayes Model – Text Classification Example

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

- The prior probabilities of a document being classified using the six documents are,

$$P(h) = \frac{4}{6} = \frac{2}{3}$$

$$and$$

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

- That is there is $\frac{2}{3}$ prior probability that a document will be classified as h and $\frac{1}{3}$ probability of not h.

Naïve Bayes Model – Text Classification Example

- The conditional probability for each term is the relative frequency of the term occurring in each class of the documents 'h class' and 'not h class'.

Testing Example:
Love Pain Joy Love Kick = ?

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

Naïve Bayes Model – Text Classification Example

Testing Example:
Love Pain Joy Love Kick = ?

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

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

$$P(h|d_7) = P(h) * (P(Love|h) * (P(Love|h) * P(Pain|h) * P(Joy|h) * P(Kick|h)$$
$$= (2/3) * (5/19) * (5/19) * (1/19) * (5/19) * (1/19) = \mathbf{-0.0000067}$$

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

Class label for testing example is: **No**

Gaussian Naïve Bayes

Gaussian Naive Bayes Algorithm – Solved Example 6

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

Based on the following data determine the gender of a person having height 6 ft., weight 130 lbs, and foot size 8 inch. (use Naive Bayes algorithm).

Gaussian Naive Bayes Algorithm – Solved Example 6

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$P(Male) = 4/8 = 0.5$

$P(Female) = 4/8 = 0.5$

Male:

$$Mean\ (Height) = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$Variance\ (Height) = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(6-5.855)^2+(5.92-5.855)^2+(5.58-5.855)^2+(5.92-5.855)^2}{4-1}$$

$$= 0.035055$$

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

Subscribe to Mahesh Huddar Visit: vtupulse.com

Gaussian Naive Bayes Algorithm – Solved Example 6

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$P(Male) = 4/8 = 0.5$

$P(Female) = 4/8 = 0.5$

$$Posterior\ (Male) = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{Evidence}$$

$$Posterior\ (Female) = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{Evidence}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$

$P(W|M) = 5.9881e^{-6}$

$P(FS|M) = 1.3112e^{-3}$

$P(H|F) = 2.2346e^{-1}$

$P(W|F) = 1.6789e^{-2}$

$P(FS|F) = 2.8669e^{-1}$

Gaussian Naive Bayes Algorithm – Solved Example 6

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$P(Male) = 4/8 = 0.5$

$P(Female) = 4/8 = 0.5$

$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$

$P(W|M) = 5.9881e^{-6}$

$P(FS|M) = 1.3112e^{-3}$

$P(H|F) = 2.2346e^{-1}$

$P(W|F) = 1.6789e^{-2}$

$P(FS|F) = 2.8669e^{-1}$

$$Posterior\ (Male) = \frac{P(M)*P(H|M)*P(W|M)*P(FS|M)}{Evidence} = 0.5 * 1.5789 * 5.9881e^{-6} * 1.3112e^{-3} = 6.1984e^{-9}$$

$$Posterior\ (Female) = \frac{P(F)*P(H|F)*P(W|F)*P(FS|F)}{Evidence} = 0.5 * 2.2346e^{-1} * 1.6789e^{-2} * 2.8669e^{-1} = 5.377e^{-4}$$

Subscribe to Mahesh Huddar Visit: vtupulse.com

Simple Linear Regression

* Equation of Regression line

$$Y' = aX + b$$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} =$$

$$b = \frac{1}{n}(\Sigma y - a\Sigma x) =$$

=>Expendeture of business (in thousands) for every year is shown in table below:

X (Year)	1	2	3	4	5
Y (Expendeture)	12	19	29	37	45

- Find the expenditure of company in 6th year using line of a model.

Sol:

Sr. No	X	Y
1	1	12
2	2	19
3	3	29
4	4	37
5	5	45

=> Expendeture of business (in thousands) for every year is shown in table below:

X (Year)	1	2	3	4	5
Y (Expendeture)	12	19	29	37	45

- Find the expenditure of company in 6th year using line of a model.

