Introduction to Analytic Methods, Types of Data Mining for Analytics
&
Introduction to Group 2

Thilanka Munasinghe

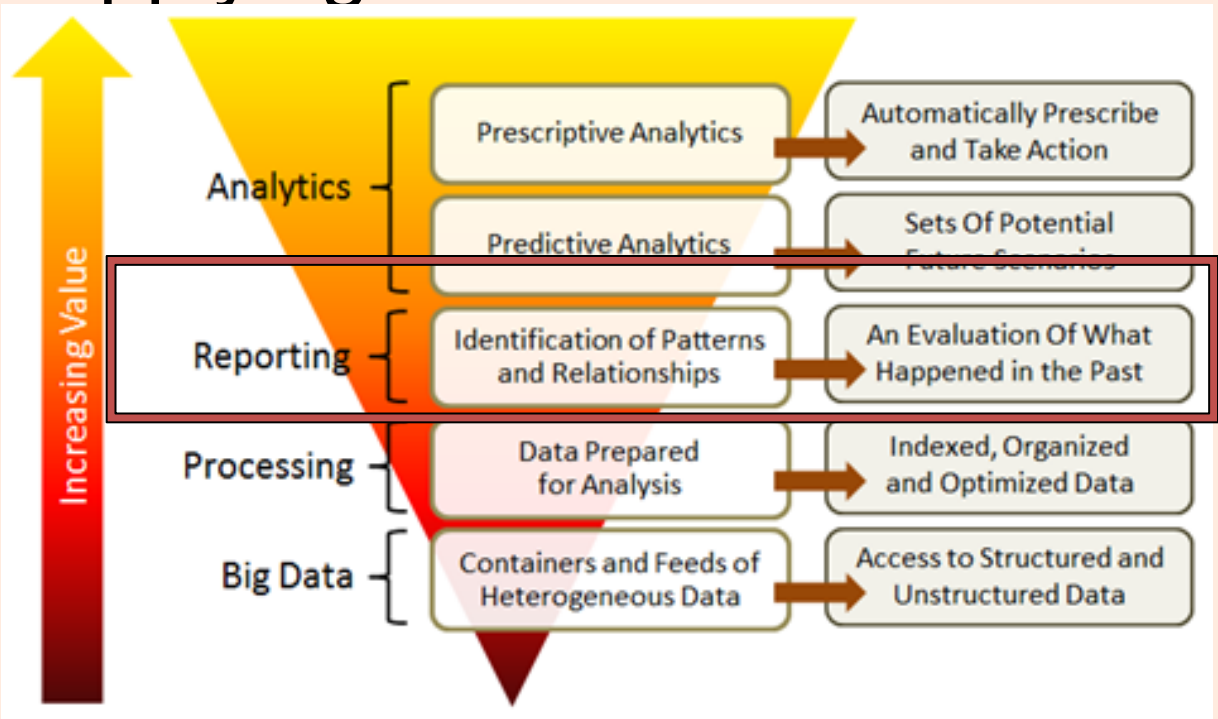
Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 1, Module 4, February 8th, 2021

Contents

- Reminder: PDA/EDA, models
- Patterns/ Relations via “Data mining”
- Interpreting results
- Saving the models
- Proceeding with applying the models



Preliminary Data Analysis

- Relates to the sample v. population
- Also called Exploratory DA
  - "EDA is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there , as well as those we believe will be there" (John Tukey)
- Distribution analysis and comparison, visual 'analysis', model testing, i.e. pretty much the things you did last lab and will do more of!

Models

- Assumptions are often used when considering models, e.g. as being representative of the *population* – since they are so often derived from a *sample* – this should be starting to make sense (a bit)
- Two key topics:
  - N=all and the open world assumption
  - Model of the thing of interest *versus* model of the data (data model; structural form)
- "All models are wrong but some are useful" (*generally attributed to the statistician ~ George Box*)

Art or science?

- The form of the model, incorporating the hypothesis determines a “form”
- Thus, as much art as science because it depends both on your world view and what the data is telling you (or not)
- We will however, be giving the models nice mathematical properties

Patterns and Relationships

Group 2 - Patterns, relations, descriptive analytics

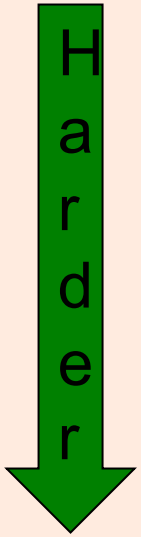
- Stepping from elementary/ distribution analysis to algorithmic-based analysis
- I.e. pattern detection via data mining: classification, clustering, rules; machine learning; support vector machines, non-parametric models
- Relations – associations between/among populations
- Outcome: model and an evaluation of its fitness for purpose

Data Mining = Patterns

- Classification (Supervised Learning)
  - Classifiers are created using labeled training samples
  - Training samples created by ground truth / experts
  - Classifier later used to classify unknown samples
- Clustering (Unsupervised Learning)
  - Grouping objects into classes so that similar objects are in the same class and dissimilar objects are in different classes
  - Discover overall distribution patterns and relationships between attributes
- Association Rule Mining
  - Initially developed for market basket analysis
  - Goal is to discover relationships between attributes
  - Uses include decision support, classification and clustering
- Other Types of Mining
  - Outlier Analysis
  - Concept / Class Description
  - Time Series Analysis

Models/ types

- Trade-off between Accuracy and Understandability
- Models range from “easy to understand” to incomprehensible
  - Decision trees
  - Rule induction
  - Multi-Regression models
  - Neural Networks
  - Deep Learning



Patterns and Relationships

In **Group 2 - Patterns, relations, descriptive analytics**

- Linear and multi-variate

- Nearest Neighbor
  - Training.. (supervised)
- K-means
  - Clustering.. (un-supervised) and classification

The Dataset(s)

- Open the *multivariate.xls* and explore the data
- What are your observations ?
- What can you say about the data set ?
- Same dataset is available as simple multivariate.csv

(http://aquarius.tw.rpi.edu/html/DA)

- Additionally, another dataset;

  nytimes/ nyt<n> and “sales”

