
DECISION TREE – ID3 ALGORITHM NUMERICAL

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Attribute: Outlook

Values (Outlook) = Sunny, Overcast, Rain

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

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971$$

$$S_{Overcast} \leftarrow [4+, 0-]$$

$$\text{Entropy}(S_{Overcast}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$S_{Rain} \leftarrow [3+, 2-]$$

$$\text{Entropy}(S_{Rain}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

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

Gain(S, Outlook)

$$\begin{aligned} &= Entropy(S) - \frac{5}{14} Entropy(S_{Sunny}) - \frac{4}{14} Entropy(S_{Overcast}) \\ &\quad - \frac{5}{14} Entropy(S_{Rain}) \end{aligned}$$

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

Attribute: Temp

Values (Temp) = Hot, Mild, Cool

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

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

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

$$\text{Entropy}(S_{Hot}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1.0$$

$$S_{Mild} \leftarrow [4+, 2-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} \leftarrow [3+, 1-]$$

$$\text{Entropy}(S_{Cool}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

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

$Gain(S, Temp)$

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

$$Gain(S, Temp) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.0289$$

vhubpulse.com

Attribute: Humidity

Values (Humidity) = High, Normal

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

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{High} \leftarrow [3+, 4-]$$

$$\text{Entropy}(S_{High}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.9852$$

$$S_{Normal} \leftarrow [6+, 1-]$$

$$\text{Entropy}(S_{Normal}) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.5916$$

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

Gain(S, Humidity)

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

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

Attribute: Wind

Values (Wind) = Strong, Weak

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

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

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

$$\text{Entropy}(S_{Strong}) = 1.0$$

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

$$\text{Entropy}(S_{Weak}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.8113$$