Sol:

Sr. No	X	Y	XY	X^2
1	1	12	12	1
2	2	19	38	4
3	3	29	87	9
4	4	37	148	16
5	5	45	225	25
	$\Sigma X = 15$	$\Sigma Y = 142$	$\Sigma XY = 510$	$\Sigma X^2 = 55$

* Equation of Regression line

$$Y' = aX + b$$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5 \times 510 - 15 \times 142}{5 \times 55 - (15)^2} = 8.4$$

$$b = \frac{1}{n}(\Sigma y - a\Sigma x) = \frac{1}{5}(142 - 8.4 \times 15) = 3.2$$

Equation of line:

$$Y' = 8.4X + 3.2$$

* for the sixth month

$$Y' = 8.4 \times 6 + 3.2$$

$$= 53.6 \text{ Thousands}$$

Q: Consider the following & obtain the Multiple Linear Equation

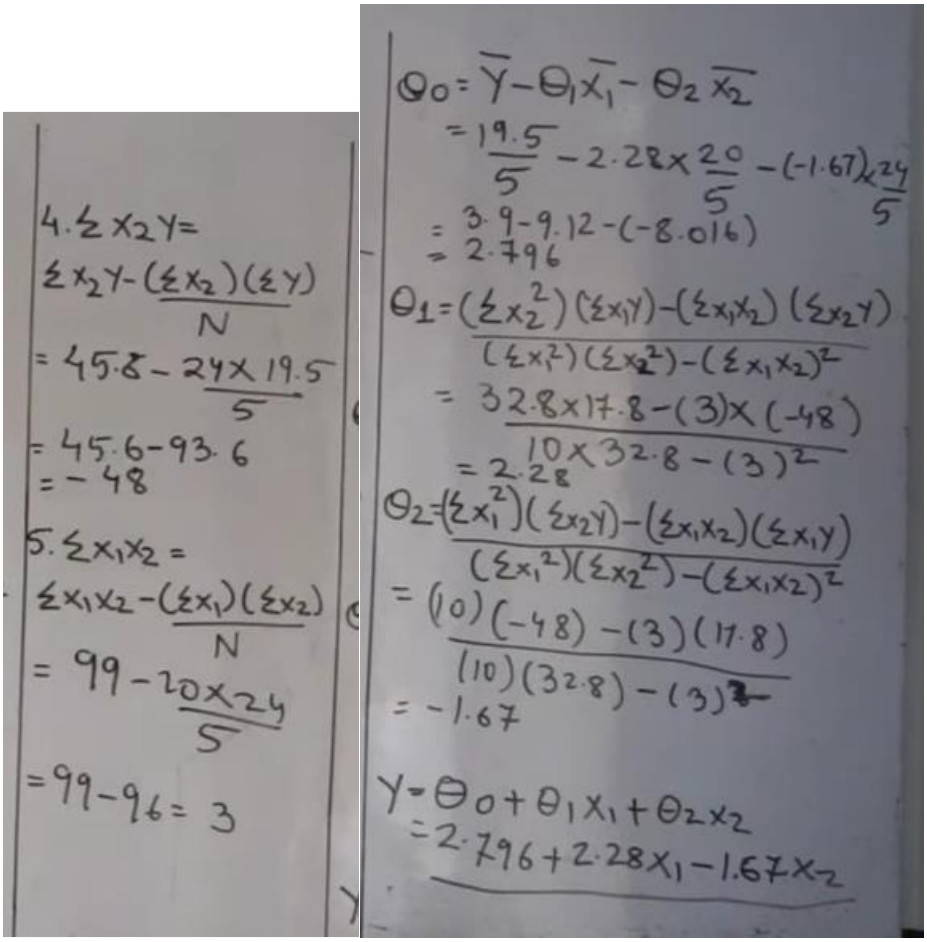
Sr.No	Y	X_1	X_2	$(X_1)^2$	$(X_2)^2$	X_1Y	X_2Y	X_1X_2
1	-3.7	3	8	9	64	-11.1	-29.6	24
2	3.5	4	5	16	25	14	17.5	20
3	2.5	5	7	25	49	12.5	17.5	35
4	11.5	6	3	36	9	69	34.5	18
5	5.7	2	1	4	1	11.4	5.7	2
	$\Sigma Y = 19.5$	$\Sigma X_1 = 20$	$\Sigma X_2 = 24$	$\Sigma(X_1)^2 = 90$	$\Sigma(X_2)^2 = 148$	$\Sigma X_1Y = 95.8$	$\Sigma X_2Y = 45.6$	$\Sigma X_1X_2 = 99$