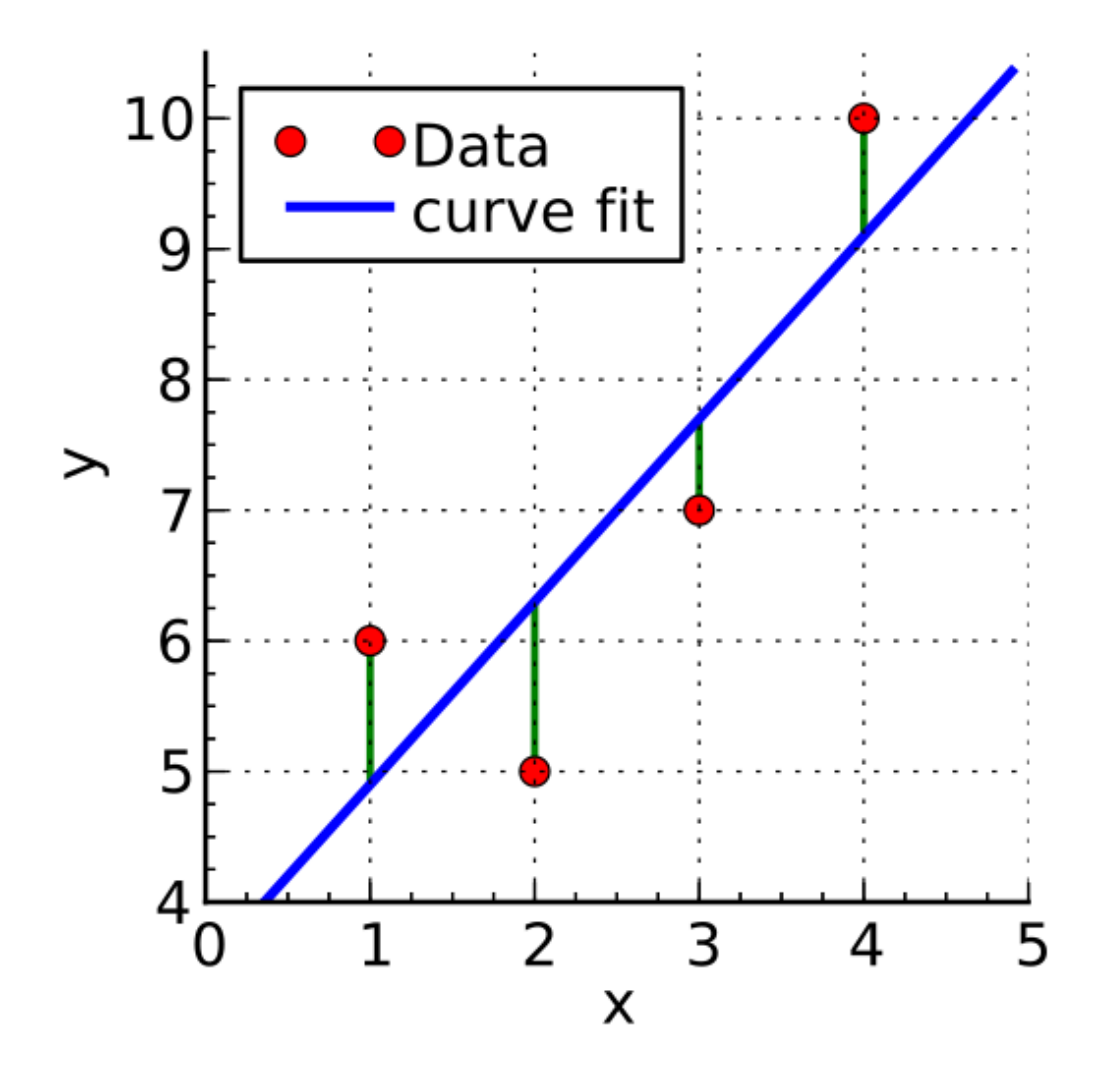
Regression in Statistics

- Regression is a statistical process for *estimating* the relationships among variables
- Includes many techniques for modeling and analyzing several variables
- When the focus is on the relationship between a dependent variable and one or more independent variables
- Independent variables are also called basis functions
- Estimation is often by constraining an objective function
- Must be tested for significance, confidence

Objective function

Constraint function(s)

regression…



Linear basis and least-squares constraints

```
> multivariate <-
read.csv("~/Documents/teaching/DataAnalytics/
data/multivariate.csv")
> attach(multivariate)
> mm<-lm(Homeowners~Immigrant)
> mm
```

dependent variable

independent variable

```
Call:
lm(formula = Homeowners ~ Immigrants)

Coefficients:
(Intercept)   Immigrants
     107495        -6657
```

	City	Income	Population	Immigrants	Homeowners	area
1	A	35460	20066	10.62	12081	50108
2	B	27038	70526	11.53	56518	141209
3	C	40337	99433	10.26	92478	199498
4	D	44833	27118	10.75	2920	54752
5	E	34590	86892	14.47	25331	173937
6	F	31205	34764	11.62	12764	70091
7	G	47102	41286	12.67	5044	82834

In Class Work

```
# Read the csv file
multivariate <- read.csv("~/Downloads/multivariate.csv")
attach(multivariate)
names(multivariate)
multivariate

# Create some Scatterplots
plot(Income,Immigrant, main = "Scatterplot")
plot(Immigrant,Homeowners)

# Fitting Linear Models using "lm" function
help(lm)
mm<-lm(Homeowners ~ Immigrant)
mm
plot(Immigrant,Homeowners)
```

In Class Work

```
# Fitting Linear Models using "lm" function
help(lm)
mm<-lm(Homeowners ~ Immigrant)
mm
plot(Immigrant,Homeowners)
abline(mm)
abline(mm, col=2, lwd=3)

summary(mm)
attributes(mm)
mm$coefficients
```

