

CDS6214

Data Science Fundamentals

Lecture 5
Data Mining

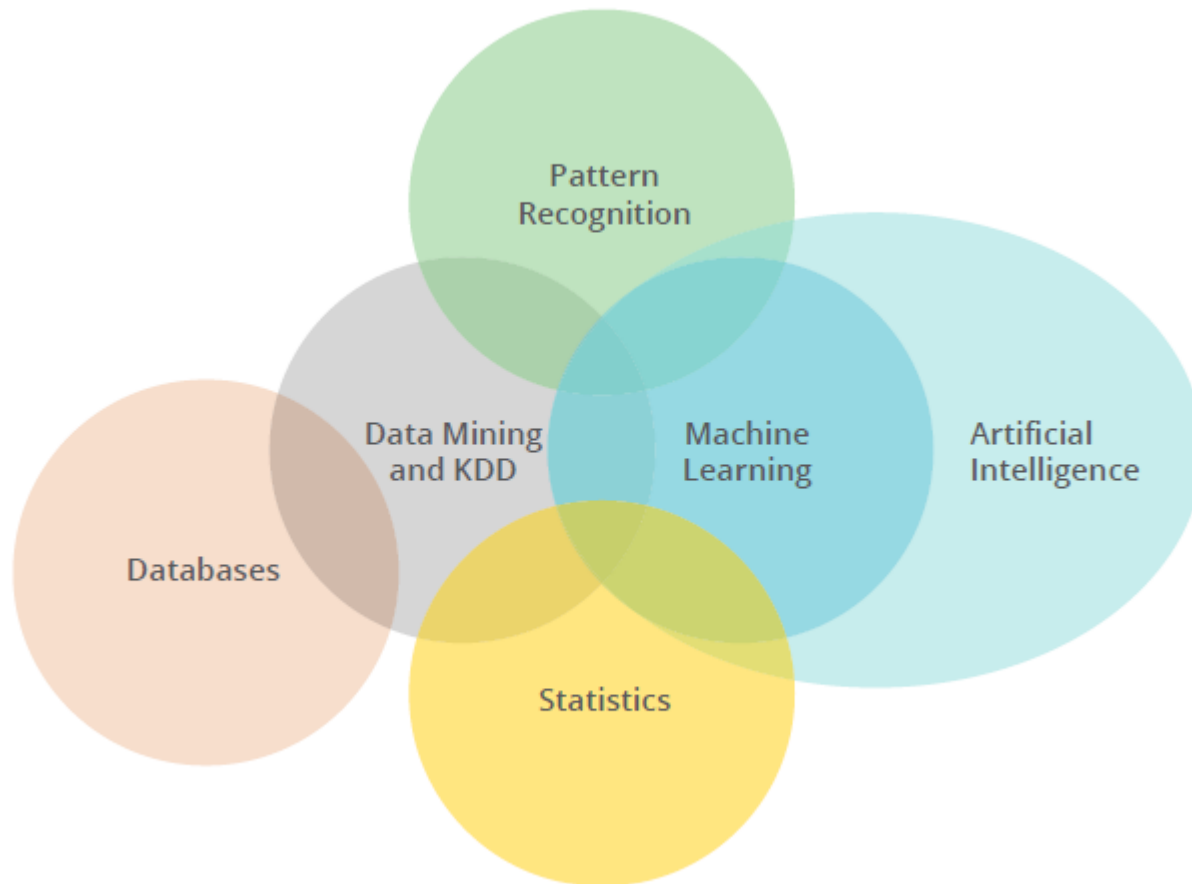
Outline

- AI vs. DM vs. ML
- **Data Mining Methods**
 - Grouping
 - Similarity Methods – Euclidean, Jaccard distances
 - Clustering – k-means Algorithm
 - Associative Rule Mining

AI vs. DM vs. ML

- ❖ **Artificial Intelligence:** the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.
- ❖ **Data Mining:** the practice of examining large databases to discover patterns or new information.
- ❖ **Machine Learning:** machine learning is the training of a model from data that generalizes a decision (prediction) against a performance measure.

