UNDERSTANDING MACHINE LEARNING APPROACHES

DIFFERENCE BETWEEN ML APPROACHES

Supervised Learning Models

- Use already existing example inputs and desired outputs to learn the overall system behavior
- Types:
  - Classification
  - Regression

Unsupervised Learning Models

- No information on desired output is provided. Algorithm is supposed to find the underlying structures automatically.
- Types:
  - Clustering

DIFFERENCE BETWEEN ML APPROACHES

Supervised Learning Models

- Classification

Classification models use the values of one or more **input** fields to predict the value of one or more output, or **target**, fields, typically discrete values.

Unsupervised Learning Models

- Clustering

These models divide the data into segments, or clusters, of records that have similar patterns of input fields. As they are only interested in the input fields, segmentation models have no concept of output or target fields.

DIFFERENCE BETWEEN ML & STATISTICS

Machine Learning Models

- A bottom-up approach
- Involves interrogating the data for information
- Determined by method and goals rather than by the users
- Typically require large datasets to work upon

Statistical Learning Models

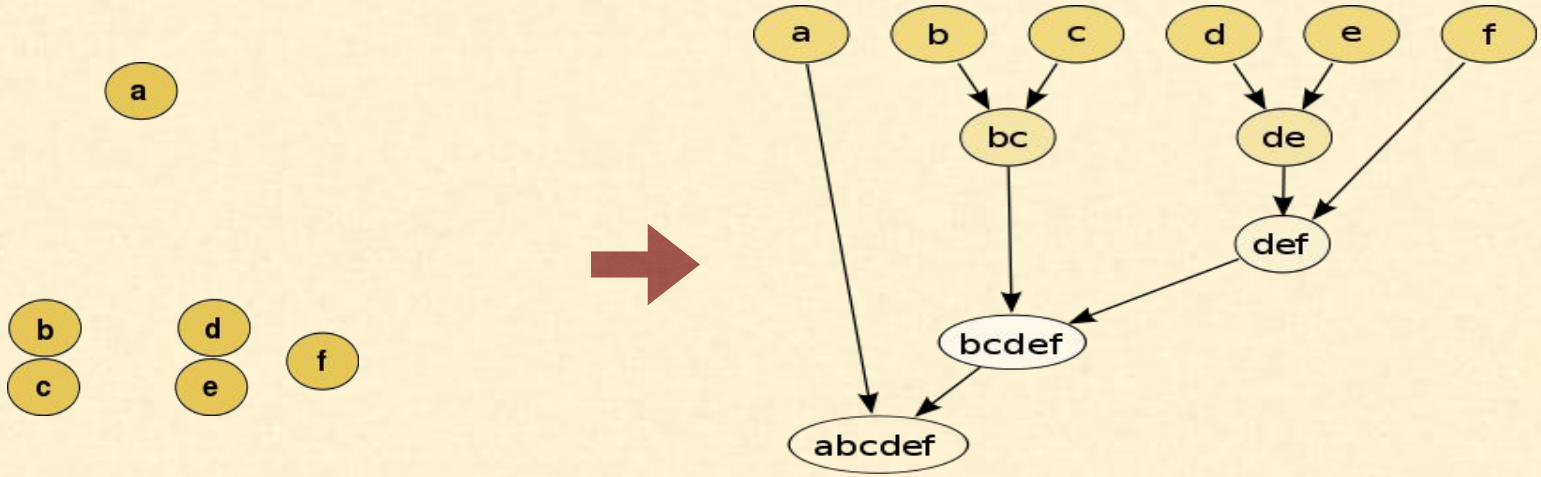
- A top-down approach
- Forming a theory about possible relationships between variables within data
- Formulating hypotheses about the same followed by testing them for affirmation
- Works with smaller datasets as well

UNSUPERVISED MACHINE LEARNING APPROACHES

CLUSTER ANALYSIS

Task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

CLUSTER ANALYSIS : HIERARCHICAL



dendrogram

CLUSTER ANALYSIS : HIERARCHICAL

Understanding Data - mtcars

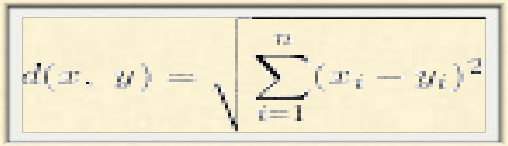
32 CARS	MPG	CYL	DISP	HP	DRAT	WT	QSEC	VS	AM	GEAR	CARB
MAZDA RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
MAZDA RX4 WAG	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
DATSUN 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
HORNET 4 DRIVE	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
HORNET SPORTABOUT	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
VALIANT	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
DUSTER 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4

CLUSTER ANALYSIS : HIERARCHICAL

Measure of Distance

It quantifies dissimilarity between sample data for numerical computation.