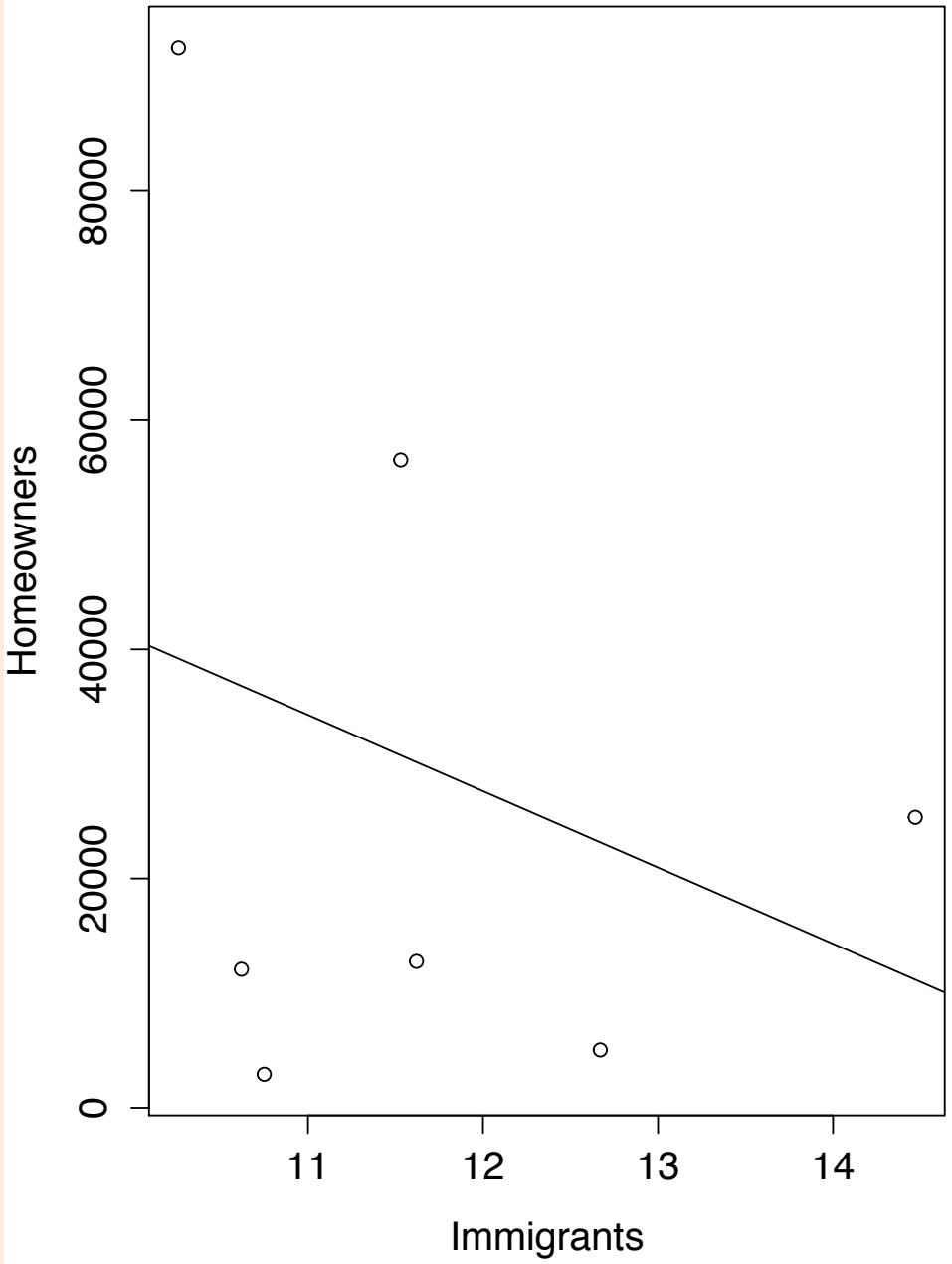
Linear fit?

- plot(Homeowners ~ Immigrants)

Find out how you can

Draw a line using

>**help(abline)**



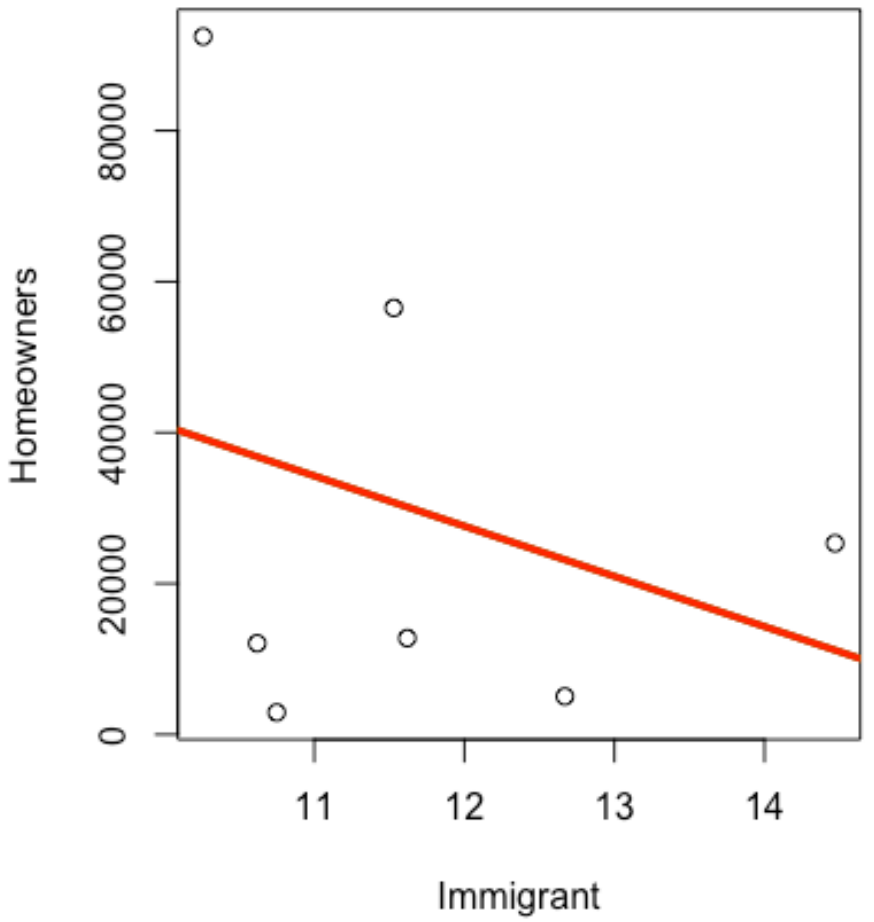
Linear fit?

- plot(Homeowners~Immigrants)

Find out how you can

Draw a line using

>**help(abline)**



Does this line fit any data points ? No

Analysis – i.e. Science question

- We want to see if there is a relation between immigrant population and the mean income, the overall population, the percentage of people who own their own homes, and the population density.
- To do so we solve the set of 7 linear equations of the form:
- %_immigrant = a x Income + b x Population + c x Homeowners/Population + d x Population/area + e

Multi-variate
(during the Lab Work)

Figure out how to include multiple independent variables

```
> HP<- Homeowners/Population
> PD<-Population/area
> mm<-lm(Immigrants~Income+Population+HP+PD)
> summary(mm)

Call:
lm(formula = Immigrants ~ Income + Population + HP + PD)
```