AI vs. DM vs. ML



Machine Learning

- ❖ **Unsupervised Learning:** The program is given a bunch of data and must find patterns and relationships. Learning is not supervised by any known facts (or called 'labels') ⇒ discover from the data itself
 - ❑ **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior e.g. k-means etc.
 - ❑ **Association:** An association rule learning problem is where you want to **discover rules that describe large portions of your data**, such as people that buy X also tend to buy Y e.g. apriori algorithm for association rule mining etc.

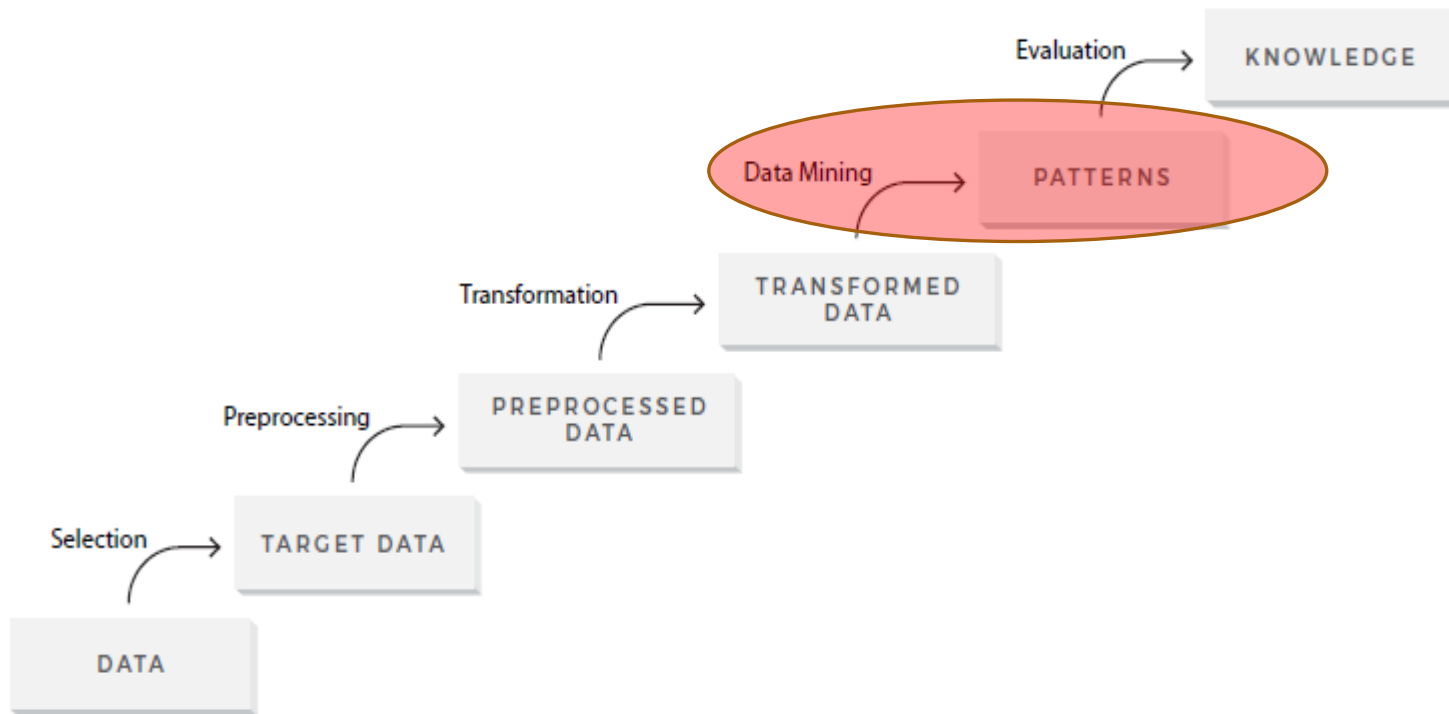
Machine Learning

- ❖ **Supervised Learning:** Learning is done on a pre-defined set of “training examples”, based on known facts (or ‘labels’), the learned model has the ability to reach an accurate conclusion when given new data.
