

Machine Learning for Finance

Prepared by Dr. Diaa Salama

FNT633
KNN

Lecture 2





مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY

Machine Learning for Finance

Lecture 2

FNT633
Intro

Prepared by
Dr. Diao Salama

www.midocean.ae

- Associate Professor, Data Science Department, Faculty of Computers Science , Misr International University, On Leave from Faculty of Computers & Artificial Intelligence, Banha University, Banha, Egypt.
- Member, National Committee of Communications and Information Technology, Academy of Scientific Research & Technology, CAIRO, Egypt.
- Member, International Telecommunication Union (ITU), United Nations (UN)
- Member, International Federation for Information Processing - IFIP
- Winner of the Arab Republic of Egypt Encouragement Award for Advanced Science and Technology in the Field of Engineering Sciences 2021
- Winner of the Benha University Encouragement Award for Advanced Science and Technology in the Field of Engineering Sciences 2019
- Winner of the DELL EMC AWARD for the academic year 2015/2016, 2016/2017, and 2017/2018



Dr.Diao Salama Ph.D

One of the top 2% Scientists worldwide 2023 by Stanford university ranking.

Email: diaasalama@midocean.edu.km

Week

03

Machine Learning

Today

- What is machine learning?
- Supervised learning
- Unsupervised learning

- What is machine learning?
- Machine Learning applications and tasks
- Supervised learning KNN

- Used for
 - Classification
 - Regression

Week

03

What is machine learning?

Machine learning!

- Give the machine the ability how to solve some problems (e.g. statistical phenomenon) based on **incomplete information**
- Here, learning is not only a question of remembering but also of generalization to unseen cases
- **Learn by example**

- Suppose you are building a Face recognition program
- Supervised learning (Classification):
 - You have a dataset of faces image and other images.
 - The dataset is labeled (classified) into Faces images, and other images.
 - Program task know is to classify new images based on the available dataset.
- Unsupervised learning (Clustering):
 - You don't have a dataset, and program role know is to cluster similar images for a new given dataset into different group, e.g. it can distinguish that faces are very different from landscapes, which are very different from

Week

03

Supervised Learning

Supervised Classification

- Referred to as classification or pattern recognition.
- The goal is to find a functional mapping between the input data (patterns or examples) X , to a class label Y .
 - i.e. $Y=f(X)$
- A pattern is described by a set of features.
 - E.g. color of eyes, or distance between the eyes.
- Pattern recognition task can be viewed as a two dimensional matrix, whose axes are the examples and the features.

Pattern classification tasks

- Data collection and representation.
- Feature selection and/or feature reduction.
- Classification.

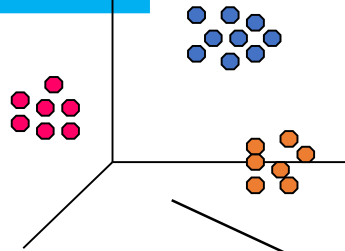
Classification algorithms

- Techniques for Low dimensional data sets
 - K-Nearest Neighbor Classification
 - Linear Discriminant Analysis
 - Decision Trees
 - Neural Networks
- Techniques for High dimensional data sets
 - Support Vector Machines
 - Boosting

Week

03

Machine Learning Tasks ...



Clustering

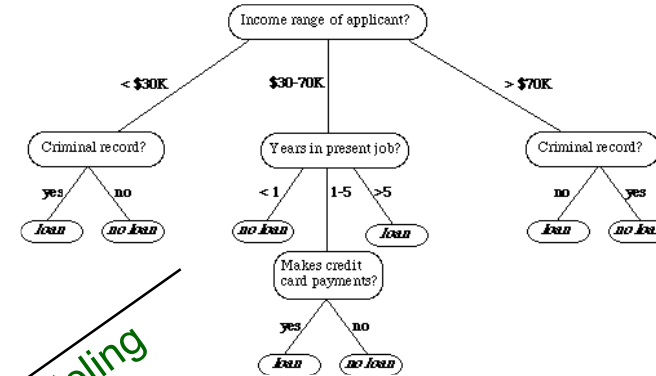
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