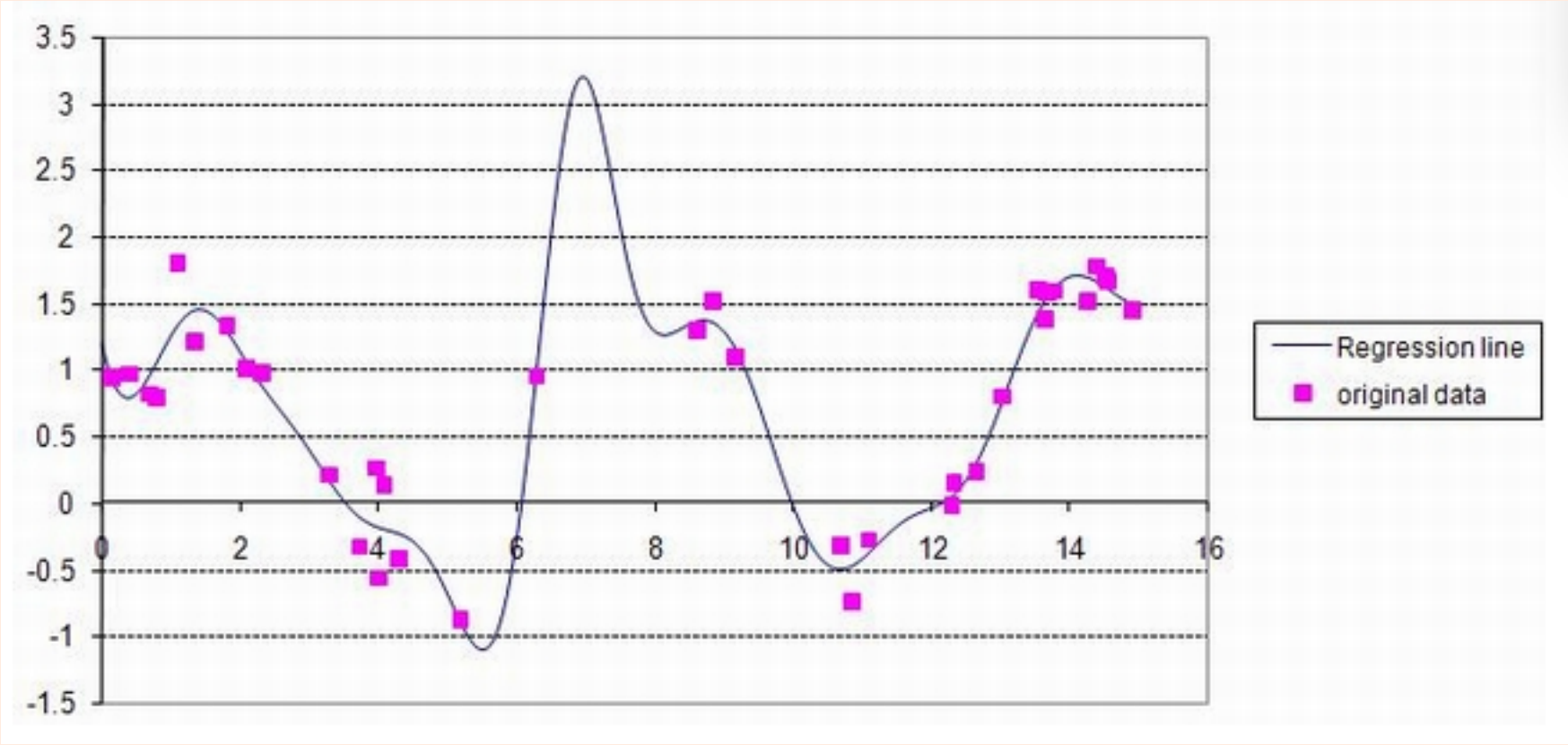
Multi-variate

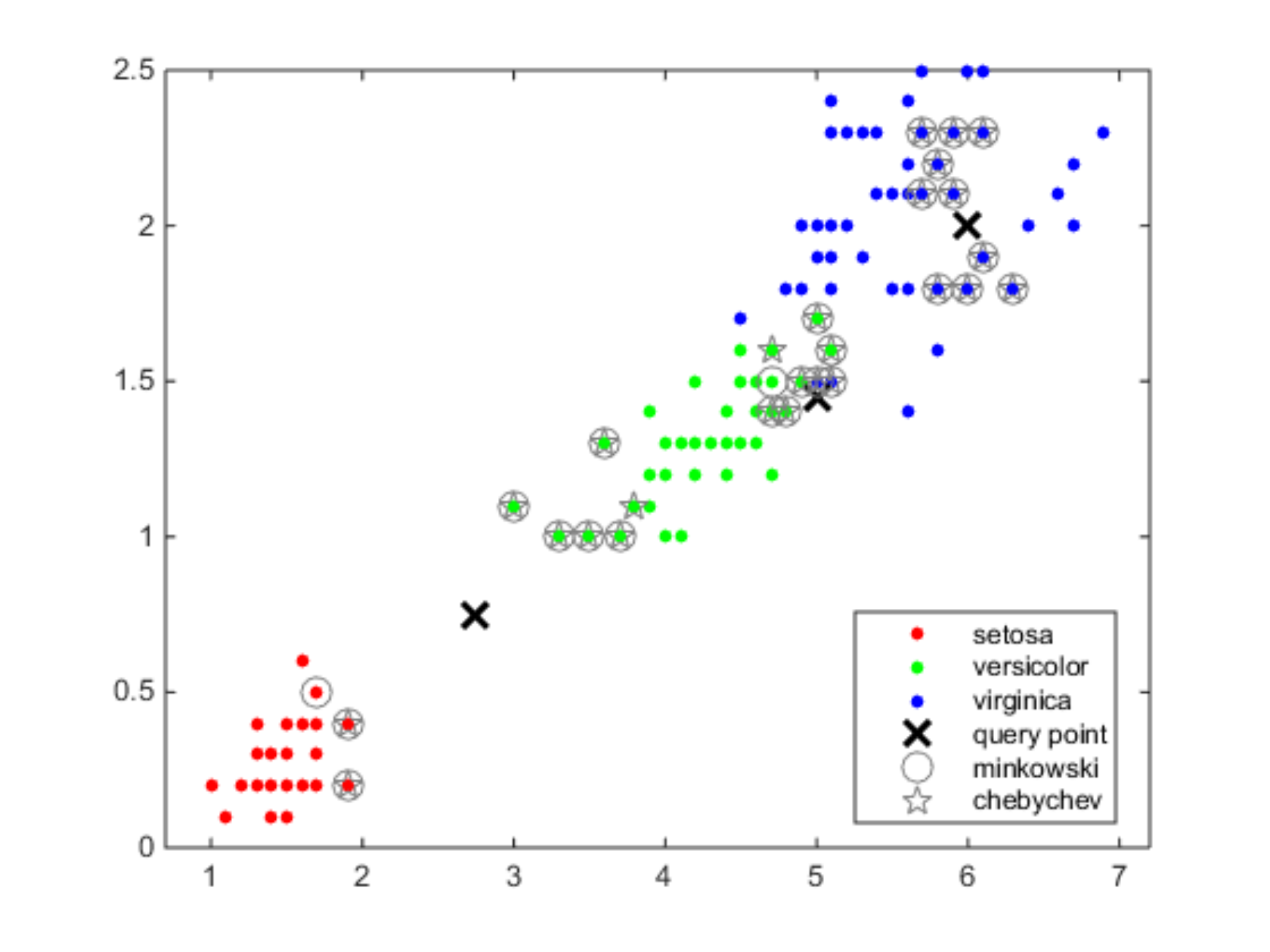
```
> cm<-coef(mm)
> cm
  (Intercept)        Income    Population            hp            pd
 2.454544e+01 -1.130049e-04  5.443904e-05 -6.533818e-02 -1.773908e-01
```

These linear model coefficients can be used with the predict.lm function to make predictions for new input variables. E.g. for the likely immigrant % given an income, population, %homeownership and population density

Oh, and you would probably try less variables?