 - ❑ **Classification:** A classification problem **predicts an output variable as a category**, such as “red” or “blue” or “disease” and “no disease”.e.g. Decision trees, support vector machine etc.
 - ❑ **Regression:** A regression problem **predicts an output variable as a real value**, such as “5.50” for dollars or “178” for weight. E.g linear regression etc.

Reinforcement Learning

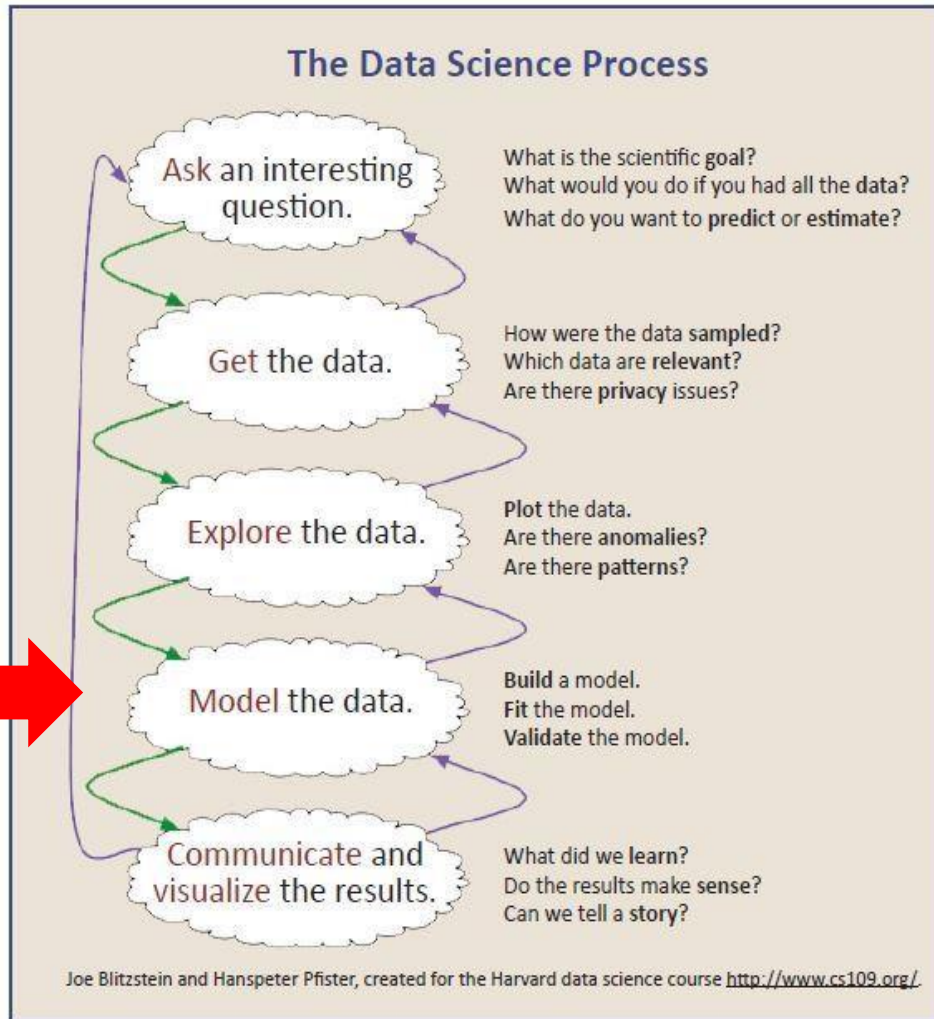
- ❖ **Reinforcement Learning:** Using this algorithm, the machine learns to make specific decisions by reinforcing experiences
 - ❖ The machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.
 - ❖ E.g. Markov Decision Process

Knowledge Discovery in Databases (KDD)



Concept introduced by Gregory Piatetsky-Shapiro in 1984
Looks familiar?