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

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

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

$Gain(S, Outlook) = 0.2464$

$Gain(S, Temp) = 0.0289$

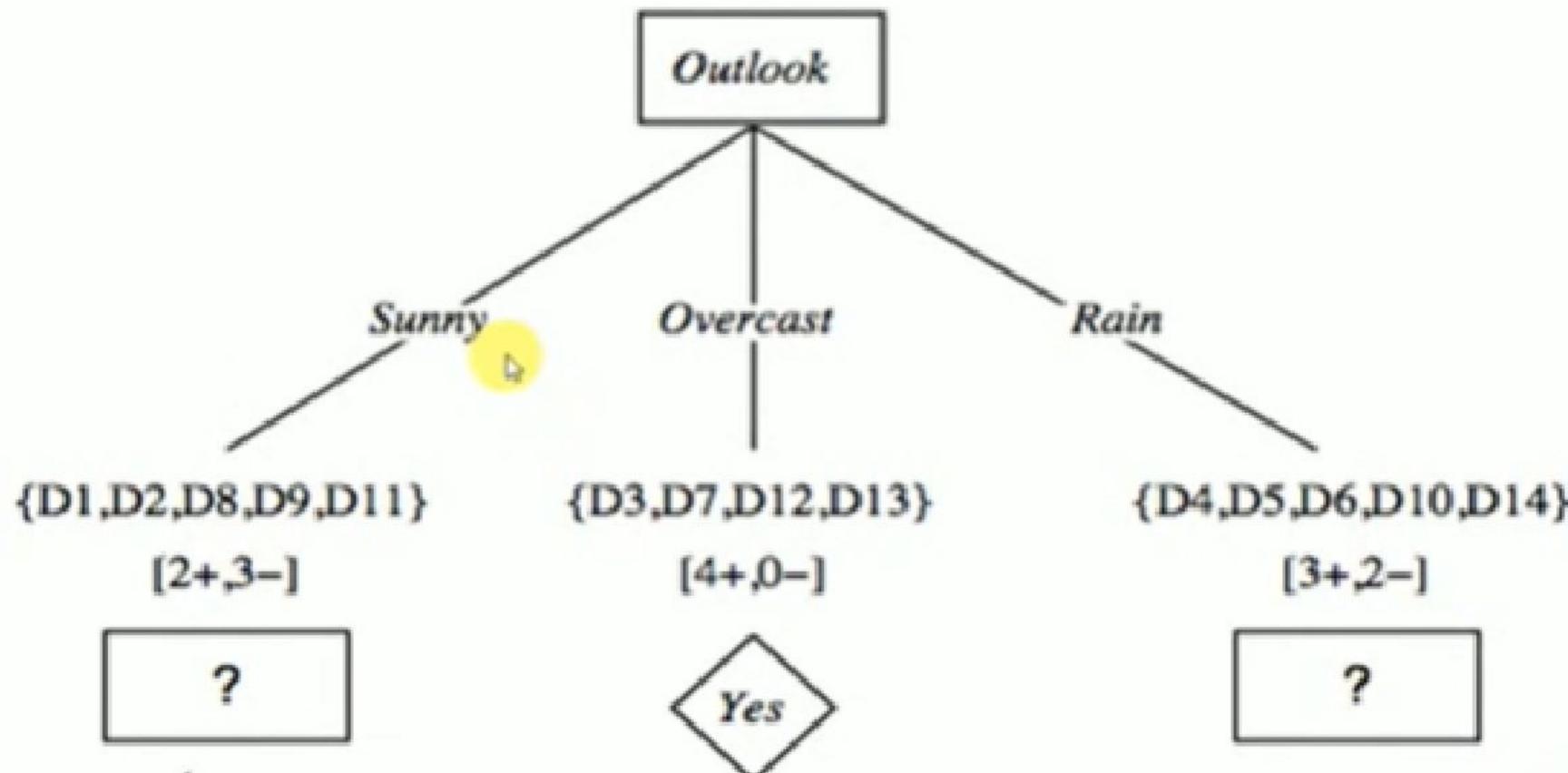


$Gain(S, Humidity) = 0.1516$

$Gain(S, Wind) = 0.0478$

$\{D_1, D_2, \dots, D_{14}\}$

[9+, 5-]



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
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: Humidity

Values (Humidity) = High, Normal

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

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

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

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