When it gets complex…




2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
setosa
versicolor
virginica
query point
minkowski
chebychev

K-nearest neighbors (knn)

- Can be used in both regression and classification (“non-parametric”)
  - Is supervised, i.e. training set and test set
- KNN is a method for classifying objects based on closest training examples in the feature space.
- **An object is classified by a majority vote of its neighbors. K is always a positive integer**. The neighbors are taken from a set of objects for which the correct classification is known.
- It is usual to use the Euclidean distance, though other distance measures such as the Manhattan distance could in principle be used instead.

Algorithm

- The algorithm on how to compute the K-nearest neighbors is as follows:
  - Determine the parameter K = number of nearest neighbors beforehand. This value **is all up to you**.
  - Calculate the distance between the query-instance and all the training samples. You can use **any distance** algorithm.
  - Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
  - Since this is supervised learning, get all the categories of your training data for the sorted value which fall under K.
  - Use the majority of nearest neighbors as the prediction value.

Distance metrics

- **Euclidean** distance is the most common use of distance. When people talk about distance, this is what they are referring to. Euclidean distance, or simply 'distance', examines the root of square differences between the coordinates of a pair of objects. This is most generally known as the Pythagorean theorem.

- The **taxicab** metric is also known as **rectilinear** distance, L1 distance or L1 norm, city block distance, **Manhattan** distance, or Manhattan length, with the corresponding variations in the name of the geometry. It represents the distance between points in a city road grid. It examines the absolute differences between the coordinates of a pair of objects.

More generally

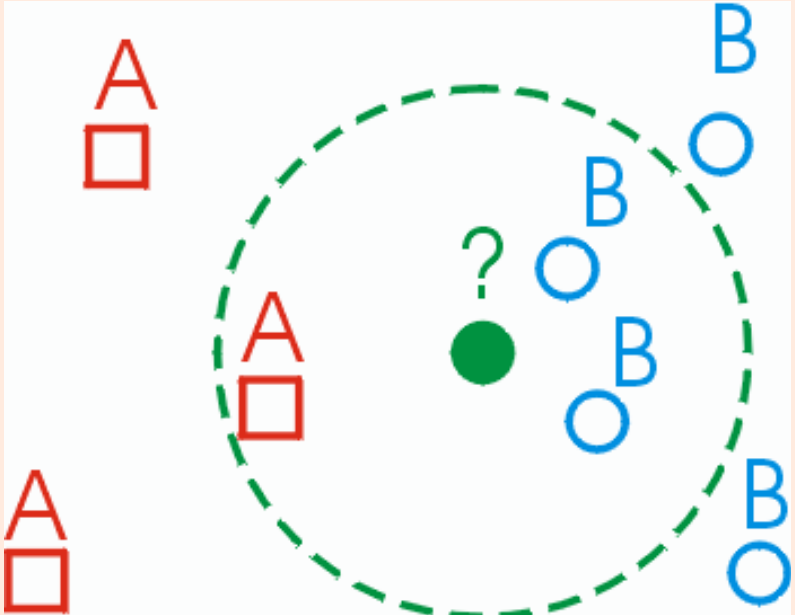
- The general metric for distance is the **Minkowski** distance. When lambda is equal to 1, it becomes the city block distance, and when lambda is equal to 2, it becomes the Euclidean distance. The special case is when lambda is equal to infinity (taking a limit), where it is considered as the Chebyshev distance.

- **Chebyshev** distance is also called the Maximum value distance, defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. In other words, it examines the absolute magnitude of the differences between the coordinates of a pair of objects.

Choice of k?

- Don’t you hate it when the instructions read: the choice of ‘k’ is all up to you ??

- Loop over different k, evaluate results…



What does “Near” mean…

- More on this in the next topic but …
  - DISTANCE – and what does that mean
  - RANGE – acceptable, expected?
  - SHAPE – i.e. the form

Training and Testing

- We are going to do much more on this going forward…
- Regression (un-supervised) – uses all the data to 'train' the model, i.e. calculate coefficients
  - Residuals are differences between actual and model for all data
- Supervision means not all the data is used to train because you want to test on the untrained set (before you predict for new values)
  - What is the 'sampling' strategy for training? (1b)

Summing up ‘knn’

- Advantages
  - Robust to noisy training data (especially if we use inverse square of weighted distance as the “distance”)
  - Effective if the training data is large
- Disadvantages
  - Need to determine value of parameter K (number of nearest neighbors)
  - Distance based learning is not clear which type of distance to use and which attribute to use to produce the best results. Shall we use all attributes or certain attributes only?
  - Computation cost is quite high because we need to compute distance of each query instance to all training samples. Some indexing (e.g. K-D tree) may reduce this computational cost.

K-means

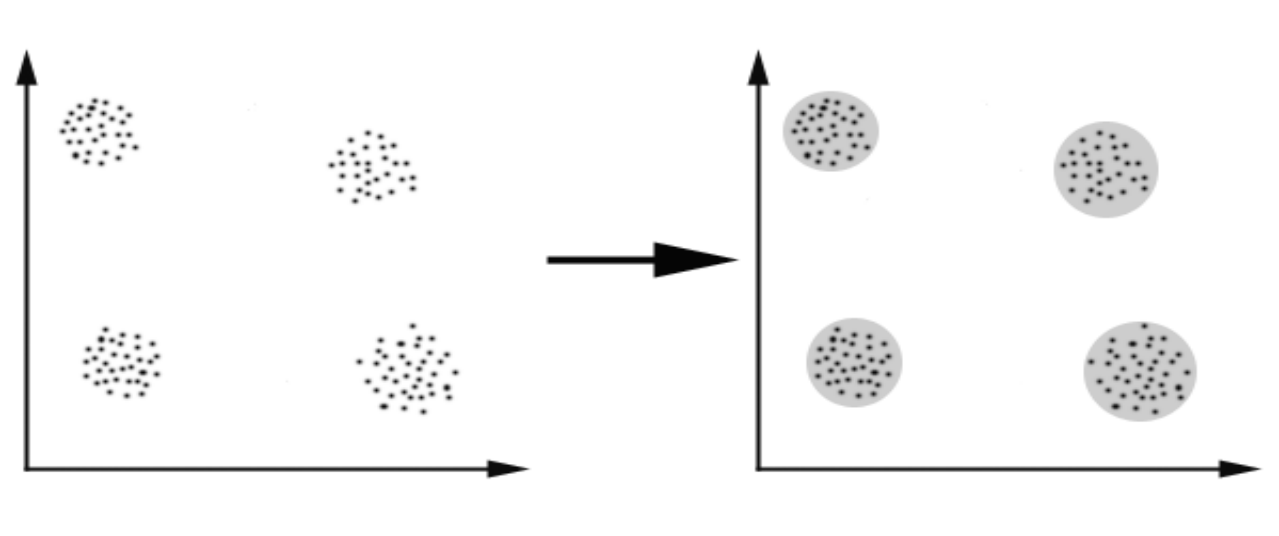
- Unsupervised classification, i.e. no classes known beforehand

- Types:
  - Hierarchical: Successively determine new clusters from previously determined clusters (parent/child clusters).