1. $\Sigma x_1^2 = \Sigma(X_1)^2 - \frac{(\Sigma X_1)^2}{N}$

$= 90 - \frac{(20)^2}{5} = 90 - 80 = 10$

2. $\Sigma x_2^2 = \Sigma(X_2)^2 - \frac{(\Sigma X_2)^2}{N} = 148 - \frac{(24)^2}{5} = 32.8$

3. $\Sigma x_1 y = \Sigma X_1 Y - \frac{(\Sigma X_1)(\Sigma Y)}{N} = 95.8 - \frac{20 \times 19.5}{5} = 17.8$



Types of Regression Models

- Regression modeling is a process of determining a relationship between one or more independent variables and one dependent or output variable.
- **Examples:**
  1. Predicting the height of a person given the age of the person.
  2. Predicting the price of the car given the car model, year of manufacturing, mileage, engine capacity, etc.

- Based on the type of functions used to represent the relationship between the dependent or output variable and independent variables, the regression models are categorized into four types. The regression models are,
  1. Simple Linear Regression
  2. Multiple Regression
  3. Polynomial Regression
  4. Logistic Regression

1. Simple Linear Regression

- Assume that there is only one independent variable x. If the relationship between x (independent variable) and y (dependent or output variable) is modeled by the relation,

$$y = a + bx$$

- then the regression model is called a linear regression model.

2. Multiple Regression

- Assume that there are multiple independent variables *say* x_1, x_2, ...x_n. If the relationship between independent variables x and dependent or output variable y is modeled by the relation,

$$y=a_0+a_1*x_1+a_2*x_2+.....+a_n*x_n$$

- then the regression model is called a multiple regression model.

3. Polynomial regression

- Assume that there is only one independent variable x. If the relationship between independent variables x and dependent or output variable y is modeled by the relation,

$$y=a_0+a_1*x+a_2*x^2+.....+a_n*x^n$$

- for some positive integer n>1, then we have a polynomial regression.

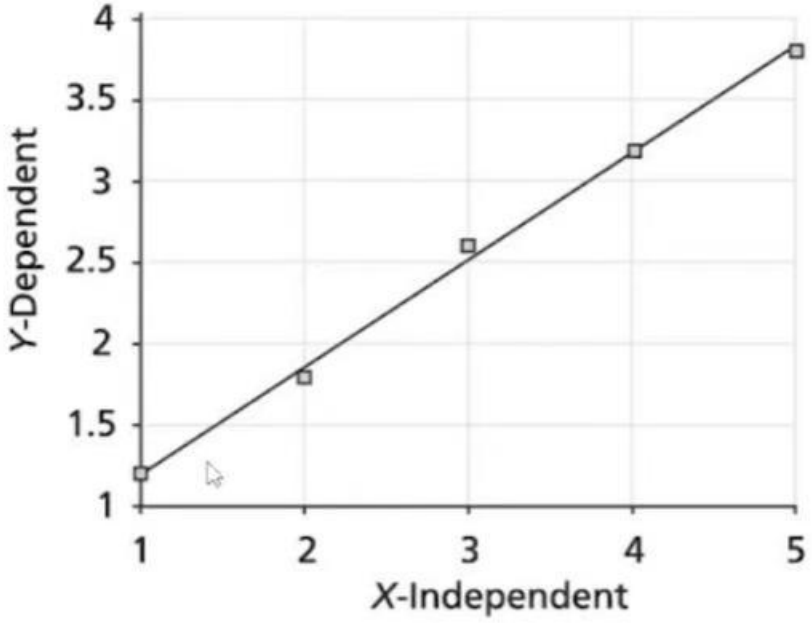
4. Logistic Regression

- Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature.

