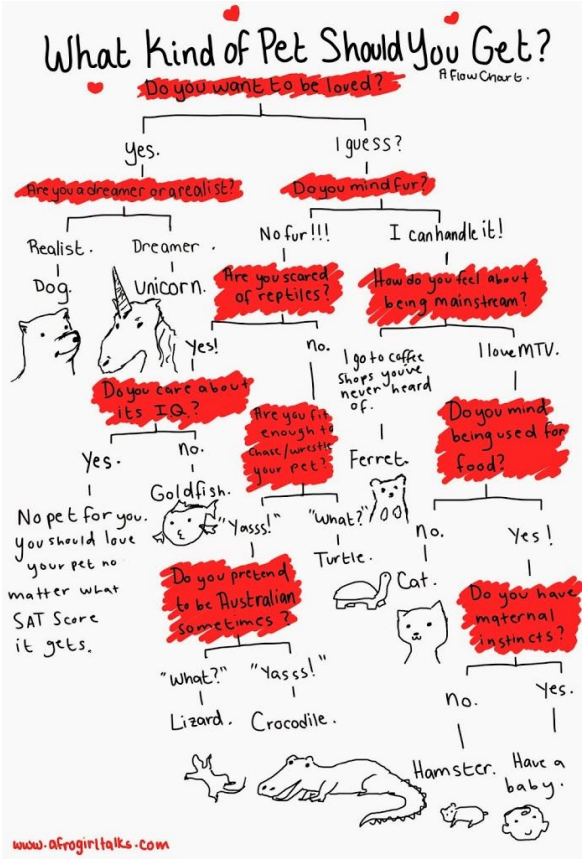
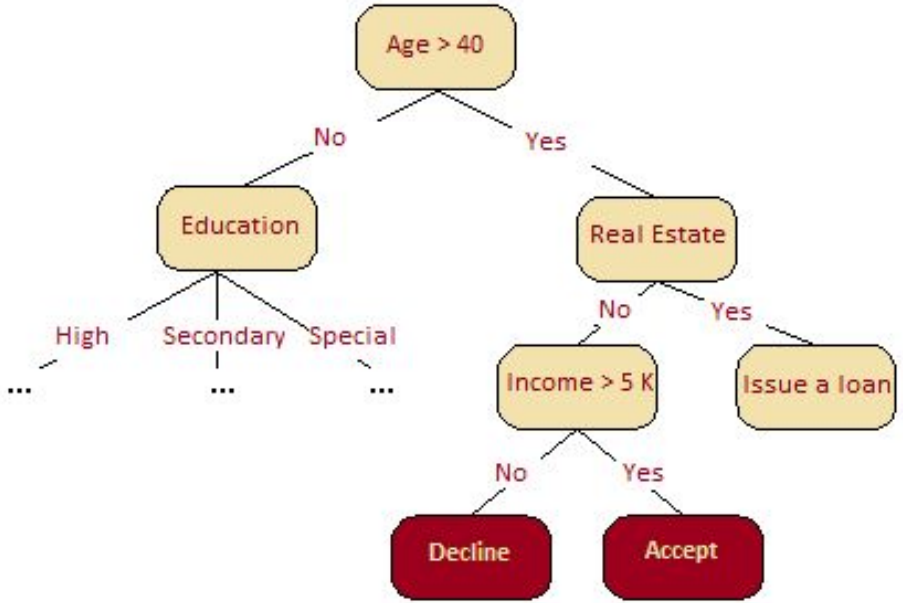
Decision Trees