A popular choice of distance metric is the Euclidean distance,

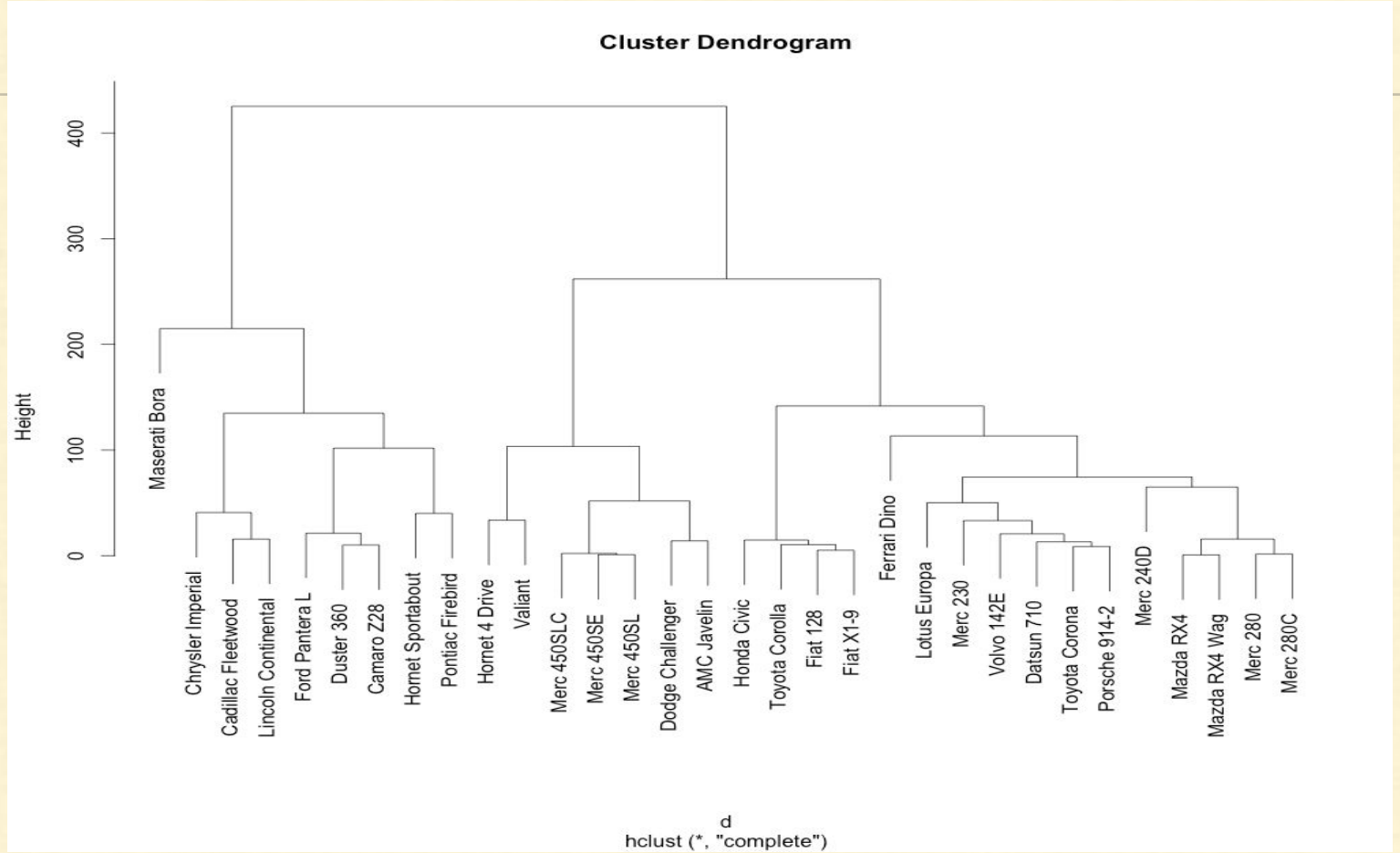



```
x <- mtcars["Honda Civic",]
y <- mtcars["Camaro Z28",]
dist(rbind(x, y))
            Honda Civic
Camaro Z28      335.89
```

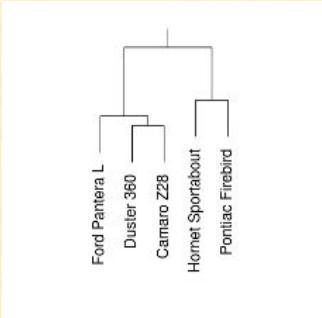
```
> r = x - y
> rq  =r*r
> sqrt(sum(rq))
```

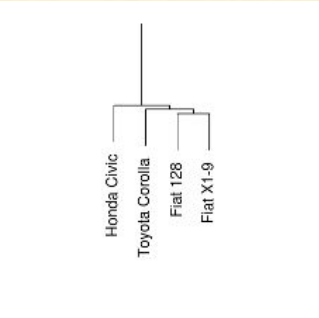
CLUSTER ANALYSIS : Hands-on

- We will explore mtcars dataset
- Go to Equiskill lab http://lab1.equiskill.com
- Check mtcars dataset
- ?mtcars
- Open file cluster_analysis.R


Cluster Dendrogram
Height
0
100
200
300
400
Maserati Bora
Chrysler Imperial
Cadillac Fleetwood
Lincoln Continental
Ford Pantera L
Duster 360
Camaro Z28
Hornet Sportabout
Pontiac Firebird
Hornet 4 Drive
Valiant
Merc 450SLC
Merc 450SE
Merc 450SL
Dodge Challenger
AMC Javelin
Honda Civic
Toyota Corolla
Fiat 128
Fiat X1-9
Ferrari Dino
Lotus Europa
Merc 230
Volvo 142E
Datsun 710
Toyota Corona
Porsche 914-2
Merc 240D
Mazda RX4
Mazda RX4 Wag
Merc 280
Merc 280C
d
hclust (*, "complete")

UNDERSTANDING RESULTS





For a data set with 4,000 elements, it takes hclust about 2 minutes to finish the job on an AMD Phenom II X4 CPU.