  - Partitional: Establish all clusters at once, at the same level.

Distance Measure

- Clustering is about finding "similarity".
- To find how similar two objects are, one needs a "distance" measure.
- Similar objects (same cluster) should be close to one another (short distance).

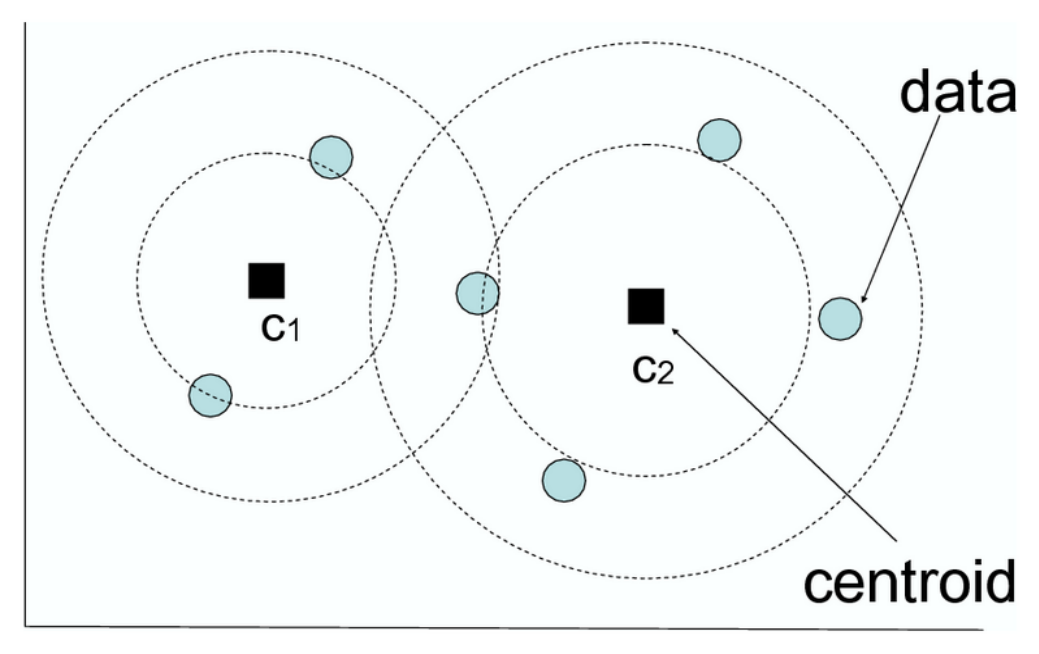


Distance Measure

- Many ways to define distance measure.
- Some elements may be close according to one distance measure and further away according to another.
- Select a good distance measure is an important step in clustering.

K-Means Clustering

- Separate the objects (data points) into K clusters.
- Cluster center (centroid) = the average of all the data points in the cluster.
- Assigns each data point to the cluster whose centroid is nearest (using distance function.)



K-Means Algorithm

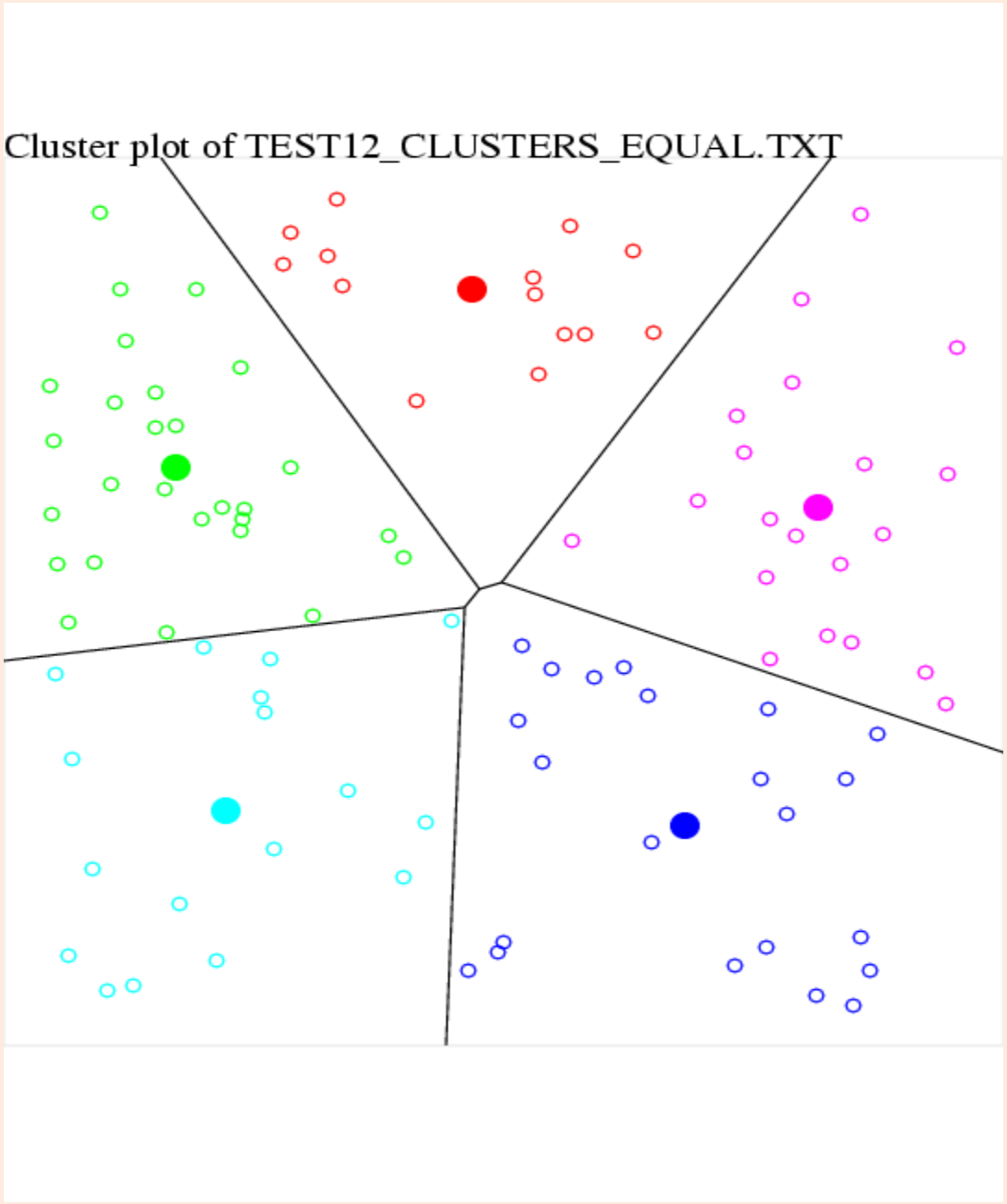
1. Place K points into the space of the objects being clustered. They represent the initial group centroids.