Linear Regression Algorithm

- Let us consider an example where the five weeks' sales data (in Thousands) is given as shown in Table.
- Apply linear regression technique to predict the 7th and 12th week sales.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8



- Linear regression equation is given by
- $y = a_0 + a_1 * x + e$
- $where$
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

- Here, there are 5 items, i.e., i = 1, 2, 3, 4, 5.

	x_i (Week)	y_j (Sales in Thousands)	x_i^2	$x_i * y_j$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Average	$\bar{x} = 3$	$\bar{y} = 2.52$	$\overline{x^2} = 11$	$\overline{xy} = 8.88$

- $\bar{x} = 3$ $\bar{y} = 2.52$ $\overline{x^2} = 11$ $\overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy})-(\bar{x})(\bar{y})}{\overline{x^2}-\bar{x}^2} = \frac{8.88-3*2.52}{11-3^2} = 0.66$
- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$
- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$

- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$
- The predicted 7th week sale (when x = 7) is,
- **y = 0.54 + 0.66 x 7 = 5.16**
- the predicted 12th week sale (when x = 12) is,
- **y = 0.54 + 0.66 x 12 = 8.46**

Linear Regression – Solved Example – Matrix Method

- Here, the independent variable X is be given as:
- X^T = [1 2 3 4]
- The dependent variable is given as follows:
- Y^T = [1 3 4 8]
- The data can be given in matrix form as follows:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad Y = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 8 \end{pmatrix}$$

x_i (Week)	y_j (Sales in Thousands)
1	1
2	3
3	4
4	8

The first column can be used for setting bias.

- The regression is given as:

$$a = ((X^T X)^{-1} X^T) Y$$

- The computation order of this equation is shown step by step as:

1. Computation of $(X^T X) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \times \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}$

2. Computation of matrix inverse of $(X^T X)^{-1} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}^{-1} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix}$

3. Computation of $((X^T X)^{-1} X^T) = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{pmatrix}$

4. Finally, $((X^T X)^{-1} X^T) Y = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \\ 4 \\ 8 \end{pmatrix} = \begin{pmatrix} -1.5 \\ 2.2 \end{pmatrix} \begin{pmatrix} Intercept \\ slope \end{pmatrix}$

- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = -1.5 + 2.2 * x$
- The predicted 5th week sale (when x = 5) is,
- $y = -1.5 + 2.2 * 5 = 9.5$

Multiple Linear Regression

Multiple Linear Regression – Solved Example

- In linear regression model we have one dependent and one independent variable.
- Multiple regression model involves multiple predictors or independent variables and one dependent variable.
- This is an extension of the linear regression problem.
- The multiple regression of two variables x1 and x2 is given as follows:

$$y = f(x_1, x_2)$$

$$y = a_0 + a_1x_1 + a_2x_2$$

- In general, this is given for 'n' independent variables as:

$$y = f(x_1, x_2, \ldots . x_n)$$

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + \varepsilon$$

Logistic Regression Algorithm

- Linear regression predicts the numerical response but is not suitable for predicting the categorical variables.
- When categorical variables are involved, it is called classification problem.
- Logistic regression is suitable for binary classification problem.

Logistic Regression – Algorithm & Solved Example

For example, the following scenarios are instances of predicting categorical variables.

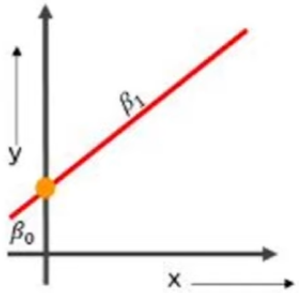
1. Is the mail spam or not spam? The answer is yes or no. Thus, categorical dependent variable is a binary response of yes or no.
2. If the student should be admitted or not is based on entrance examination marks. Here, categorical variable response is admitted or not.
3. The student being pass or fail is based on marks secured.