Data Science Process



What should we do next?

- Identify the question
- Collect and pre-process the data
- **Model the data**
- Infer and visualize results

Part of Data Mining = Unsupervised ML

Grouping

Clustering

Associative Rule Mining

data mining methods

Grouping

Dividing a data set into smaller subsets of related observations or groups is important for exploratory data analysis and data mining for a number of reasons:

- ❖ **Finding hidden relationships:** Grouping methods organize observations in different ways.
 - ❖ Example: A data set of retail transactions is grouped and these groups are often used to find nontrivial associations, such as customers who purchase doormats and umbrellas at the same time.

Grouping

- ❖ **Becoming familiar with the data:** Grouping methods allows us to discover which types of observations are present in the data.
- ❖ Example: A database of medical records can be used to create a general model for predicting a number of medical conditions. Before creating the model, the data set is characterized by grouping the observations. This reveals that a significant portion of the data consists of young female patients having flu.

Grouping

- ❖ **Segmentation**: Techniques for grouping data may lead to divisions that simplify the data for analysis.
 - ❖ Example, when building a model that predicts car fuel efficiency, it may be possible to group the data to reflect the underlying technology platforms the cars were built on. Generating a model for each of these 'platform-based' subsets will result in simpler models.

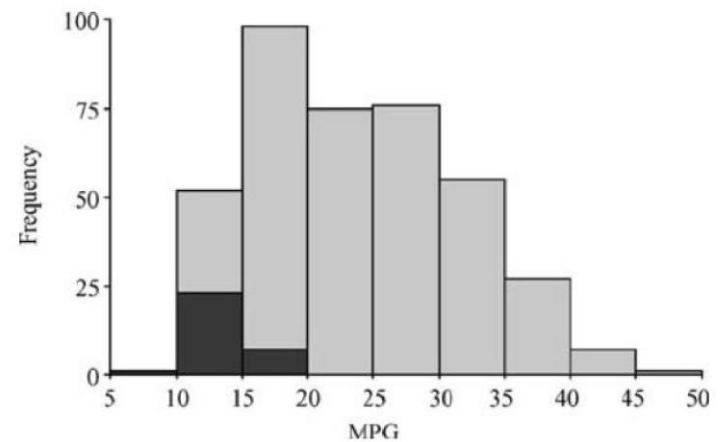
Grouping by Value/Range

- ❖ One way of creating a group is to search or query the data set. Each query would bring back a subset of observations. This set could then be examined to determine whether some interesting relationship exists.
- ❖ Example: In looking for hidden relationships that influence car fuel efficiency, we may query the data set in a variety of ways. The query could be:
 - ❖ By a single value, e.g. number-of-cylinders = 4
 - ❖ By a range of values, e.g. Weight < 4000
 - ❖ By Boolean combination of query terms to create more complex queries, e.g. number-of-cylinders = 4 && Weight < 4000

Grouping by Value/Range

- ❖ **Query 1:** All cars where Horsepower is greater than or equal to 160 AND Weight is greater than or equal to 4000.

