Regression Trees

Zahoor Tanoli (PhD)
CUI COMSATS Attock
xahoor@ciit-attock.edu.pk

Definition

- Decision trees are power tools to classify the problem
- Can be adopted to solve regression problems
- Decision trees which built for a data set where the the target column could be real number are called regression trees
- For regression trees, information gain/gain ratios and gini index wouldn't work
- What to do?

Let Look into Data Set

- We will be using golf or tennis data set
- Golf playing decision was nominal (yes/No)
- We can count number of instances of class
- When the target column is Number of Golf Players and stores real number so can't count
- Instead of count, we can handle problem by **standard deviation**

Changed Data Set

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

What to Do?

- **Standard deviation**
- Golf players = {25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30}
- Average of golf players = (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)/14 = **39.78**
- Standard deviation of golf players = $\sqrt{}$[((25 – 39.78)2 + (30 – 39.78)2 + (46 – 39.78)2 + … + (30 – 39.78)2)/14] = **9.32**

Finding Decision Feature

- Outlook
- Temperature
- Humidity
- Wind

Outlook

- Outlook can be sunny, overcast and rain

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Golf players for sunny outlook = {25, 30, 35, 38, 48}
- Average of golf players for sunny outlook = (25+30+35+38+48)/5 = **35.2**
- Standard deviation of golf players for sunny outlook
- = √(((25 – 35.2)2 + (30 – 35.2)2 + …)/5) = **7.78**

Outlook Continued...

- Overcast

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Golf players for overcast outlook = {46, 43, 52, 44}

Average of golf players for overcast outlook = (46 + 43 + 52 + 44)/4 = **46.25**

Standard deviation of golf players for overcast outlook

= √(((46-46.25)2+(43-46.25)2+…)= **3.49**

Outlook Rainy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Golf players for overcast outlook = {45, 52, 23, 46, 30}

Average of golf players for overcast outlook

= (45+52+23+46+30)/5 = **39.2**

Standard deviation of golf players for rainy outlook

= √(((45 – 39.2)2+(52 – 39.2)2+…)/5) = **10.87**

Summarized Standard Deviation

Outlook	Stdev of Golf Players	Instances
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

Weighted standard deviation for outlook

= (4/14)x3.49 + (5/14)x10.87 + (5/14)x7.78 = **7.66**

Standard deviation reduction for outlook

= 9.32 – 7.66 = 1.66

Temperature Feature

- Temperature can be hot, cool or mild

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
13	Overcast	Hot	Normal	Weak	44

- Golf players for hot temperature = {25, 30, 46, 44}
- Standard deviation of golf players for hot temperature = **8.95**

Temperature Continued...

- For Cool Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38

- Golf players for cool temperature = {52, 23, 43, 38}
- Standard deviation of golf players for cool = **10.51**

Temperature Continued...

- For Mild Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

- Golf players for mild = {45, 35, 46, 48, 52, 30}
- Standard deviation of golf players = **7.65**

Summarized Standard Deviation

- Temperature Feature

Temperature	Stdev of Golf Players	Instances
Hot	8.95	4
Cool	10.51	4
Mild	7.65	6

- Weighted standard deviation = (4/14)x8.95 + (4/14)x10.51 + (6/14)x7.65 = **8.84**

- Standard deviation reduction = 9.32 – 8.84 = **0.47**

Humidity Feature

Humidity high

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

- Golf players for high = {25, 30, 46, 45, 35, 52, 30}
- Standard deviation for high humidity = **9.36**

Humidity Continued...

Normal Humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
13	Overcast	Hot	Normal	Weak	44

- Golf players for normal humidity = {52, 23, 43, 38, 46, 48, 44}
- Standard deviation for golf players for normal humidity = **8.73**

Summarized Standard Deviation

Humidity Feature

Humidity	Stdev of Golf Player	Instances
High	9.36	7
Normal	8.73	7

- Weighted standard deviation = (7/14)x9.36 + (7/14)x8.73 = **9.04**
- Standard deviation reduction = 9.32 – 9.04 = **0.27**

Wind Feature

Strong Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

- Golf players for strong = {30, 23, 43, 48, 52, 30}
- Standard deviation for strong wind = **10.59**

Wind Continued...