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

$$\text{Gain}(S_{\text{Sunny}}, \text{Humidity}) = 0.97 - \frac{3}{5} 0.0 - \frac{2}{5} 0.0 = 0.97$$

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: Wind

Values (Wind) = Strong, Weak

$$S_{Sunny} = [2+, 3-]$$

$$\text{Entropy}(S) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$S_{Strong} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [1+, 2-]$$

$$\text{Entropy}(S_{Weak}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9183$$

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

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

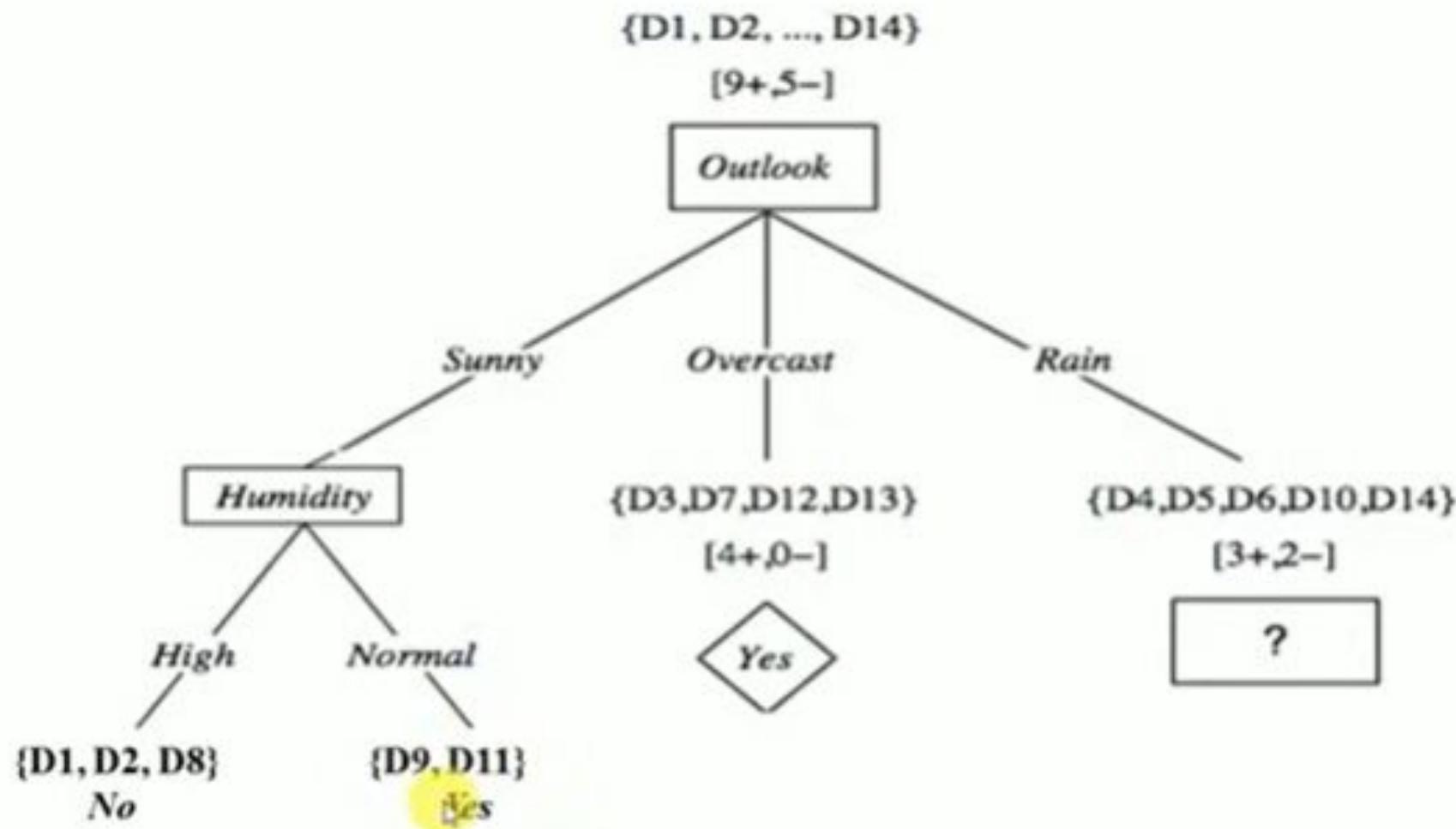
$$\text{Gain}(S_{Sunny}, \text{Wind}) = 0.97 - \frac{2}{5} 1.0 - \frac{3}{5} 0.918 = 0.0192$$