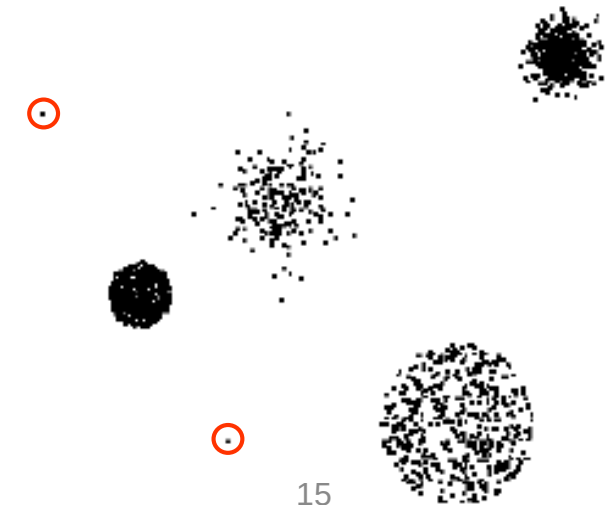
Association
Rules



Predictive Modeling



Anomaly
Detection



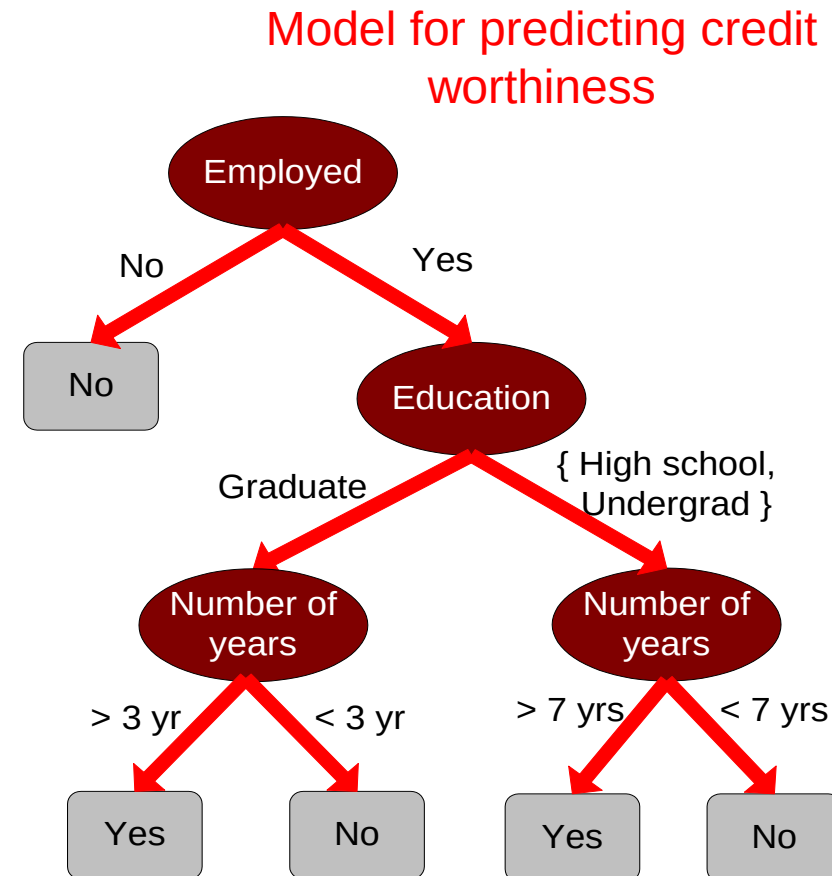
The business understanding (Define objectives)

- Define the project goals by asking and refining "sharp" questions that are relevant, specific, and unambiguous. Data science is a process that uses names and numbers to answer such questions. You typically use data science or machine learning to answer five types of questions:
 1. How much or how many? (regression)
 2. Which category? (classification)
 3. Which group? (clustering)
 4. Is this weird? (anomaly detection)
 5. Which option should be taken? (recommendation)
- Determine which of these questions you're asking and how answering it achieves your business goals.

- Find a model for class attribute as a function of the values of other attributes

Class

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...



