# Leertaak 2: Chris de Windt & Serkan Orhan

## Opdracht 1

Zie leertaak2Recurise.zip

## Opdracht 2

Zie leertaak2Bomen.zip

## Opdracht 7

Binair zoeken gaat veel sneller dan lineair zoeken. Soms 10 keer zo snel. Code staat in leertaak2LineairZoeken.zip

## Opdracht 8

De orde van het linieaire zoekalgoritme is O(n).

## Opdracht 9

De orde van het binaire zoekalgoritme is O(log n). Dit komt doordaat je elke keer halveert.

## Opdracht 10

Om de best split uit te rekenen moet je de information gain voor de Airco en de ABS berekenen.  
eerst moet de inforation content van de hele trainingset worden berekent en daarna features.

T abs,ja T abs,nee T airco,ja T airco,nee moeten worden berekent.

I = ∑f\*log(f) A = (hoog, midden, lag) dit zijn de categorieën

De sequence voor de hele trainingset is (hoog, hoog, midden, laag)

F1 = 0.5, F2 = 0.25, F3 = 0.25 F is de frequentie

F = hoe vaak komt een categorie voor in een sequence/ aantal categorieën in sequence

(0,5\*log(0,5)) + (0,25\*log(0,25)) + (0,25\*log(0,25)) = -1,5 => It = -1,5

Tairco,ja sequence is (hoog,hoog)

F1 = 1

I airco,ja = 1\*log(1) => 0

Tairco,nee sequence is (midden, laag)

F1 = 0.5 F2 = 0.5

I airco,nee = (0.5\*log(0.5)) + (0.5\*log(0.5)) => -1

Tabs,ja sequence is (hoog, midden)

F1 = 0.5, F2 = 0.5

I abs,ja = (0.5\*log(0.5)) + (0.5\*log(0.5)) => -1

Tabs,nee sequence is (hoog, laag)

F1 = 0.5, F2 = 0.5

I abs,nee = (0.5\*log(0.5)) + (0.5\*log(0.5)) => -1

Nu kan het verschil worden berekent van de It en Ifeature. Om het verschil te bereken moet voor beide features het gewogen gemiddelde worden berekent

M = aantal waardes voor een feature T = trainingset

Information gain is dan de – de gewogen gemiddelde van een feature

Gewogen gemiddelde voor airco

(2/4 \* 0) + (2/4 \* -1) => -0,5

Gewogen gemiddelde voor ABS2

(2/4\* -1) + (2/4 \* -1) => -2

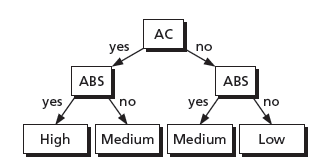
Nu kunnen we de information gain van allebei berekenen

I Airco = -1,5 - -0,5 = -1,5 + 0.5 => -1

I ABS = -1,5 - -2 = -1,5 + 2 => 0.5

Nu kan de best split worden gekozen door te kijken naar de infomation gain I. hoe lager hoe meer informatie het bevat. De gene met de meeste informatie is de best split. We zien dat airco -1 is dus de meeste informatie bevat. De best split is Airco.

Van het algoritme komt hetzelfde beslisboom uit als dat in het boek staat.

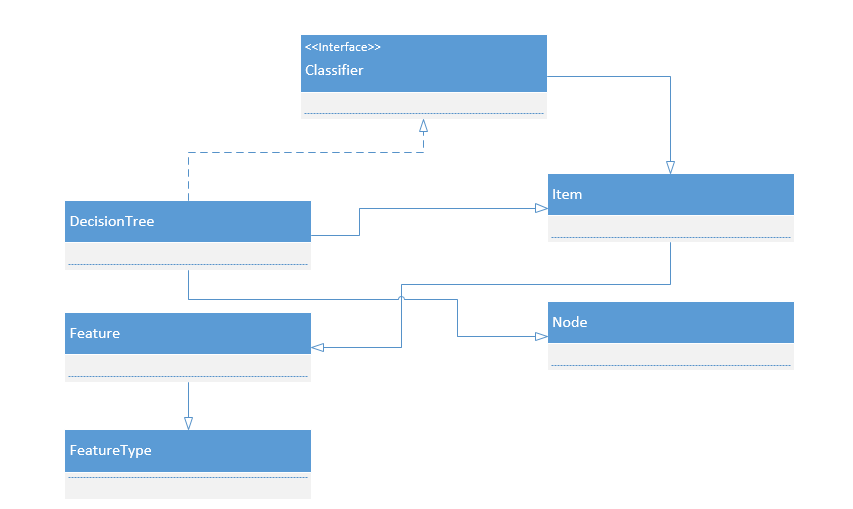


## Opdracht 11

De eerste stap van de algoritme wordt altijd 1x uitgevoerd. De tweede stap wordt n keer uitgevoerd afhankelijk van de boom. De derde stap wordt ook 1 keer uitgevoerd. Dus:

1\* (c\*n) \* 1 = 2 \* cn => cn de constante kunnen we weghalen. De O notatie is O(n)

## Opdracht 13



## Opdracht 15 & 16

Zie leertaak2Classifier.zip en TrainingSets.zip