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

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

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

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



Decision Tree IDS Algorithm				
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

Values (Temp) = Hot, Mild, Cool

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 0-]$$

$$\text{Entropy}(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [2+, 1-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$S_{Cool} \leftarrow [1+, 1-]$$

$$\text{Entropy}(S_{Cool}) = 1.0$$

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

$$\text{Gain}(S_{Rain}, Temp)$$

$$= \text{Entropy}(S) - \frac{0}{5} \text{Entropy}(S_{Hot}) - \frac{3}{5} \text{Entropy}(S_{Mild})$$

$$- \frac{2}{5} \text{Entropy}(S_{Cool})$$

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

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

Attribute: Humidity

Values (Humidity) = High, Normal

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

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

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

b)

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

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

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

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

Attribute: Wind

Values (wind) = Strong, Weak

$$S_{Rain} = [3+, 2-]$$

$$\text{Entropy}(S_{Sunny}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$S_{Strong} \leftarrow [0+, 2-]$$

$$\text{Entropy}(S_{Strong}) = 0.0$$

$$S_{Weak} \leftarrow [3+, 0-]$$

$$\text{Entropy}(S_{Weak}) = 0.0$$

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

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

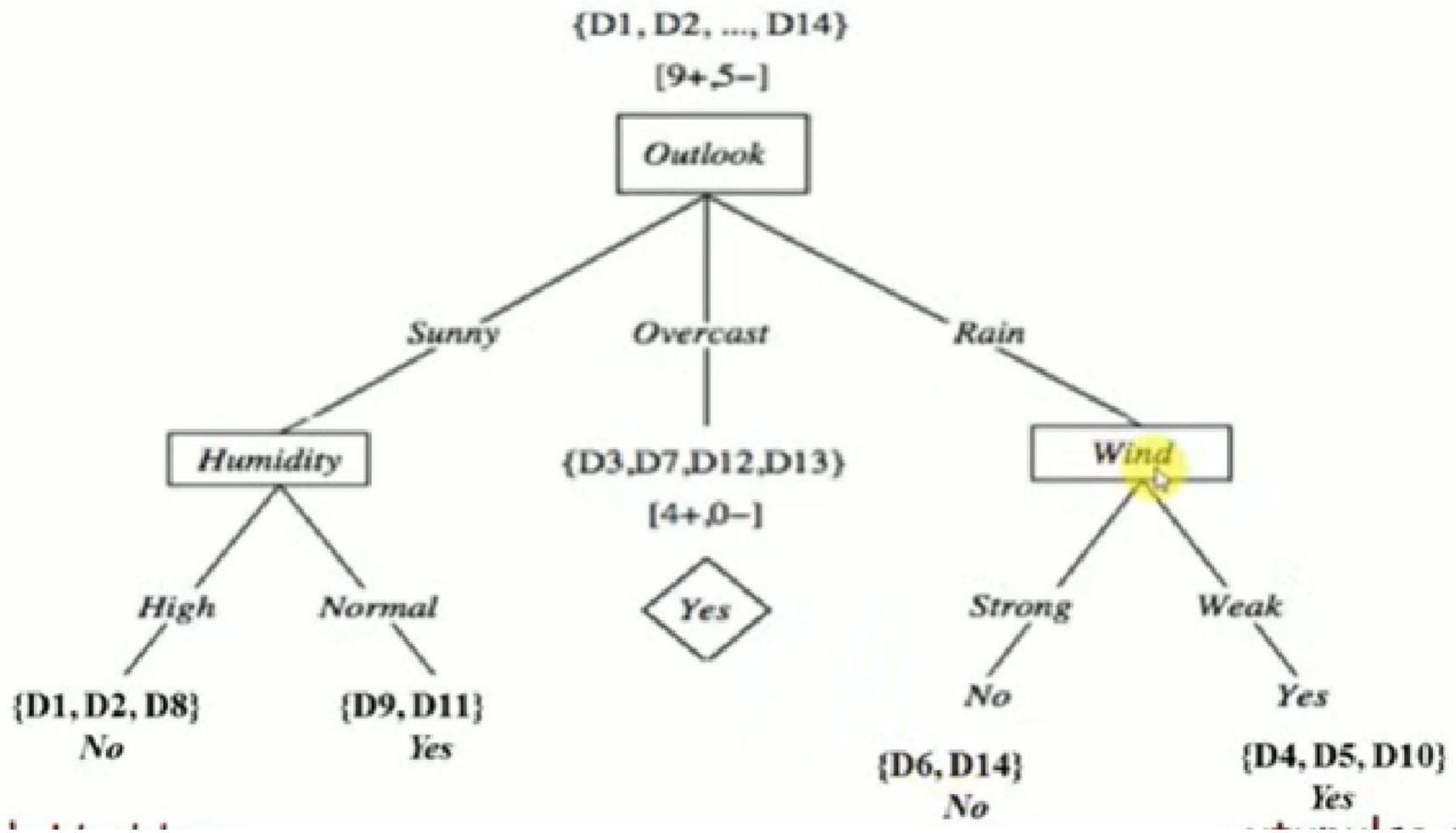
$$\text{Gain}(S_{Rain}, \text{Wind}) = 0.97 - \frac{2}{5} 0.0 - \frac{3}{5} 0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
DI0	Mild	Normal	Weak	Yes
DI4	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$



Naive Bayes Theorem and Conditional Probability

Bayes' Theorem is a mathematical formula that helps determine the conditional probability of an event based on prior knowledge and new evidence.

Bayes theorem (also known as the Bayes Rule or Bayes Law) is used to determine the conditional probability of event A when event B has already occurred.

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

- Where,
- $P(A)$ and $P(B)$ are the probabilities of events A and B also $P(B)$ is never equal to zero,
- $P(A|B)$ is the probability of event A when event B happens,
- $P(B|A)$ is the probability of event B when A happens.

Day	<i>Outlook</i>	<i>Temperature</i>	<i>Humidity</i>	<i>Wind</i>	<i>PlayTennis</i>
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

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

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

$$P(\text{PlayTennis} = \text{yes}) = 9/14 = .64$$

$$P(\text{PlayTennis} = \text{no}) = 5/14 = .36$$

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

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

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

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

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

$$v_{NB} = \operatorname{argmax}_{v_j \in \{yes, no\}} P(v_j) \prod_i P(a_i | v_j)$$

$$\begin{aligned} &= \operatorname{argmax}_{v_j \in \{yes, no\}} P(v_j) \\ &\quad \cdot P(Outlook = sunny | v_j) P(Temperature = cool | v_j) \\ &\quad \cdot P(Humidity = high | v_j) P(Wind = strong | v_j) \end{aligned}$$