How Does the Logistic Regression Algorithm Work?

- Consider the following example:
- An organization wants to determine an employee's salary increase based on their performance.
- For this purpose, a linear regression algorithm will help them decide.
- Plotting a regression line by considering the employee's performance as the independent variable, and the salary increase as the dependent variable will make their task easier.

- Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance?
- The above linear graph won't be suitable in this case.
- As such, we clip the line at zero and one, and convert it into a sigmoid curve (S curve).
- Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

- To understand logistic regression, let's go over the odds of success.
- $Odds\ (\theta) = \frac{Probability\ of\ an\ event\ happening}{Probability\ of\ an\ event\ not\ happening}$
- $Odds\ (\theta) = \frac{p}{1-p}$
- The values of odds range from **zero to ∞** and the values of probability lies between **zero and one**.



- Consider the equation of a straight line:
- $\boldsymbol{y = \beta 0 + \beta 1 * x}$
- Now to predict the odds of success, we take log on odds formula:

$$\log\left(\frac{p(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x$$

- Exponentiating both the sides, we have:

$$e^{\ln(x)} = x$$

$$e^{ln}\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

$$\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

- $Let\ Y = e^{\beta 0+\beta 1 * x}$
- $Then \quad \frac{p(x)}{1-p(x)} = Y$
- $p(x) = Y(1 - p(x))$
- $p(x) = Y - Y(p(x))$
- $p(x) + Y(p(x)) = Y$
- $p(x)(1+Y) = Y$
- $p(x) = \frac{Y}{1+Y}$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

The equation of the sigmoid function is:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

- The sigmoid curve obtained from the above equation is as follows:

- The student dataset has entrance mark based on the historic data of those who are selected or not selected.
- Based on the logistic regression, the values of the learnt parameters are $\beta_0 = 1\ and\ \beta_1 = 8.$
- Assuming marks of x = 60, compute the resultant class.

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

$$\beta_0+\beta_1 x = 481$$

$$p(x) = \frac{1}{1+e^{-481}} = 0.44$$

- If we assume the threshold value as 0.5, then it is observed that 0.44 < 0.5, therefore, the candidate with marks 60 is not selected.

- The dataset of pass or fail in an exam of 5 students is given in the table.
- Use logistic regression as classifier to answer the following questions.

1. Calculate the probability of pass for the student who studied 33 hours.
2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

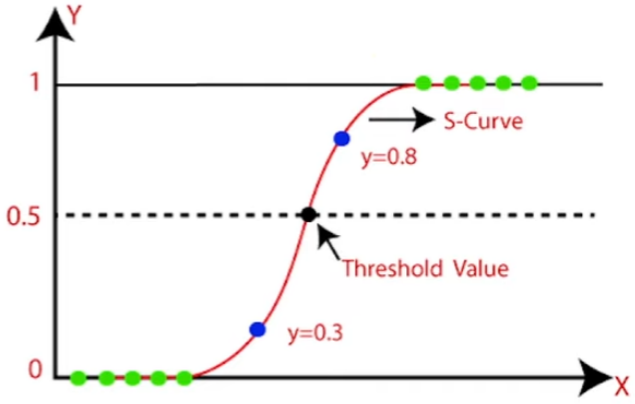
Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(odds) = -64 + 2 * hours$$

- We use Sigmoid Function in logistic regression

- $$s(x) = \frac{1}{1+e^{-x}}$$



1. Calculate the probability of pass for the student who studied 33 hours.

- $p = \frac{1}{1+e^{-z}}$ $\qquad s(x) = \frac{1}{1+e^{-x}}$
- $z = -64 + 2 * 33 = -64 + 66 = 2$
- $p = \frac{1}{1+e^{-2}} = 0.88$
- That is, if student studies 33 hours, then there is **88% chance** that the student will pass the exam

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

$$\log(odds) = z = -64 + 2 * hours$$

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- $\boldsymbol{p = \frac{1}{1+e^{-z}} = 0.95}$
- $0.95 * (1 + e^{-z}) = 1$
- $0.95 * e^{-z} = 1 - 0.95$
- $e^{-z} = \frac{0.05}{0.95} = 0.0526$
- $\ln(e^{-z}) = \ln(0.0526)$