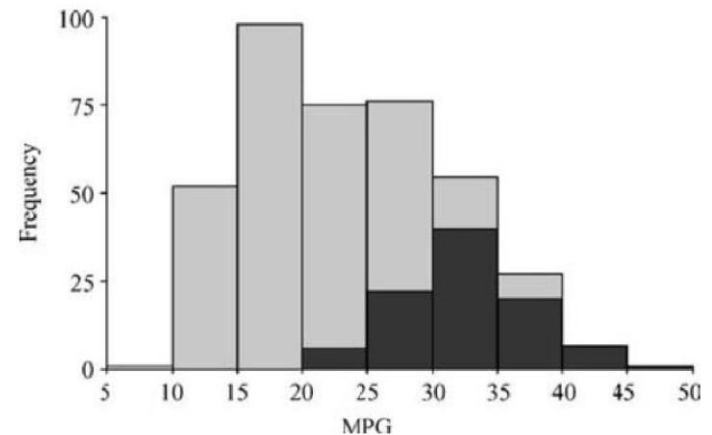
Names	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model/Year	Origin	MPG
Ford Galaxie 500	8	429	198	4,341	10	1970	1	15
Chevrolet Impala	8	454	220	4,354	9	1970	1	14
Plymouth Fury III	8	440	215	4,312	8.5	1970	1	14
Pontiac Catalina	8	455	225	4,425	10	1970	1	14
Ford F250	8	360	215	4,615	14	1970	1	10
Chevy C20	8	307	200	4,376	15	1970	1	10
Dodge D200	8	318	210	4,382	13.5	1970	1	11
Hi 1200d	8	304	193	4,732	18.5	1970	1	9
Pontiac Catalina Brougham	8	400	175	4,464	11.5	1971	1	14
Dodge Monaco (SW)	8	383	180	4,955	11.5	1971	1	12



Grouping by Value/Range

- ❖ **Query 2:** All cars where Horsepower is less than 80 AND Weight is less than 2500.

Names	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model/Year	Origin	MPG
Volkswagen 1131								
Deluxe Sedan	4	97	46	1,835	20.5	1970	2	26
Chevrolet Vega (SW)	4	140	72	2,408	19	1971	1	22
Peugeot 304	4	79	70	2,074	19.5	1971	2	30
Fiat 124B	4	88	76	2,065	14.5	1971	2	30
Toyota Corolla 1200	4	71	65	1,773	19	1971	3	31
Datsun 1200	4	72	69	1,613	18	1971	3	35
Volkswagen model 111	4	97	60	1,834	19	1971	2	27
Plymouth Cricket	4	91	70	1,955	20.5	1971	1	26
Volkswagen type 3	4	97	54	2,254	23.5	1972	2	23
Renault 12 (SW)	4	96	69	2,189	18	1972	2	26



Grouping by Value/Range

- ❖ Can you think of a drawback of using query to group observations?

Similarity Measures

- ❖ Intuition: Any method of grouping needs to have an understanding for **how similar** observations are to each other.
- ❖ To determine how similar two observations are to each other we need to **compute the distance** between them.
- ❖ Example:
 - ❖ Euclidean Distance (handles continuous variables)
 - ❖ Jaccard Distance (handles binary variables)
 - ❖ There are other distance measures possible

Euclidean Distance

- ❖ Euclidean Distance (d) calculates the distance between p and q , where each observation has n variables.

$$d = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

- ❖ Three observations with values for five variables

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

Euclidean Distance

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

The Euclidean distance between A and B is

$$d_{A-B} = \sqrt{(0.7 - 0.6)^2 + (0.8 - 0.8)^2 + (0.4 - 0.5)^2 + (0.5 - 0.4)^2 + (0.2 - 0.2)^2}$$
$$d_{A-B} = 0.17$$

The Euclidean distance between A and C is

$$d_{A-C} = \sqrt{(0.7 - 0.8)^2 + (0.8 - 0.9)^2 + (0.4 - 0.7)^2 + (0.5 - 0.8)^2 + (0.2 - 0.9)^2}$$
$$d_{A-C} = 0.83$$

The Euclidean distance between B and C is

$$d_{B-C} = \sqrt{(0.6 - 0.8)^2 + (0.8 - 0.9)^2 + (0.5 - 0.7)^2 + (0.4 - 0.8)^2 + (0.2 - 0.9)^2}$$
$$d_{B-C} = 0.86$$

- ❖ The distance between A and B is 0.17, indicating that there is **more similarity** between these observations than A and C (0.83). C is not so closely related to either A or B.

Jaccard Distance

You will need to form a contingency table.