2. Assign each object to the group that has the closest centroid.

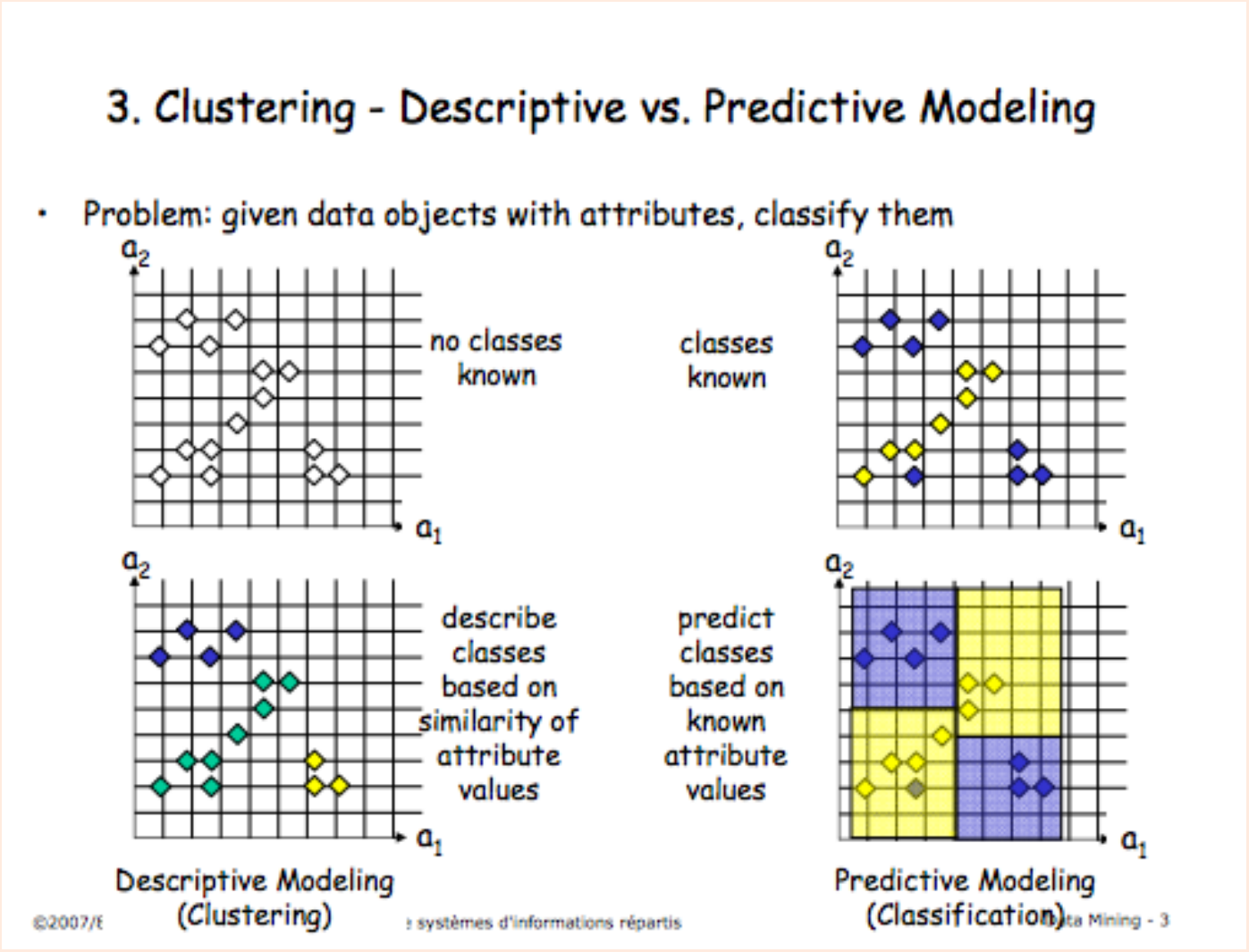
3. Recalculate the positions of the K centroids.

4. Repeat Steps 2 & 3 until the group centroids no longer move.

K-Means Algorithm: Example Output



Describe v. Predict



More on this later in Group 2 …

K-means

"Age","Gender","Impressions","Clicks","Signed_In"

36,0,3,0,1

73,1,3,0,1

30,0,3,0,1

49,1,3,0,1

47,1,11,0,1

47,0,11,1,1

(nyt datasets)

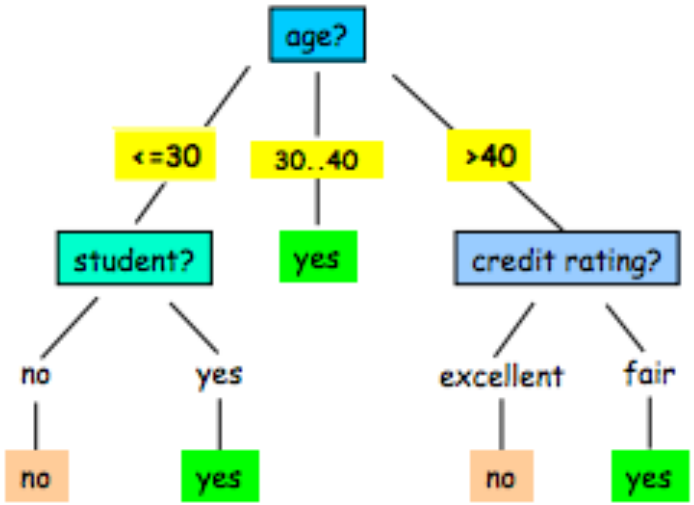
Model e.g.: If Age<45 and Impressions >5 then Gender=female (0)

Age ranges? 41-45, 46-50, etc?