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

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

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

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

No.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sport	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

$$X = \{ \text{Red, SUV, Domestic} \}$$

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

$$P(X|\text{Yes}) = ?$$

$$P(X|\text{No}) = ?$$

$$P(\text{Red}|\text{Yes}) = \frac{P(\text{Yes}|\text{Red}) \cdot P(\text{Red})}{P(\text{Yes})} = \frac{\frac{3}{5} \cdot \frac{5}{10}}{\frac{5}{10}} = \frac{3}{5}$$

$$P(\text{SUV}|\text{Yes}) = \frac{P(\text{Yes}|\text{SUV}) \cdot P(\text{SUV})}{P(\text{Yes})} = \frac{\frac{1}{5} \cdot \frac{4}{10}}{\frac{4}{10}} = \frac{1}{5}$$

$$P(\text{Domestic}|\text{Yes}) = \frac{P(\text{Yes}|\text{Domestic}) \cdot P(\text{Domestic})}{P(\text{Yes})} = \frac{\frac{2}{5} \cdot \frac{5}{10}}{\frac{4}{10}} = \frac{2}{5}$$

$$\left(\because P + q = 1 \Rightarrow q = 1 - P \right)$$

$$P(\text{Red}|\text{No}) = 1 - \frac{3}{5} = \frac{2}{5}, \quad P(\text{SUV}|\text{No}) = 1 - \frac{1}{5} = \frac{4}{5}$$

$$P(\text{Domestic}|\text{No}) = 1 - \frac{2}{5} = \frac{3}{5}$$

$$P(X|Yes) = P(Yes) \cdot P(Red|Yes) \cdot P(SUV|Yes) \cdot P(Domestic|Yes)$$

$$\frac{1}{2} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} = \frac{3}{125} = 0.024$$

$$P(X|No) = P(No) \cdot P(Red|No) \cdot P(SUV|No) \cdot P(Domestic|No)$$

$$= \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} = \frac{12}{125} = 4 \times 0.024 = 0.096$$

$$P(X|No) > P(X|Yes)$$

Therefore No ✓

Kmean

- Suppose that the data mining task is to cluster points into three clusters,
- where the points are
- $A_1(2, 10), A_2(2, 5), A_3(8, 4), B_1(5, 8), B_2(7, 5), B_3(6, 4), C_1(1, 2), C_2(4, 9)$.
- The distance function is Euclidean distance.
- Suppose initially we assign A_1, B_1 , and C_1 as the center of each cluster,
respectively.

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

New Centroids:

A1: (2, 10) ✓

B1: (6, 6) —

C1: (1.5, 3.5) ✗

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:
 A1: (2, 10)
 B1: (6, 6)
 C1: (1.5, 3.5)

	Data Points		Distance to					Cluster	New Cluster
			2	10	6	6	1.5		
A1	2	10	0.00		5.66		6.52	1	1
A2	2	5	5.00		4.12		1.58	3	
A3	8	4	8.49		2.83		6.52	2	
B1	5	8	3.61		2.24		5.70	2	
B2	7	5	7.07		1.41		5.70	2	
B3	6	4	7.21		2.00		4.53	2	
C1	1	2	8.06		6.40		1.58	3	
C2	4	9	2.24		3.61		6.04	2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	
A2	2	5	4.61		4.51		1.58		3	
A3	8	4	7.43		1.95		6.52		2	
B1	5	8	2.50		3.13		5.70		2	
B2	7	5	6.02		0.56		5.70		2	
B3	6	4	6.26		1.35		4.53		2	
C1	1	2	7.76		6.39		1.58		3	
C2	4	9	1.12		4.51		6.04		1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	
A2	2	5		4.33		5.04		1.58	3	
A3	8	4		6.62		1.05		6.52	2	
B1	5	8	1.67		4.18		5.70		1	
B2	7	5		5.21		0.67		5.70	2	
B3	6	4		5.52		1.05		4.53	2	
C1	1	2		7.49		6.44		1.58	3	
C2	4	9	0.33		5.55		6.04		1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