Presentation by Berk Sudan

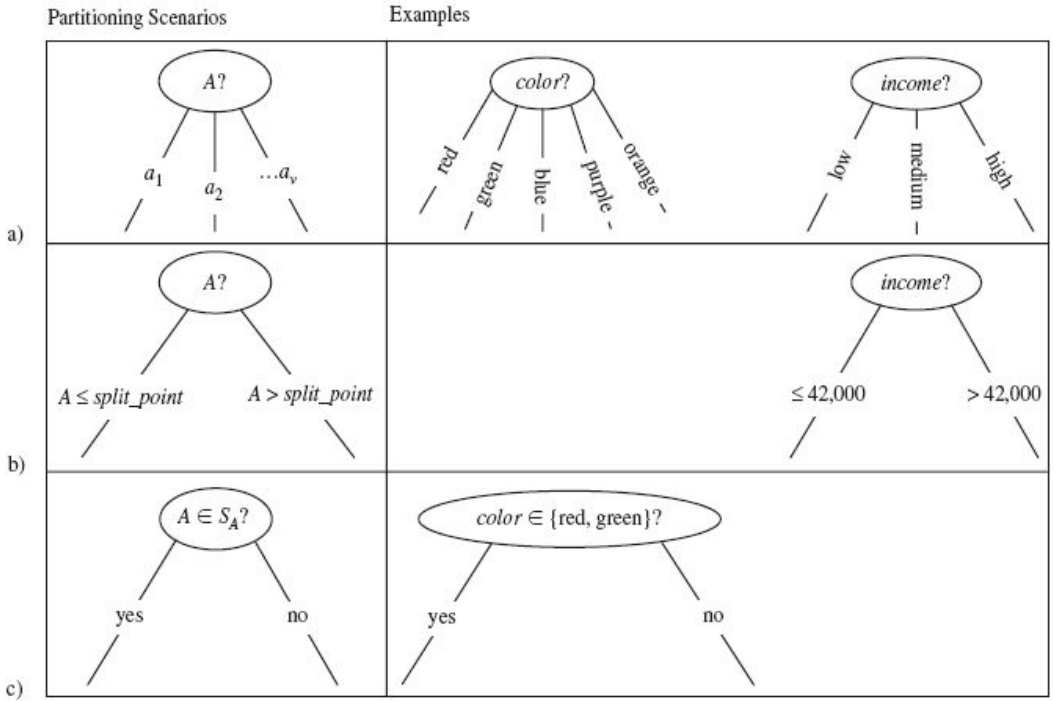

What Kind of Pet Should You Get?
A Flow Chart.
Do you want to be loved?
Yes.
I guess?
Are you a dreamer or a realist?
Do you mind Fur?
Realist.
Dreamer.
No fur!!!
I can handle it!
Dog.
Unicorn.
Are you scared of reptiles?
How do you feel about being mainstream?
Yes!
no.
I go to coffee shops you've never heard of.
I love MTV.
Do you care about its IQ?
Are you fit enough to chase/wrestle your pet?
Do you mind being used for food?
Yes.
no.
Ferret.
Goldfish.
No pet for you. You should love your pet no matter what SAT Score it gets.
"Yasss!"
"What?"
no.
Yes!
Turtle.
Cat.
Do you pretend to be Australian sometimes?
Do you have maternal instincts?
"What?"
"Yasss!"
no.
Yes.
Lizard.
Crocodile.
Hamster.
Have a baby.
www.afrogirltalks.com

Decision to Grant a Loan



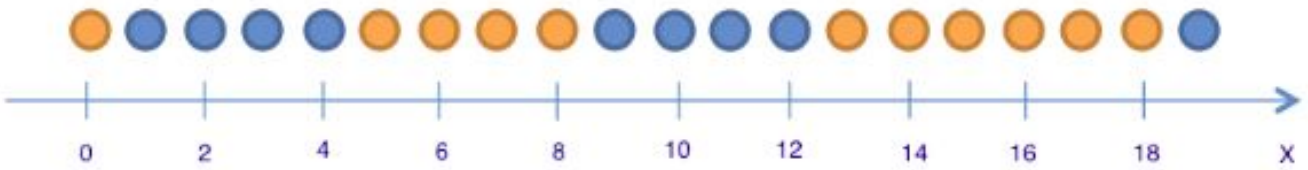
Ref: *https://mlcourse.ai/articles/topic3-dt-knn/*