Decision tree classifier

Classification by Decision Tree Induction

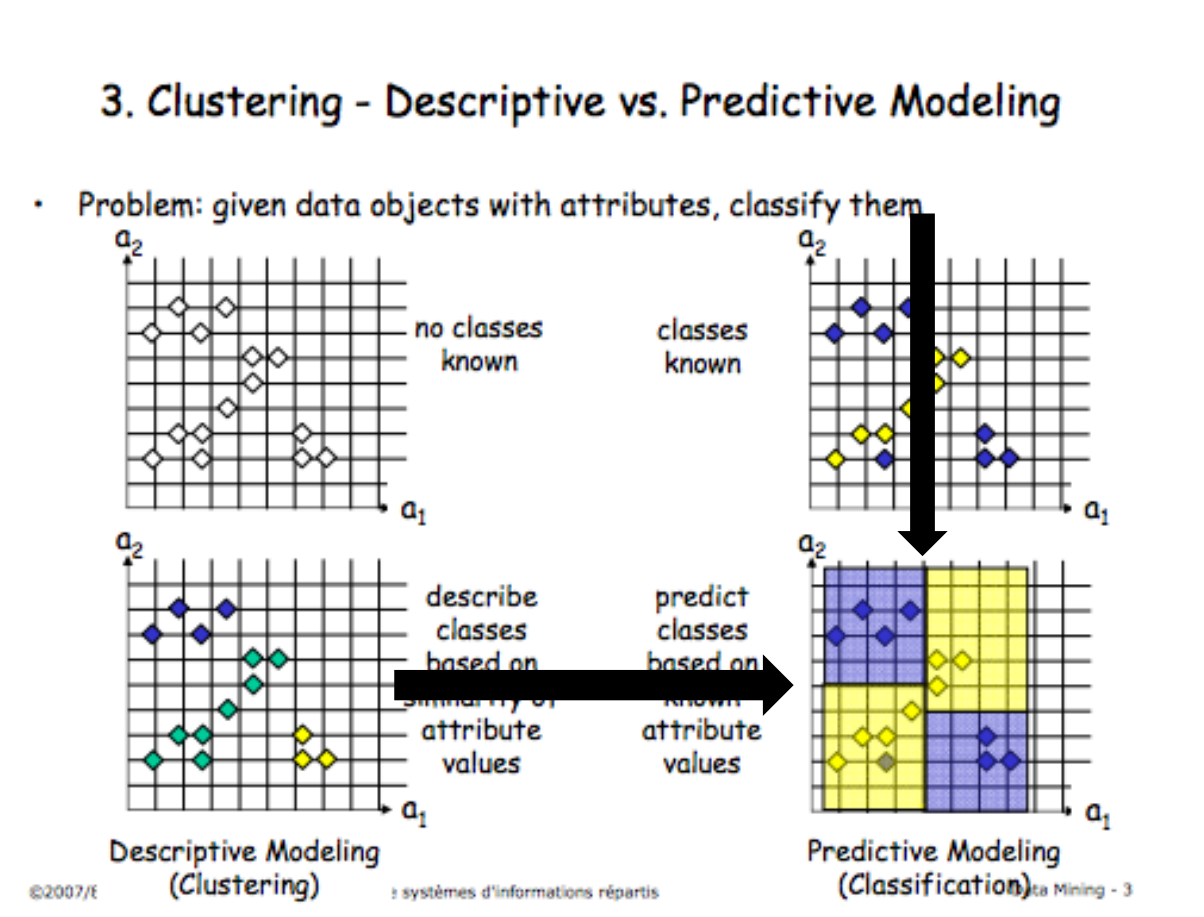
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



buys_computer ?

More on this later in Group 2 …

Predict = Decide



More on this later in Group 2 …

Visualization

- Scatter Plot – Paired data (x,y)
- Describe the relationship between numerical variables.
- Make a note on the direction of the data points
  - Positive direction
  - Negative Direction
- Check for unusual observations
- See the relationship - Linear or Non-linear

We'll do more during up coming lectures/labs..

- We will move to Group 2: Patterns, relations, descriptive analytics

Possible Project Ideas for the Data Analytic Course

- Sustainable Development Goals (SDG) using UN Data
- https://www.un.org/sustainabledevelopment/sustainable-development-goals/
- Watch:https://www.youtube.com/watch?time_continue=4&v=0XTBYMfZyrM&feature=emb_logo&ab_channel=UnitedNations
- https://www.un.org/sustainabledevelopment/
- https://data.un.org/
- https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html
- Watch:https://www.youtube.com/watch?v=yobFJniliOs&feature=emb_logo&ab_channel=WesternDigitalCorporation
- Watch: https://www.youtube.com/watch?v=v-zGHqMyd7o&feature=emb_logo&ab_channel=UNGlobalPulse

UN Data

https://www.un.org/sustainabledevelopment/

Project Dataset Selection

Project Dataset selection (some of you have already chosen a dataset by now, if not, you need to do it today)

Dataset Check-in (documenting your dataset:URLs/Weblinks/Repository information)

Next Week: Virtual One-on-One with the instructor to check-in your Project dataset

Dataset cleaning/preparation

Conducting EDA on your project dataset and preliminary analysis begins

Dataset search

- If you have not chosen a dataset so far, please search online and select datasets using search tools such as

  https://datasetsearch.research.google.com/

- If you need help choosing a dataset, please come and talk to me during the class time or during virtual office hours, so that I can guide/help you to select datasets.

- **NOTE: 6000-Level students MUST have TWO datasets (minimum two datasets) used during final project.**

More places to find data:

- US Government Data: https://www.data.gov/
- US Department of Agriculture: https://www.nass.usda.gov/Data_and_Statistics/index.php
- Center of Disease Control (CDC): https://www.cdc.gov/datastatistics/index.html
- US Financial Data: https://www.federalreserve.gov/data.htm
- European Union Open Data Portal: https://data.europa.eu/euodp/en/data/

Some Conferences that you can submit your Data Analytics project

- **IEEE GHTC2021** (Global Humanitarian Technology Conference) This is aUN Sustainable Development Goals related conference. **Paper deadline May 1st 2021**

https://ieeeghtc.org/

- **IEEE DSAA2021** (Data Science and Advanced Analytics)
- Paper deadline May 23rd 2021
- https://dsaa2021.dcc.fc.up.pt/calls/important-dates

- **ACM GoodIT2021** (International Conference on Information Technology for Social Good

http://www.grc.upv.es/goodit2021/authors/#important-dates

Read: **Chapter 3 - Linear Regression**

- **Read: Chapter 3 (Linear Regression), Introduction to Statistical Learning with Applications in R, 7th Edition**
- Next: Lab2 (Assignment 2) on Thursday.