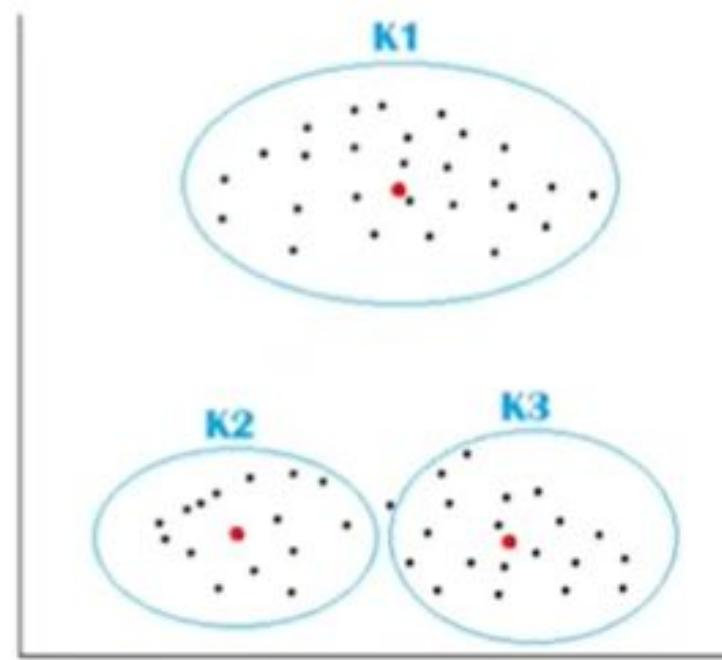
C1: (1.5, 3.5)

Data Points	Distance to						Cluster	New Cluster
	3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94	7.56	6.52	1	1	1
A2	2	5	4.33	5.04	1.58	3	3	3
A3	8	4	6.62	1.05	6.52	2	2	2
B1	5	8	1.67	4.18	5.70	1	1	1
B2	7	5	5.21	0.67	5.70	2	2	2
B3	6	4	5.52	1.05	4.53	2	2	2
C1	1	2	7.49	6.44	1.58	3	3	3
C2	4	9	0.33	5.55	6.04	1	1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

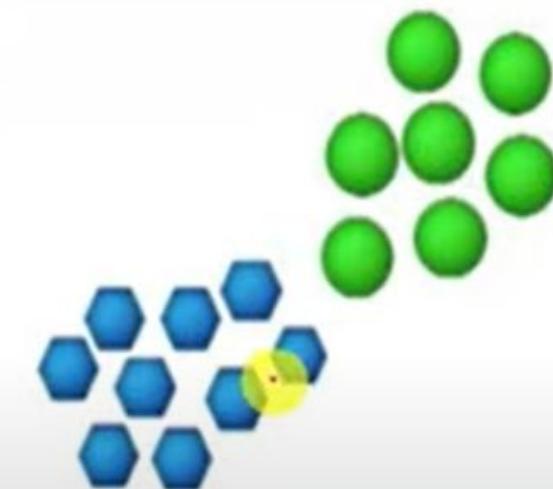
Clustering

- Clustering is a **distance-based unsupervised machine learning algorithm** where data points that are close to each other are grouped in a given number of clusters/groups.



- In hard clustering each datapoint is assigned only a single cluster.
- The K-Means, K-Medoid clustering algorithms are hard clustering algorithms.