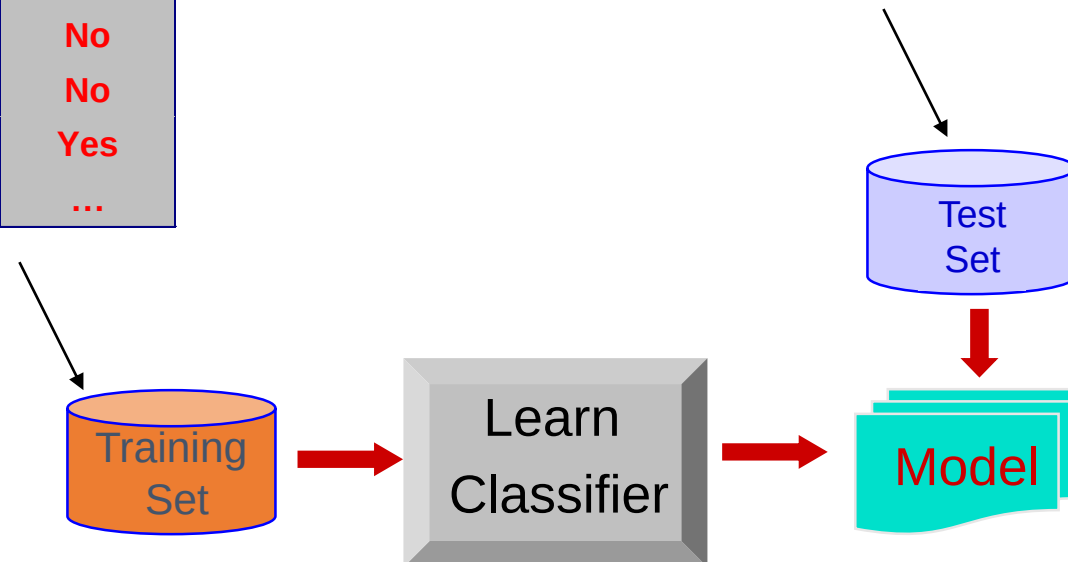
Classification Example

- Find a model for class attribute as a function of the values of other attributes

categorical *categorical* *quantitative* *class*

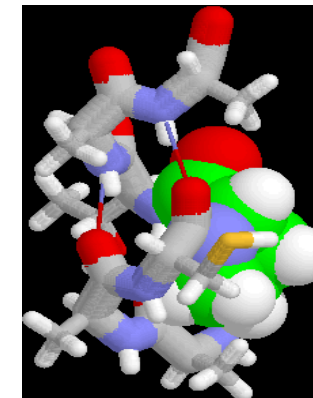
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...	...	...	...	...



Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

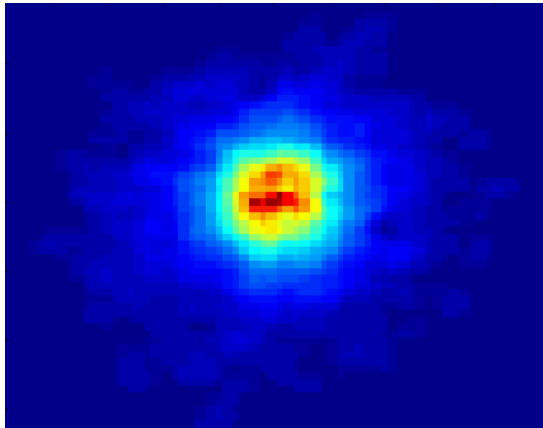


- Fraud Detection
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

- Churn prediction for telephone customers
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

- Sky Survey Cataloging
 - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
 - Approach:
 - Segment the image.
 - Measure image attributes (features) - 40 of them per object.
 - Model the class based on these features.
 - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Early



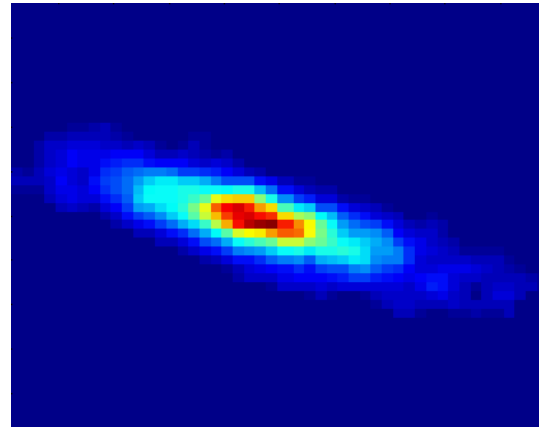
Class:

- Stages of Formation

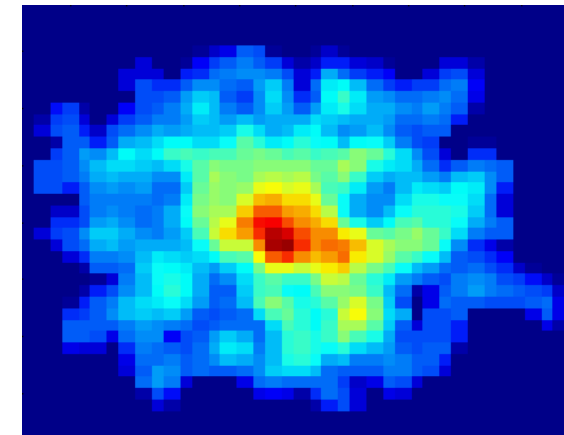
Attributes:

- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late

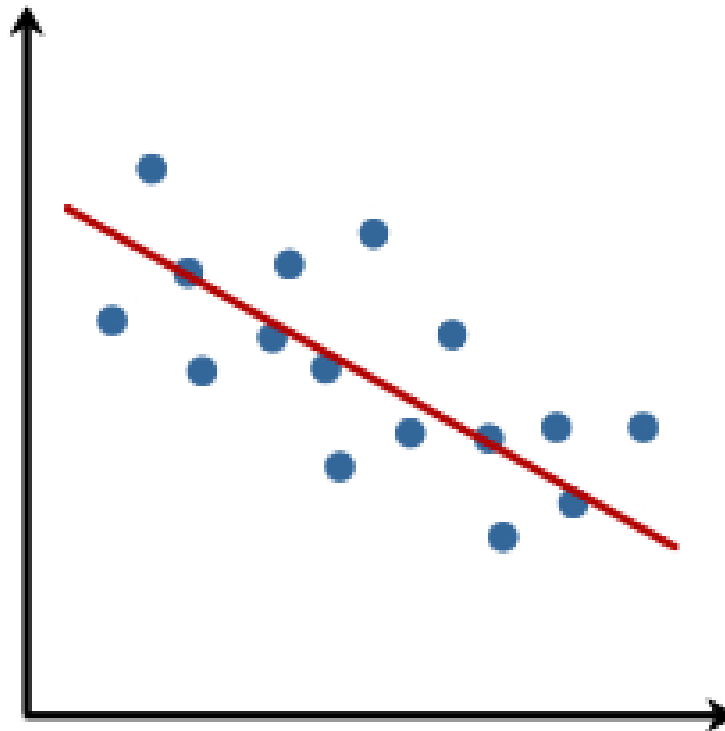


Data Size:

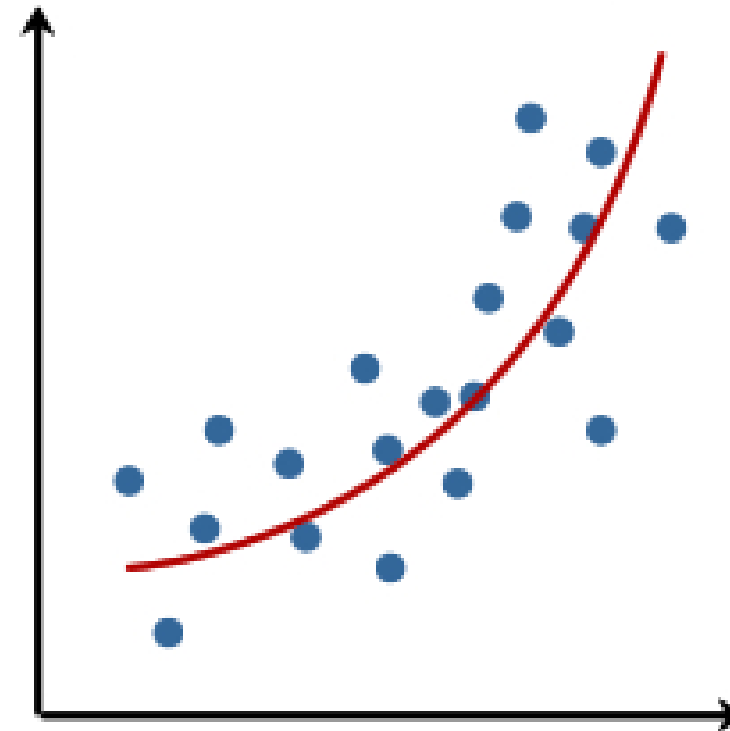
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