Weak Wind

1	Sunny	Hot	High	Weak	25
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
13	Overcast	Hot	Normal	Weak	44

- Golf players for weakk wind= {25, 46, 45, 52, 35, 38, 46, 44}
- Standard deviation for golf players for weak wind = **7.87**

Summarized Standard Deviation

Wind Feature

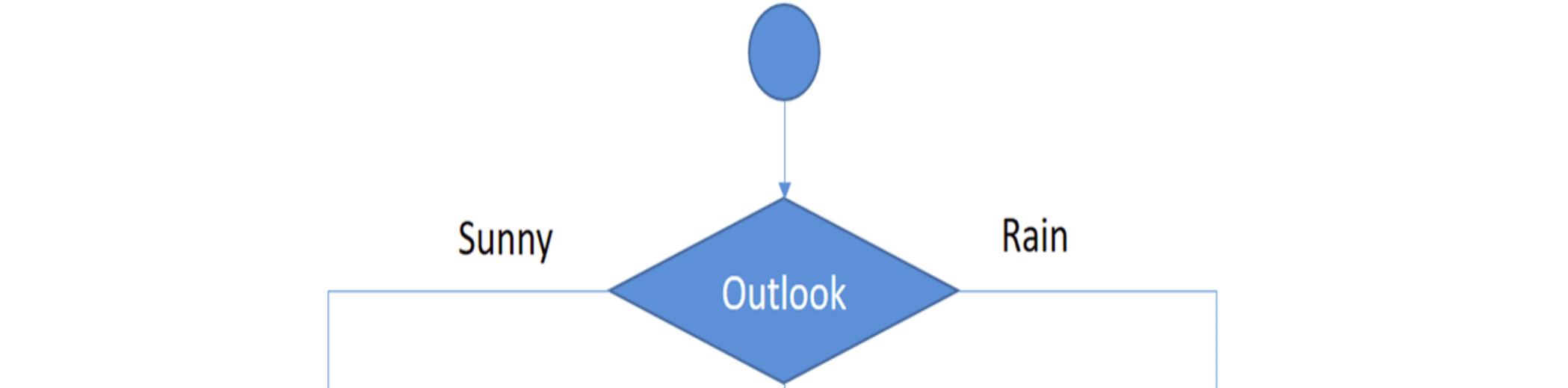
Wind	Stdev of Golf Player	Instances
Strong	10.59	6
Weak	7.87	8

- Weighted standard deviation for wind = (6/14)x10.59 + (8/14)x7.87 = 9.03
- Standard deviation reduction wind = 9.32 – 9.03 = **0.29**

Comparison of Standard Deviation

Feature	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29

- The winner is **outlook** because it has the highest reduction value
- Put outlook decision at the top of decision tree
- Let's look at the decision tree



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Each Branch as Separate Dataset

- Outlook (Sunny)

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Golf players for sunny outlook = {25, 30, 35, 38, 48}
- Standard deviation for sunny outlook = 7.78
- This standard deviation value is global for this sub data set.

Sunny outlook and Hot Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

- Standard deviation for sunny outlook and hot temperature = 2.5

Sunny outlook and Cool Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38

- Standard deviation for sunny outlook and cool temperature = 0

Sunny outlook and Mild Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

- Standard deviation for sunny outlook and mild temperature = 6.5

Summary of standard deviations

- Temperature feature when outlook is sunny

Temperature	Stdev for Golf Players	Instances
Hot	2.5	2
Cool	0	1
Mild	6.5	2

- Weighted standard deviation = (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = 3.6
- Standard deviation reduction = 7.78 – 3.6 = **4.18**

Sunny outlook and high humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35

- Standard deviation for sunny outlook and high humidity = 4.08

Sunny outlook and normal humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

- Standard deviation for sunny outlook and normal humidity = 5

Summarizing standard deviations

Humidity	Stdev for Golf Players	Instances
High	4.08	3
Normal	5.00	2

- Weighted standard deviations for sunny outlook and humidity = (3/5)x4.08 + (2/5)x5 = **4.45**
- Standard deviation reduction for sunny outlook and humidity = 7.78 – 4.45 = **3.33**

Sunny outlook and Strong Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
11	Sunny	Mild	Normal	Strong	48

- Standard deviation for sunny and strong wind = 9

Sunny outlook and Weak Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38

- Standard deviation for sunny and weak wind = 5.56

Use of Decision Trees