Hard Clustering



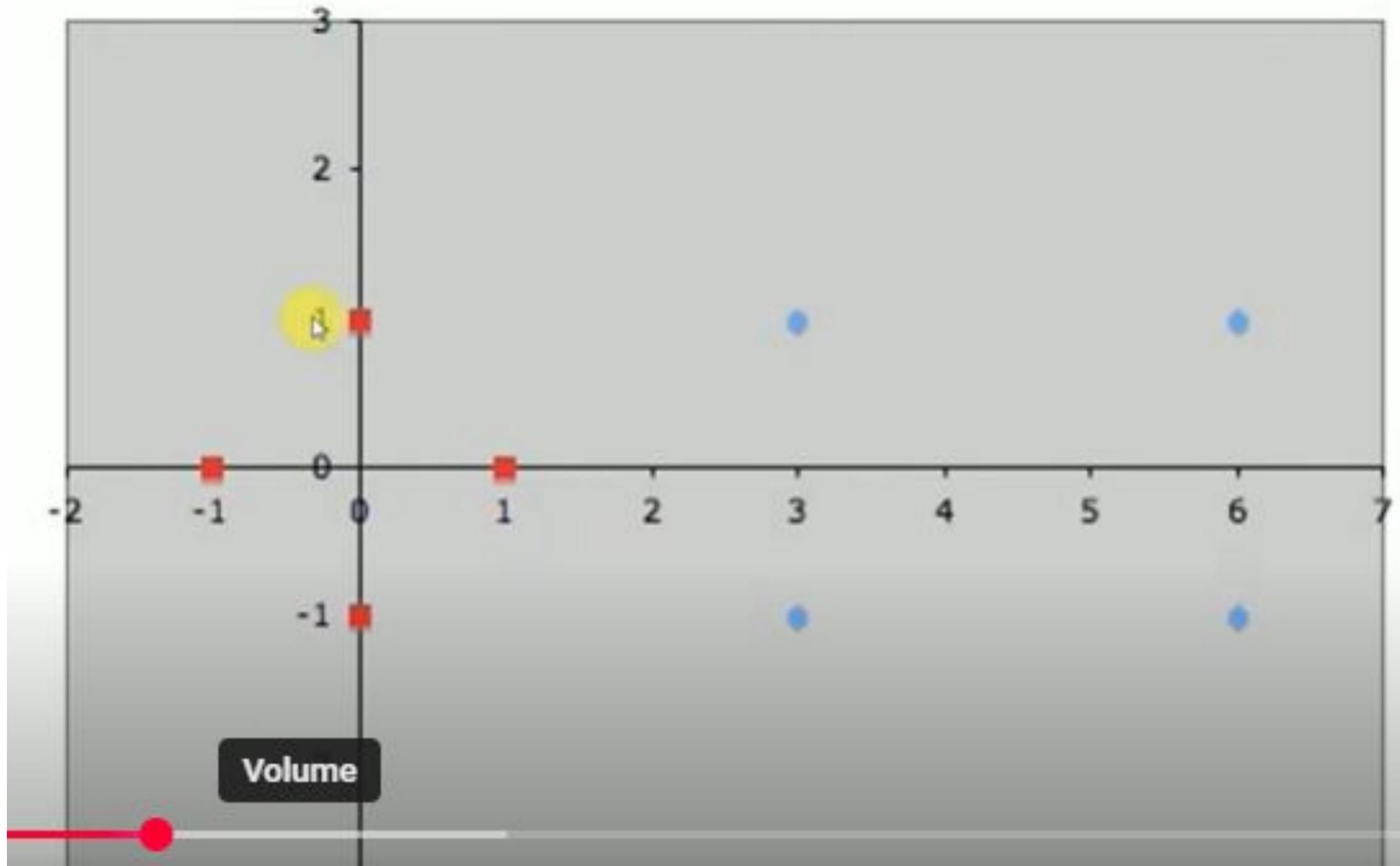
Support Vector Machine - Linear Example Solved

Suppose we are given the following positively labeled data points,

$$\left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} -3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} -6 \\ -1 \end{pmatrix} \right\}$$

and the following negatively labeled data points,

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right\}$$



Support Vector Machine - Linear Example Solved

- Each vector is augmented with a 1 as a bias input
- So, $s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then $\tilde{s}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$
- Similarly,
- $s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, then $\tilde{s}_2 = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$ and $s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$, then $\tilde{s}_3 = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$

Support Vector Machine

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_1 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_1 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_1 = -1$$

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_2 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_2 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_2 = +1$$

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_3 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_3 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_3 = +1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = 1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = 1$$

$$\alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3+0+1) = -1$$

$$\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = 1$$

$$\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = 1$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1$$