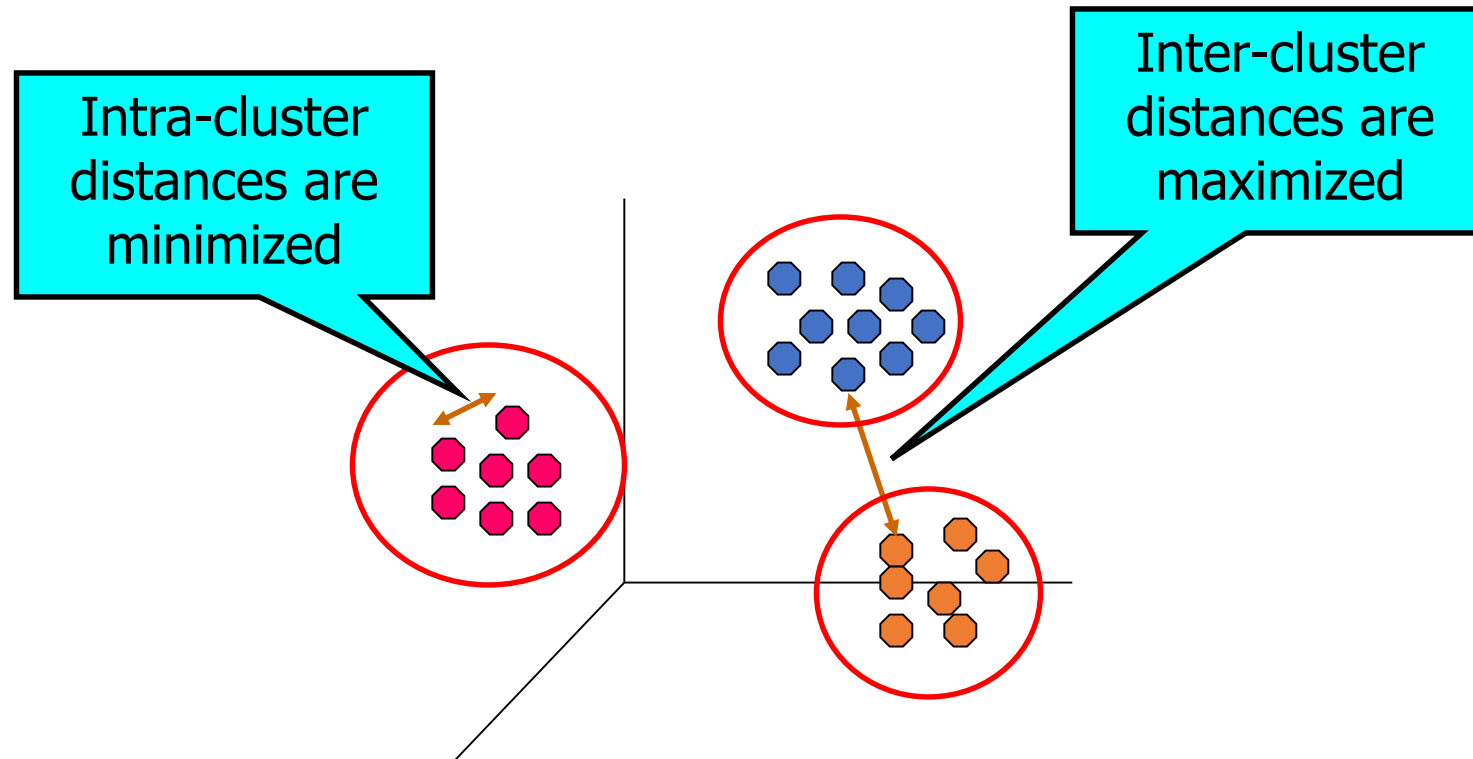
Linear



No linear relationship



- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

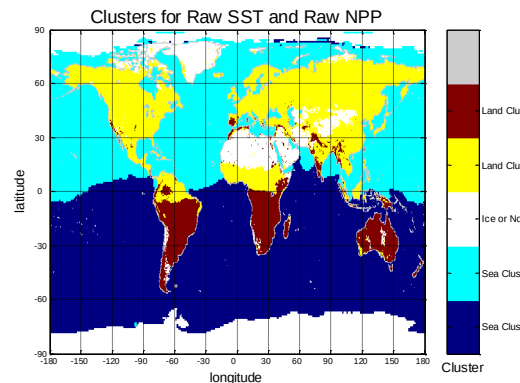


■ Understanding

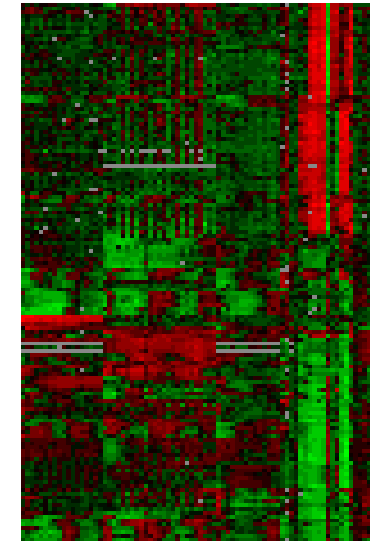
- Custom profiling for targeted marketing
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

■ Summarization

- Reduce the size of large data sets



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



- Market Segmentation:
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

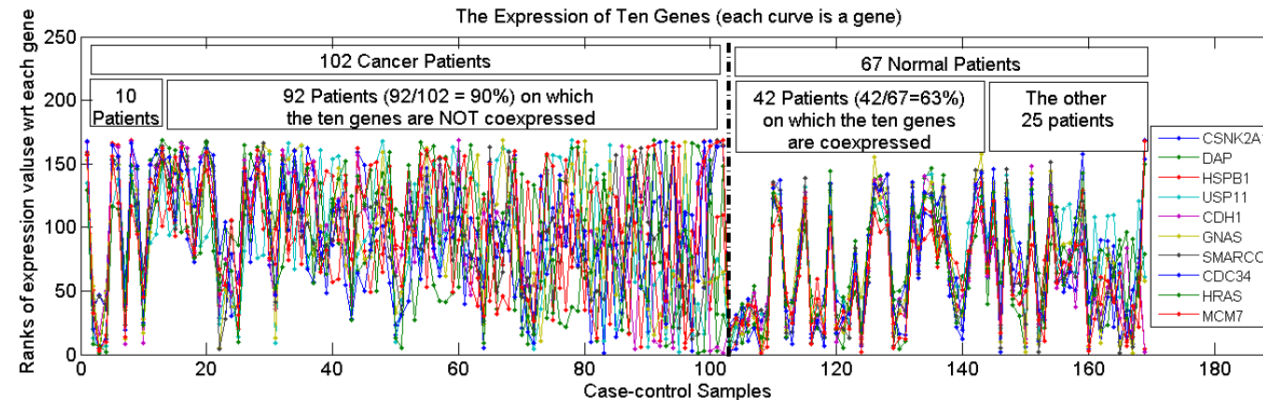
Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