$\boldsymbol{\ln(e^x) = x}$

$-z = ln(0.0526) = -2.94$

$\boldsymbol{z = 2.94}$

- $z = 2.94$
- $\log(odds) = z = -64 + 2 * hours$
- $2.94 = -64 + 2 * hours$
- $2 * hours = 2.94 + 64$
- $2 * hours = 66.94$
- $hours = \frac{66.94}{2}$
- $\boldsymbol{hours = 33.47\ Hours}$

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- The student should study **at least 33.47 hours,** so that he will pass the exam with more than 95% probability

Linear Regression	Logistic Regression
Used to solve regression problems	Used to solve classification problems
The response variables are continuous in nature	The response variable is categorical in nature
It helps estimate the dependent variable when there is a change in the independent variable	It helps to calculate the possibility of a particular event taking place
It is a straight line	It is an S-curve (S = Sigmoid)

- **Data Analysis and Visualization Tools: Comprehensive Analysis**
- **Programming Languages and Tools Overview**
- **1. Python**
- **Capabilities**
  - **General-purpose programming language**
  - Extremely versatile for data analysis, machine learning, and scientific computing
  - Extensive libraries for data manipulation and analysis
- **Advantages**
  - Open-source and free
  - Massive community support
  - Rich ecosystem of libraries:
    - Pandas for data manipulation
    - NumPy for numerical computing
    - Matplotlib and Seaborn for visualization
    - Scikit-learn for machine learning
  - Excellent for complex statistical analysis
  - Supports big data processing with libraries like Dask
- **Disadvantages**
  - Slower execution compared to compiled languages
  - Steeper learning curve for beginners
  - Memory-intensive for very large datasets
  - Requires more coding compared to specialized tools
- **Performance Metrics**
  - Best for: Complex data analysis, machine learning, scientific computing
  - Dataset Handling: Up to several GB with efficient libraries

- Performance Formula:

 # Performance depends on hardware and library optimization

performance = (computational_efficiency * library_optimization) / dataset_complexity

- **2. R**
- **Capabilities**
  - Specialized statistical programming language
  - Designed specifically for statistical computing and graphics
- **Advantages**
  - Powerful statistical analysis tools
  - Comprehensive statistical libraries
  - Excellent data visualization capabilities
  - Strong in academic and research environments
  - CRAN repository with extensive packages
- **Disadvantages**
  - Steeper learning curve
  - Less versatile compared to Python
  - Slower performance for large datasets
  - Limited general-purpose programming capabilities
- **Performance Metrics**
  - Best for: Statistical analysis, academic research, specialized statistical modeling
  - Dataset Handling: Moderate-sized datasets
  - Performance Formula:
- r
- Copy

- *# Statistical computation efficiency*
- statistical_performance = (package_complexity * analysis_depth) / computational_time
- **3. Power BI**
- **Capabilities**
  - Business intelligence and data visualization tool
  - Microsoft's interactive visualization platform
- **Advantages**
  - User-friendly interface
  - Drag-and-drop visualization
  - Seamless integration with Microsoft ecosystem
  - Real-time data processing
  - Powerful DAX (Data Analysis Expressions) language
- **Disadvantages**
  - Cost (requires licensing)
  - Limited advanced statistical capabilities
  - Less flexible for complex programming
  - Primarily Windows-focused
- **Performance Metrics**
  - Best for: Business reporting, interactive dashboards
  - Dataset Handling: Up to 1GB in desktop version
  - Performance Formula:
- Copy
- visualization_quality = (data_connectivity * rendering_speed) / complexity_level
- **4. Tableau**