$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1$$

$$\alpha_1 = -3.5$$

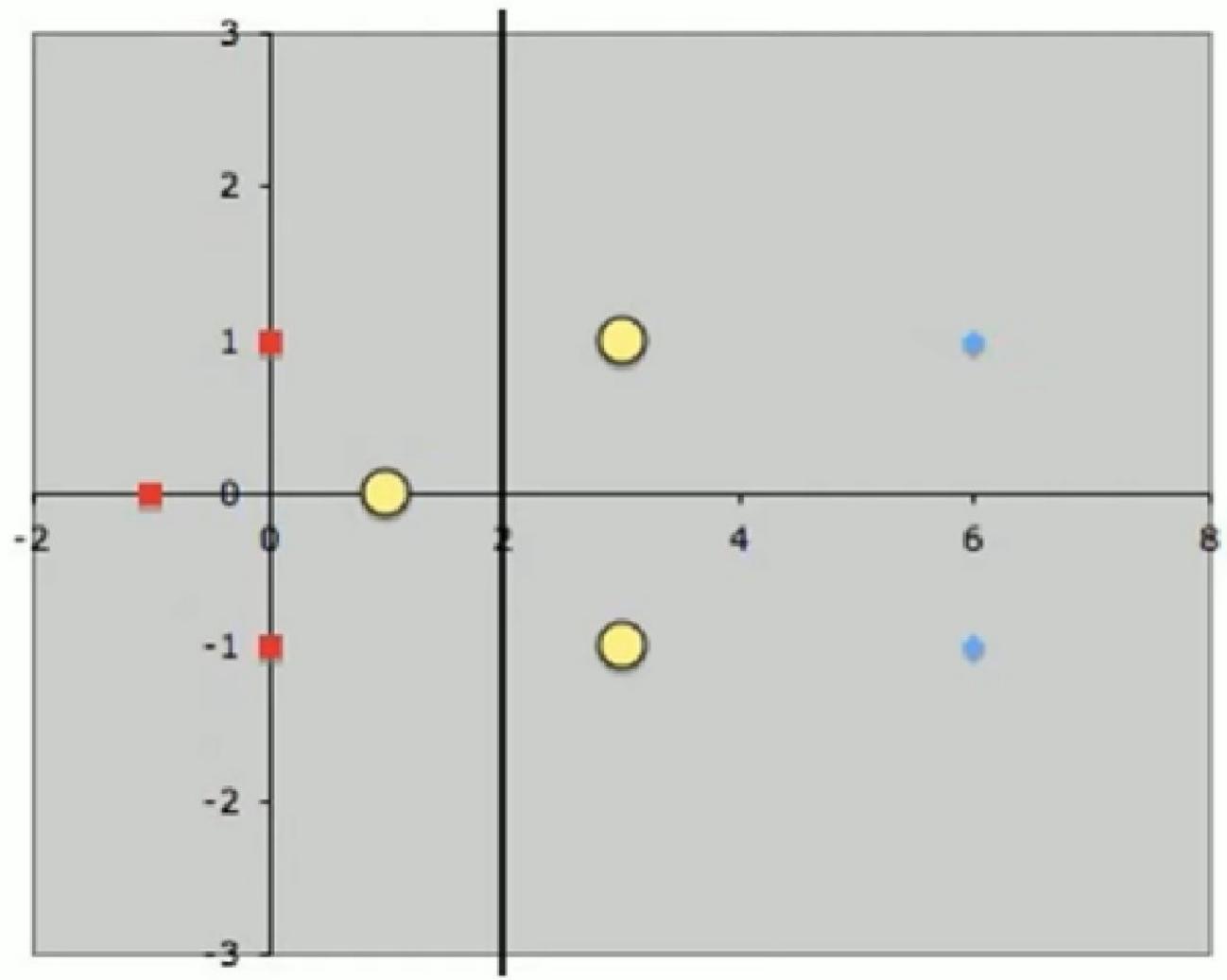
$$\alpha_2 = 0.75$$

$$\alpha_3 = 0.75$$

$$\begin{aligned}
 \tilde{w} &= \sum_i \alpha_i \tilde{s}_i \\
 &= -3.5 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}
 \end{aligned}$$

- Finally, remembering that our vectors are augmented with a bias.
- We can equate the last entry in \tilde{w} as the hyperplane offset b and write the separating
- Hyperplane equation $y = wx + b$
- with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $b = -2$.

 Subscribe



KNN

(**k -Nearest Neighbors**)

KNN Classifier

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	Versicolor
7.3	2.9	Virginica
6.0	2.7	Versicolor
5.8	2.8	Virginica
6.3	2.3	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

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	Versicolor
7.3	2.9	Virginica
6.0	2.7	Versicolor
5.8	2.8	Virginica
6.3	2.3	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

Sepal Length	Sepal Width	Species
5.2	3.1	?

Step 1: Find Distance

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

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

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

Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608

Step 2: Find Rank

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	Versicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Versicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Versicolor	1.36	12
5.1	2.5	Versicolor	0.60	4
6.3	2.5	Versicolor	1.25	11
5.5	2.4	Versicolor	0.75	7
Play (k)				

Test point is belonging to class which has minimum distance if k is not defined.

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	Versicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Versicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Versicolor	1.36	12
5.1	2.5	Versicolor	0.60	4
6.3	2.5	Versicolor	1.25	11
5.5	2.4	Versicolor	0.75	7

Step 3: Find the Nearest Neighbor

If k = 1 – Setosa

If k = 2 – Setosa

If k = 5 – Setosa