- Market-basket analysis
 - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
 - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
 - Rules are used to find combination of patient symptoms and test results associated with certain diseases

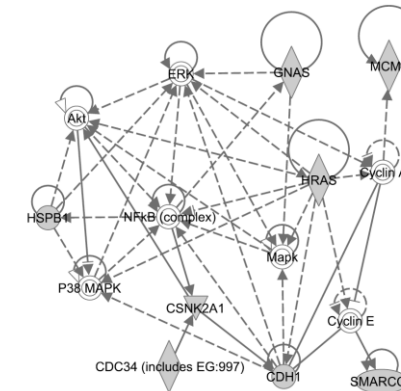
- An Example Subspace Differential Coexpression Pattern from lung cancer dataset
 - An Example Subspace Differential Coexpression Pattern from lung cancer dataset



Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]

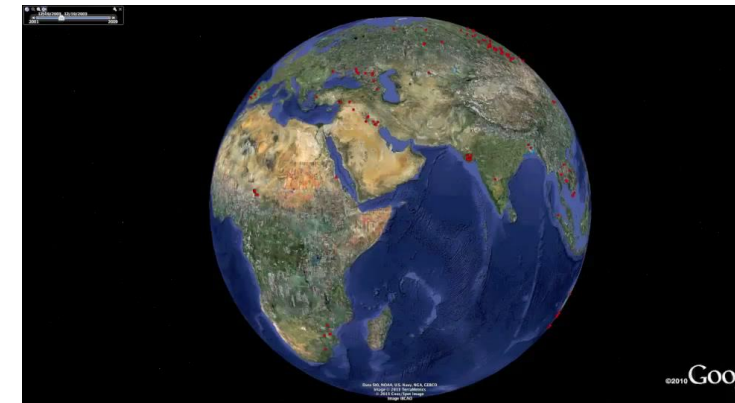
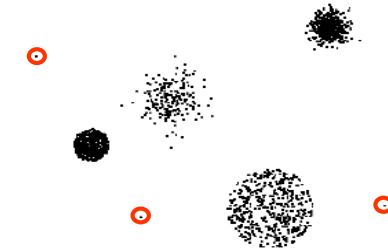
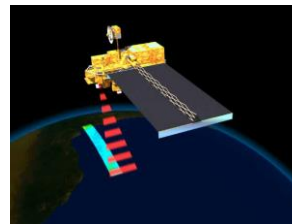
Enriched with the TNF/NFB signaling pathway
which is well-known to be related to lung cancer
P-value: 1.4×10^{-5} (6/10 overlap with the pathway)

[Fang et al PSB 2010]



Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection
 - Network Intrusion Detection
 - Identify anomalous behavior from sensor networks for monitoring and surveillance.
 - Detecting changes in the global forest cover.



Week

03

K-Nearest Neighbor Classification

K-NN

- Used for
 - Classification
 - Regression

K-Nearest Neighbor Classification

- K nearest neighbors measured by a distance function.
- These functions are valid only with continuous variables.

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^k |x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

KNN Example

- Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target.

Age	Loan	Default
25	\$40,000	N
35	\$60,000	N
45	\$80,000	N
20	\$20,000	N
35	\$120,000	N
52	\$18,000	N
23	\$95,000	Y
40	\$62,000	Y
60	\$100,000	Y
48	\$220,000	Y
33	\$150,000	Y

- Know we want to classify unknown case ((Age=48 and Loan=\$142,000) using Euclidean distance.)

!!!!

KNN Example

- Using Euclidean distance: E.g. distance between (48, \$142k) and (33, 150k)
- □ $\text{Sqrt}[(48-33)^2 + (142000-150000)^2] = 8000.01$
- So for all the dataset

Age	Loan	Default	Distance	
25	\$40,000	N	102000	
35	\$60,000	N	82000	
45	\$80,000	N	62000	
20	\$20,000	N	122000	
35	\$120,000	N	22000	2
52	\$18,000	N	124000	
23	\$95,000	Y	47000	
40	\$62,000	Y	80000	
60	\$100,000	Y	42000	3
48	\$220,000	Y	78000	
33	\$150,000	Y	8000	1
48	\$142,000	?		