Partitioning Scenarios



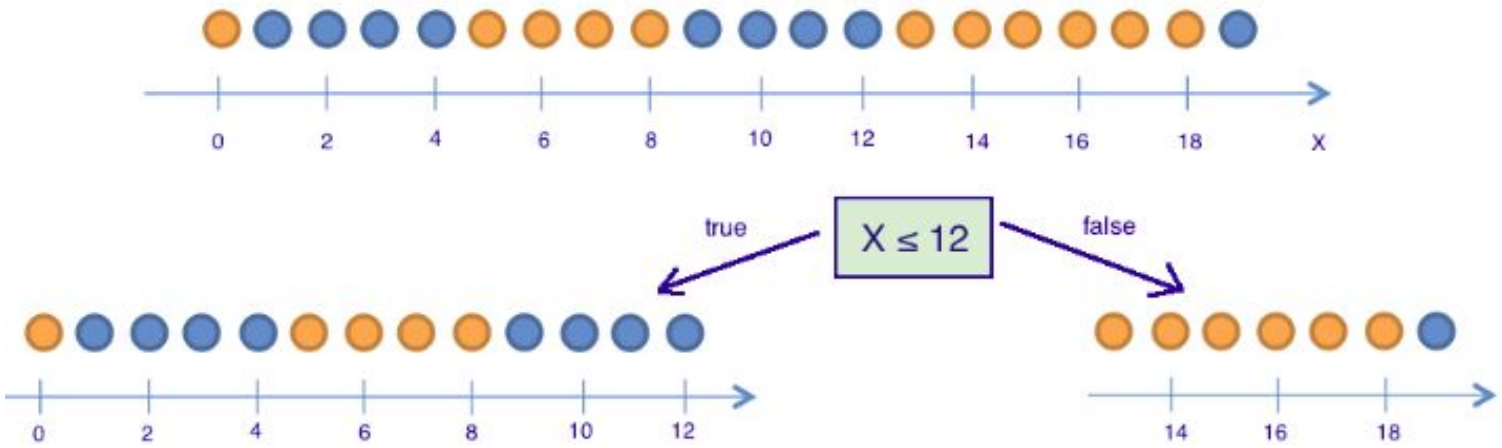
Ref: *Doç. Dr. Songül Varlı, Introduction to Data Mining Lecture Slides, 2018*