- **Capabilities**
  - Advanced data visualization and business intelligence tool
  - Interactive and intuitive visualization platform
- **Advantages**
  - Exceptional data visualization
  - Easy to use, minimal coding required
  - Supports multiple data sources
  - Robust interactive dashboards
  - Good for non-technical users
- **Disadvantages**
  - Expensive licensing
  - Limited advanced statistical analysis
  - Less flexible for custom programming
  - Performance issues with extremely large datasets
- **Performance Metrics**
  - Best for: Business intelligence, data storytelling
  - Dataset Handling: Moderate to large datasets
  - Performance Formula:

  visualization_effectiveness = (data_connection_speed * user_interaction) / data_complexity
- **Comparative Analysis**
- **Dataset Handling Capabilities**
  1. Python: Best for large, complex datasets
  2. R: Good for statistical datasets
  3. Power BI: Limited to moderate-sized datasets
  4. Tableau: Moderate dataset handling

- **Complexity of Analysis**
    1. Python: Highest flexibility and depth
    2. R: Strong statistical analysis
    3. Power BI: Basic to intermediate analysis
    4. Tableau: Visualization-focused analysis
- **Cost Considerations**
    1. Python: Free
    2. R: Free
    3. Power BI: Paid licensing
    4. Tableau: Most expensive
- **Recommended Use Cases**
    - **Python**: Machine learning, scientific computing, complex data analysis
    - **R**: Statistical research, academic analysis
    - **Power BI**: Business reporting, organizational dashboards
    - **Tableau**: Data visualization, business intelligence
- **Learning Curve**
    1. Python: Moderate to steep
    2. R: Steep
    3. Power BI: Easy
    4. Tableau: Easy to moderate
- **Practical Recommendations**
    - For comprehensive data science: Learn Python
    - For statistical research: Focus on R
    - For business reporting: Master Power BI or Tableau
    - For maximum versatility: Learn multiple tools

- **Key Takeaway**
- No single tool is perfect. Choose based on specific project requirements, organizational needs, and personal learning goals.

- **Real-World Scenarios: Tool Selection in Data Analysis and Visualization**
- **Scenario 1: Scientific Research and Complex Data Analysis**
- **Preferred Tool: Python**
- **Scenario**: Climate Change Research at a Major University
  - **Context**: Researchers need to analyze massive datasets involving global temperature, atmospheric CO2 levels, and oceanic current patterns
  - **Why Python?**
    - Handles large, complex scientific datasets
    - Supports advanced numerical computations
    - Integrates machine learning for predictive modeling
    - Libraries like NumPy, SciPy perfect for scientific computing
- **Scenario 2: Pharmaceutical Clinical Trials**
- **Preferred Tool: R**
- **Scenario**: Statistical Analysis of Drug Efficacy Trials
  - **Context**: Analyzing patient response rates, treatment effectiveness, and statistical significance
  - **Why R?**
    - Specialized statistical analysis packages
    - Built-in statistical testing functions
    - Strong visualization for medical research
    - Precise statistical modeling capabilities