Euclidean Distance

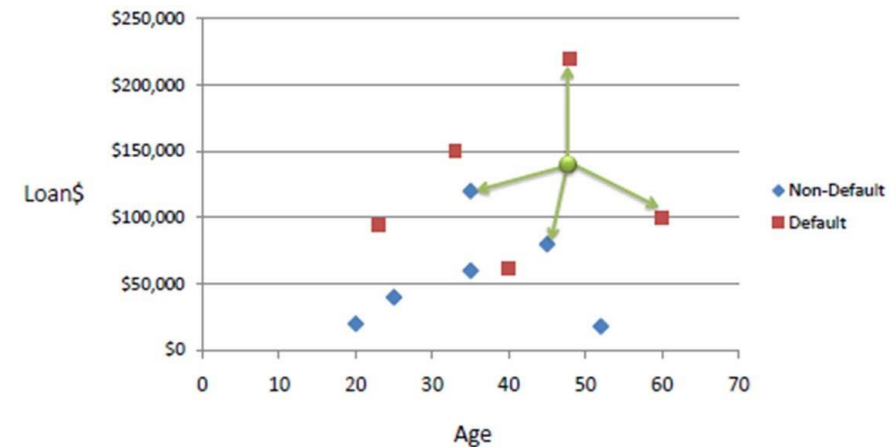
$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- Know we choose the number of neighbors K
 - If K=1 then this case is Y
 - If K=3 then we have 2 Y and 1 N which give us Y

Age	Loan	Default	Distance
25	\$40,000	N	102000
35	\$60,000	N	82000
45	\$80,000	N	62000
20	\$20,000	N	122000
35	\$120,000	N	22000
52	\$18,000	N	124000
23	\$95,000	Y	47000
40	\$62,000	Y	80000
60	\$100,000	Y	42000
48	\$220,000	Y	78000
33	\$150,000	Y	8000
48	\$142,000	?	

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$



1- Classification using KNN:

- Steps:
 - Find distance.
 - Find Rank.
 - Find the output according to the value of k.
- Euclidean Distance:

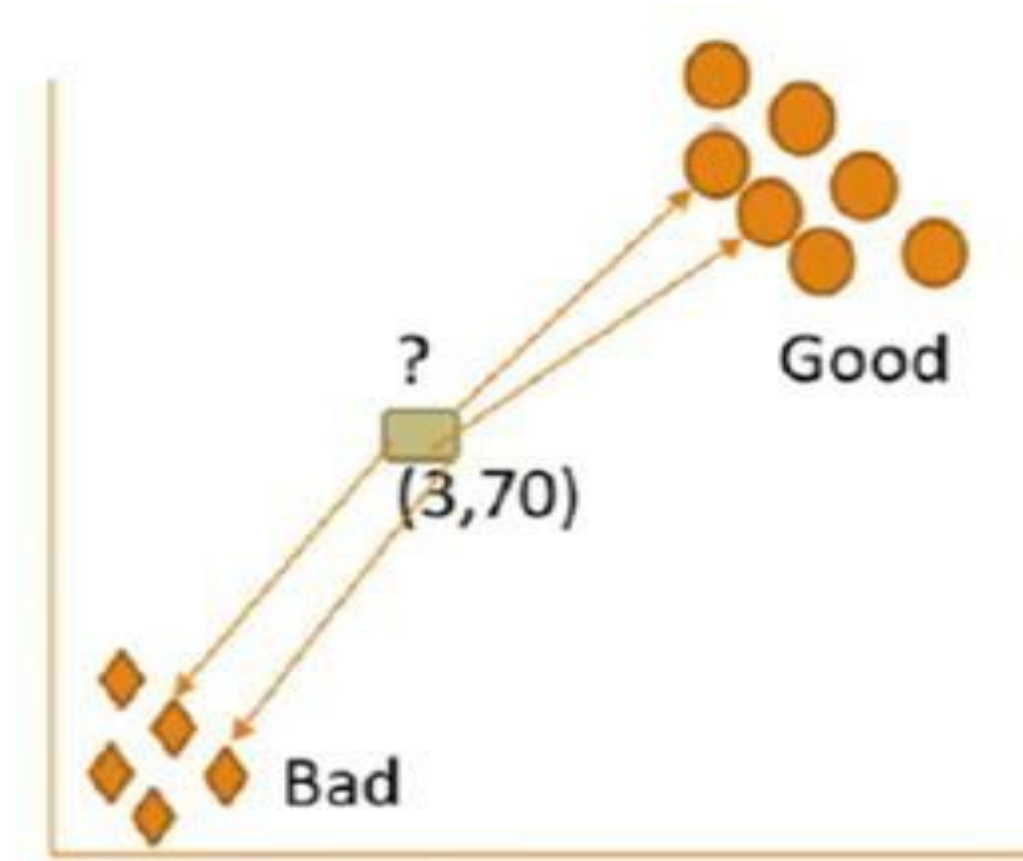
$$d = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$