Partitioning Example

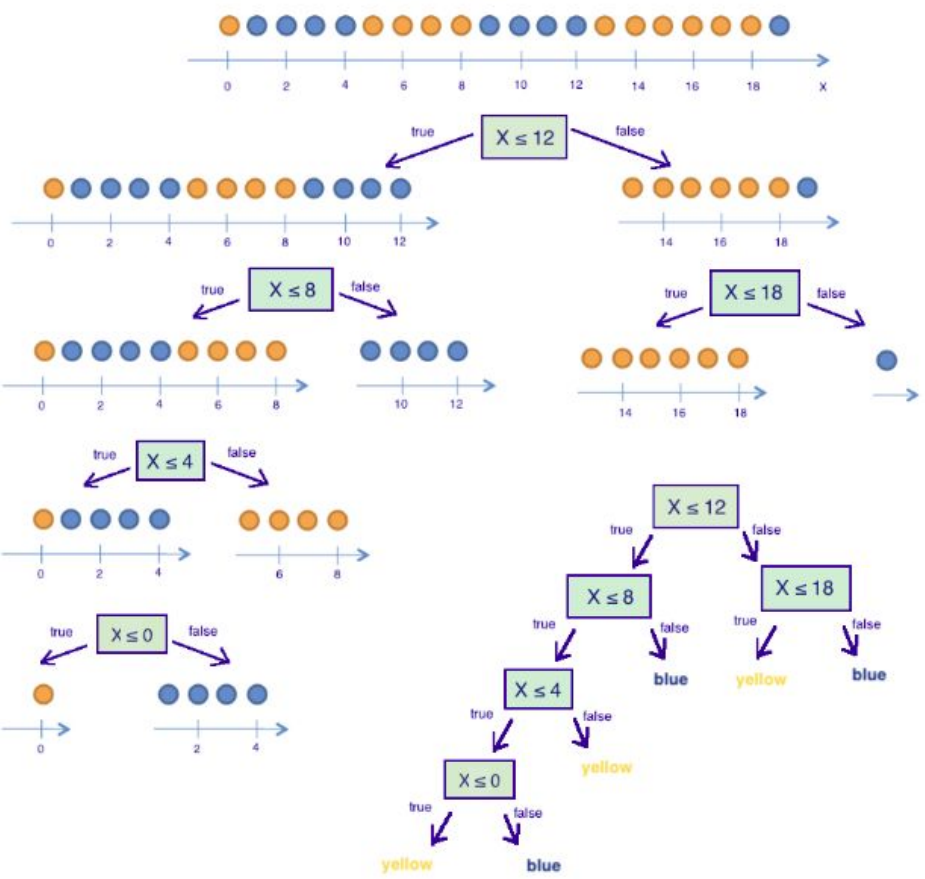


Ref: *https://mlcourse.ai/articles/topic3-dt-knn/*

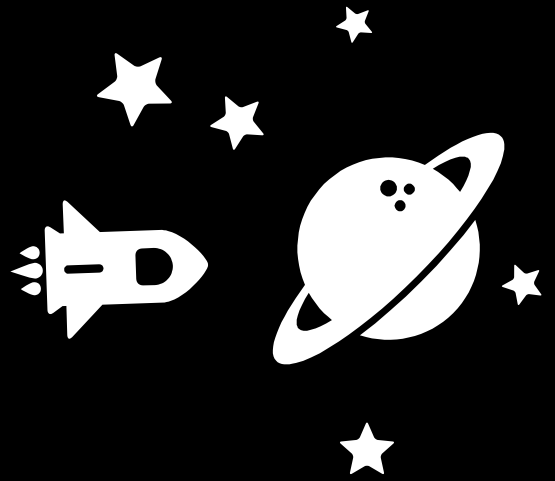
Partitioning Example



Ref: *https://mlcourse.ai/articles/topic3-dt-knn/*


0 2 4 6 8 10 12 14 16 18 X
X ≤ 12
true
false
0 2 4 6 8 10 12
14 16 18
X ≤ 8
X ≤ 18
0 2 4 6 8
10 12
14 16 18
X ≤ 4
0 2 4
6 8
X ≤ 0
0
2 4
X ≤ 12
X ≤ 8
X ≤ 18
X ≤ 4
X ≤ 0
blue
yellow
blue
yellow
yellow
blue

Information Gain with Example



Entropy (Expected Information)

Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Ref: *Doç. Dr. Songül Varlı, Introduction to Data Mining Lecture Slides, 2018*

Information Gain

$$\text{Information gain} = \text{entropy (parent)} - [\text{weightes average}] * \text{entropy (children)}$$

Ref: *https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01*

Self Driving Car Example

	Grade	Bumpiness	Speed Limit	Speed
2	steep	bumpiness	yes	slow
3	steep	smooth	yes	slow
4	flat	bumpiness	no	fast
5	steep	smooth	no	fast

- Grade, Bumpiness and Speed Limit are the features and **Speed** is label.
- Total **4** observation.

Ref: *https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01*

Self Driving Car Example

	Grade	Bumpiness	Speed Limit	Speed
2	steep	bumpiness	yes	slow
3	steep	smooth	yes	slow
4	flat	bumpiness	no	fast
5	steep	smooth	no	fast

$$Entropy_{parent} = -\sum P_{slow} log_2(P_{slow}) + P_{fast} log_2(P_{fast})$$

Ref: *https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01*

Self Driving Car Example

	Grade	Bumpiness	Speed Limit	Speed
2	steep	bumpiness	yes	slow
3	steep	smooth	yes	slow
4	flat	bumpiness	no	fast
5	steep	smooth	no	fast

QUESTION: Gain(T, Grade)= ?