		Observation 2	
		1	0
Observation 1	1	$Count_{11}$	$Count_{10}$
	0	$Count_{01}$	$Count_{00}$

The formula to calculate Jaccard distance is:

$$d = \frac{Count_{10} + Count_{01}}{Count_{11} + Count_{10} + Count_{01}}$$

- $Count_{11}$: Count of all variables that are 1 in “Observation 1” and 1 in “Observation 2”.
- $Count_{10}$: Count of all variables that are 1 in “Observation 1” and 0 in “Observation 2”.
- $Count_{01}$: Count of all variables that are 0 in “Observation 1” and 1 in “Observation 2”.
- $Count_{00}$: Count of all variables that are 0 in “Observation 1” and 0 in “Observation 2”.

Jaccard Distance

Jaccard Distance

- You will need to form a contingency table.

		Observation 2	
		1	0
Observation 1	1	Count ₁₁	Count ₁₀
	0	Count ₀₁	Count ₀₀

The formula to calculate Jaccard distance is:

$$d = \frac{Count_{10} + Count_{01}}{Count_{11} + Count_{10} + Count_{01}}$$

- Count₁₁: Count of all variables that are 1 in “Observation 1” and 1 in “Observation 2”.
- Count₁₀: Count of all variables that are 1 in “Observation 1” and 0 in “Observation 2”.
- Count₀₁: Count of all variables that are 0 in “Observation 1” and 1 in “Observation 2”.
- Count₀₀: Count of all variables that are 0 in “Observation 1” and 0 in “Observation 2”.

Table of observations with values for five binary variables

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	1	1	0	0	1
B	1	1	0	0	0
C	0	0	1	1	1

Can you calculate the Jaccard distance between A and B ?

Grouping: Issues to consider

- ❖ **Supervised vs. unsupervised methods** to find hidden relationships / association
 - ❖ Methods that do not use any data to guide how the groups are generated are 'unsupervised'.
 - ❖ Methods that make use of a response variable to guide group generation are 'supervised'.
- ❖ **Types of variables:**
 - ❖ Certain grouping methods will only accept categorical data whereas others only accept continuous data
 - ❖ You might need to do data transformation!

Grouping: Issues to consider

- ❖ **Data size limit**

- ❖ Some methods work with smaller datasets. When dataset is too large, it becomes extremely inefficient or not able to group well

- ❖ **Interpretable and actionable**

- ❖ Certain grouping methods generate results that are easy to interpret whereas other methods require additional analysis to interpret the results