Example

- Consider $k = 3$

Name	Cigarettes	Weight	Heart Attack	Distance	
A	7	70	Bad	$\sqrt{(3-7)^2 + (70-70)^2}$ $= 4$	①
B	7	40	Bad	$\sqrt{(3-7)^2 + (70-40)^2}$ $= 30.27$	④
C	3	40	Good	$\sqrt{(3-3)^2 + (70-40)^2}$ $= 30.00$	②
D	1	40	Good	$\sqrt{(3-1)^2 + (70-40)^2}$ $= 30.07$	③
E	3	70	Good		

- Consider $k = 4$



Example2 :

Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

Solution cont.:

Height (CM)	Weight (KG)	Class	Distance	Rank
169	58	Normal	1.4	1
170	55	Normal	2	2
173	57	Normal	3	3
174	56	Underweight	4.1	4
167	51	Underweight	6.7	5
173	64	Normal	7.6	6
172	65	Normal	8.2	7
182	62	Normal	13	8
176	69	Normal	13.4	9
170	57	?		

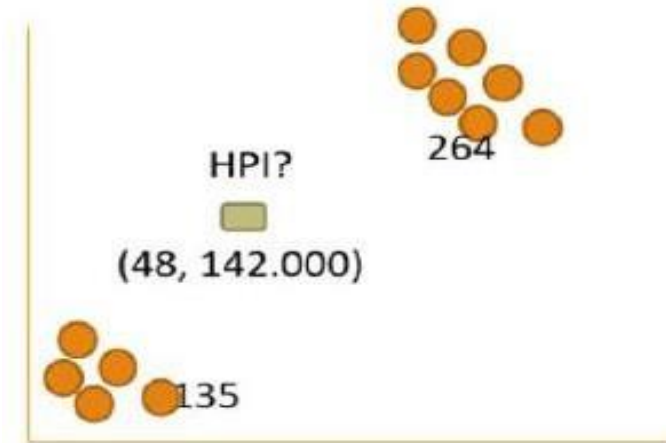
Solution cont.:

Height (CM)	Weight (KG)	Class	Distance	Rank
169	58	Normal	1.4	1
170	55	Normal	2	2
173	57	Normal	3	3
174	56	Underweight	4.1	4
167	51	Underweight	6.7	5
173	64	Normal	7.6	6
172	65	Normal	8.2	7
182	62	Normal	13	8
176	69	Normal	13.4	9
170	57	?		

- If $K=1$, Normal
- If $K=2$, Normal
- If $K=3$, Normal
- If $K=4$, Normal
- If $K=5$, Normal

2- Regression Using KNN:

Age	Loan	House Price Index	Distance
25	\$40,000	135	102000
35	\$60,000	256	82000
45	\$80,000	231	62000
20	\$20,000	267	122000
35	\$120,000	139	22000
52	\$18,000	150	124000
23	\$95,000	127	47000
40	\$62,000	216	80000
60	\$100,000	139	42000
48	\$220,000	250	78000
33	\$150,000	264	8000
48	\$142,000	?	



2- Regression Using KNN:

K=1

- using the training set to classify an unknown case
- (Age=33 and Loan=\$150,000) {Euclidean distance}.
- If K=1 then the nearest neighbor is the last case in the set
- with HPI=264.
- $D = \text{Sqrt}[(48-33)^2 + (142000-150000)^2] = 8000.01 \gg \text{HPI} = 264$

K=3

- the prediction for HPI is equal to the average of HPI for the top three neighbors
- $\text{HPI} = (264+139+139)/3 = 180.7$

Age	Loan	House Price Index	Distance
25	\$40,000	135	102000
35	\$60,000	256	82000
45	\$80,000	231	62000
20	\$20,000	267	122000
35	\$120,000	139	22000 2
52	\$18,000	150	124000
23	\$95,000	127	47000
40	\$62,000	216	80000
60	\$100,000	139	42000 3
48	\$220,000	250	78000
33	\$150,000	264	8000 1
48	\$142,000	?	

Example: Assume $k = 3$

x1	x2	weight
7	7	6
7	4	3
3	4	2
1	3	1
2	1	?

2. Data preparation

- The data is explored, putted into a suitable place and prepared to
- use in our machine learning training. In this step, we understand
- the nature of data that we must work with. We need to understand the characteristics, format, and quality of data.
- A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
- Data wrangling is the process of cleaning and converting raw data into a useable format.
- In real-world applications, collected data may have various issues

Solution: $K = 3$

$$\text{Weight} = \frac{3+2+1}{3} = 2$$

x1	x2	weight	Distance	Rank
7	7	6	7.81	4
7	4	3	5.83	3
3	4	2	3.16	2
1	3	1	2.23	1
2	1	2		

KNN disadvantages

- Computationally expensive
- Requires a large memory to store the training data.



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