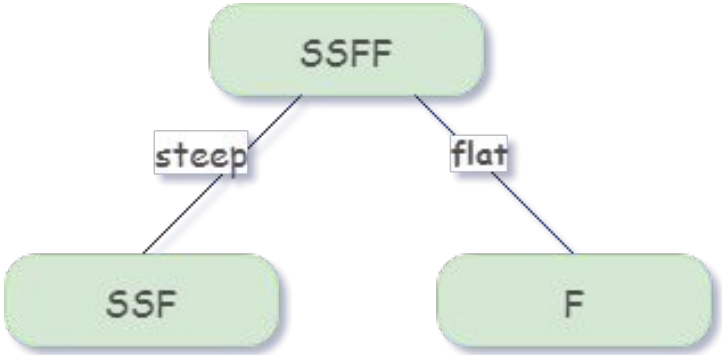
Note: T = set of training instances.

log_2{1/3} = -1.6

log_2{2/3} = -0.6

Ref: *https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01*

Self Driving Car Example

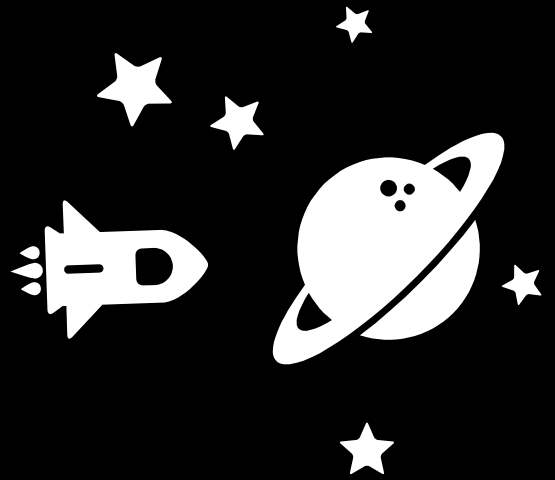


QUESTION: Gain(T, Grade)= ?

Note: T = set of training instances.

Ref: *https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01*

Gini Impurity

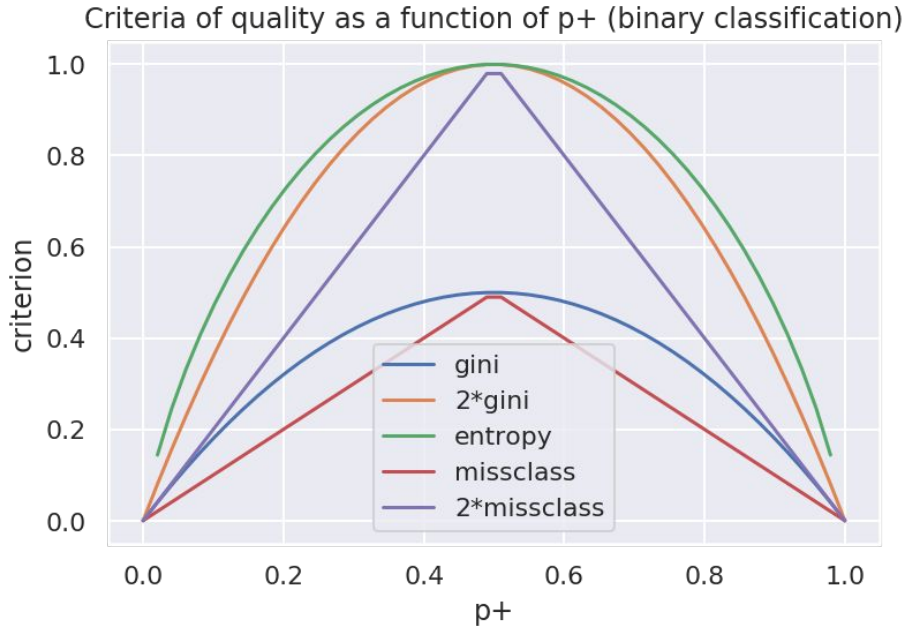


Gini Impurity - Formula

- Gini uncertainty (Gini impurity): $G = 1 - \sum_k (p_k)^2$. Maximizing this criterion can be interpreted as the maximization of the number of pairs of objects of the same class that are in the same subtree (not to be confused with the Gini index).
- Misclassification error: $E = 1 - \max_k p_k$

Ref: *https://mlcourse.ai/articles/topic3-dt-knn/*

Criteria for Split



Ref: *https://mlcourse.ai/articles/topic3-dt-knn/*

End of Presentation

Presented by Berk Sudan