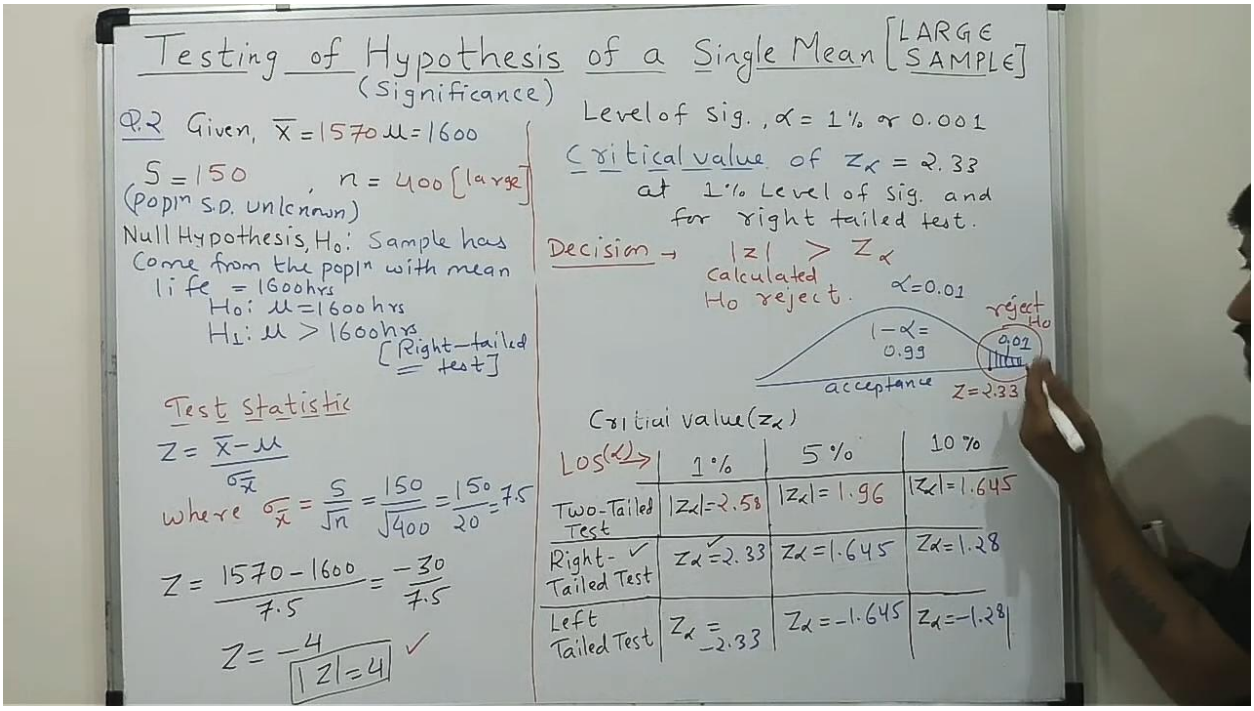
- **Scenario 3: Corporate Financial Reporting**
- **Preferred Tool: Power BI**
- **Scenario**: Quarterly Financial Dashboard for Multinational Corporation
  - **Context**: Creating interactive dashboards showing revenue, expenses, and global performance
  - **Why Power BI?**
    - Seamless Microsoft ecosystem integration
    - Real-time data connection
    - Easy-to-create interactive dashboards
    - Quick financial data visualization
    - DAX language for complex calculations
- **Scenario 4: E-commerce Customer Behavior Analysis**
- **Preferred Tool: Tableau**
- **Scenario**: Understanding Customer Purchase Patterns for Online Retail
  - **Context**: Visualizing customer segmentation, purchase frequencies, and geographical spending trends
  - **Why Tableau?**
    - Intuitive drag-and-drop interface
    - Powerful geographic data visualization
    - Interactive dashboard creation
    - No deep coding knowledge required
    - Quick insights generation
- **Scenario 5: Social Media Sentiment Analysis**
- **Preferred Tool: Python**
- **Scenario**: Analyzing Public Perception of a New Product Launch

- **Context**: Processing large volumes of social media posts, extracting sentiment
- **Why Python?**
    - Natural language processing libraries
    - Web scraping capabilities
    - Machine learning for sentiment classification
    - Scalable data processing

- **Scenario 6: Academic Research in Economics**
- **Preferred Tool: R**
- **Scenario**: Econometric Modeling of Economic Indicators
    - **Context**: Regression analysis, time series forecasting
    - **Why R?**
        - Comprehensive econometric packages
        - Advanced statistical modeling
        - Precise mathematical computations
        - Specialized research-oriented libraries
- **Key Decision Factors**
    1. **Data Complexity**
    2. **Analysis Requirements**
    3. **Visualization Needs**
    4. **Computational Resources**
    5. **Team Expertise**

AIH SOLVED EXAMPLES

The mean life time of a sample of 400 fluorescent light tube produced by a company is found to be 1570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean lifetime of the bulbs produced by the company is 1600 hours against the alternative hypothesis that it is greater than 1600 hours at 1% level of significance.



Example 14·30. *Random samples drawn from two countries gave the following data relating to the heights of adult males :*

	Country A	Country B
Mean height (in inches)	67·42	67·25
Standard deviation (in inches)	2·58	2·50
Number in samples	1,000	1,200

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?