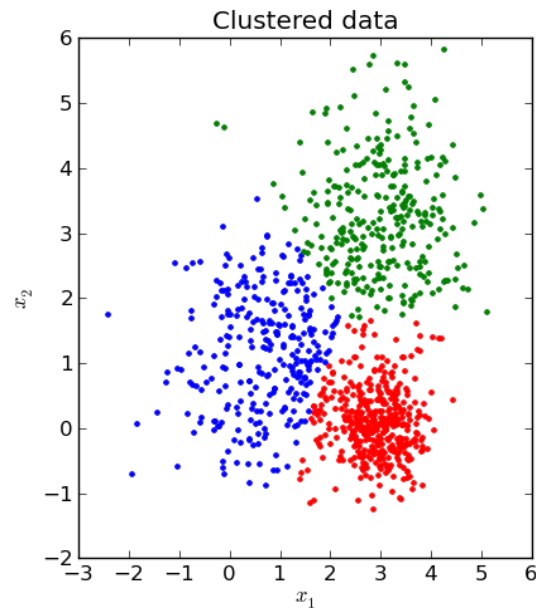
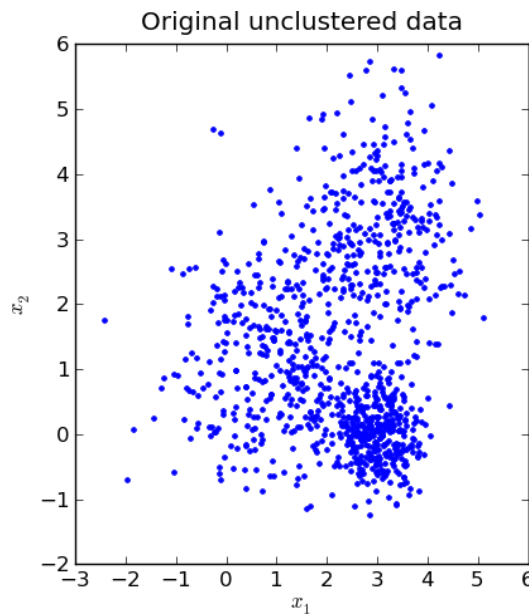
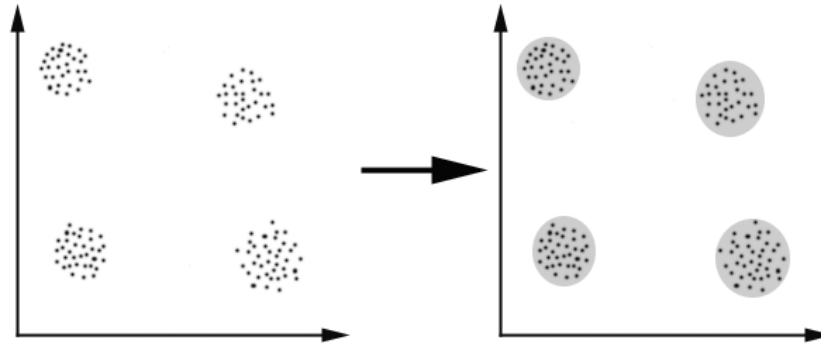
- ❖ **Overlapping groups**

- ❖ In some grouping methods, a same observation may be a member of multiple groups; in some methods, observations only exclusive to one group

Clustering

- ❖ Given a set of data points, group them into clusters so that:
 - ❖ Points within each cluster are similar to each other
 - ❖ Points from different clusters are dissimilar
- ❖ Similarity is defined using a distance measure i.e. Euclidean, Cosine, Jaccard, Edit distance, ...
- ❖ A good clustering algorithm will produce:
 - ❖ High intra-class similarity
 - ❖ Low inter-class similarity

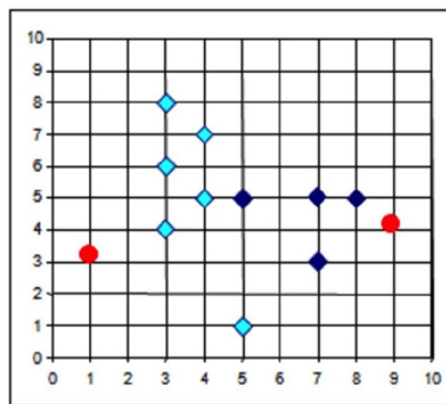
Clustering



K-means Clustering

1. Randomly pick **k** number of points for each cluster
⇒ known as **centroids**.
2. Each data point is **assigned** to the cluster with the closest centroids.
3. **Compute** the centroid of each cluster based on existing cluster members, this results in new centroids.
4. As we have new centroids, **repeat step 2 and 3**. Repeat this process **until convergence occurs** i.e. centroids do not change.

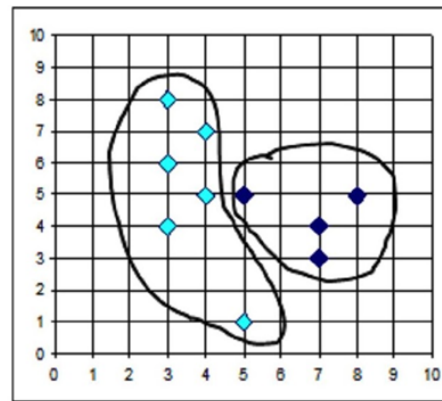
K-Means Clustering



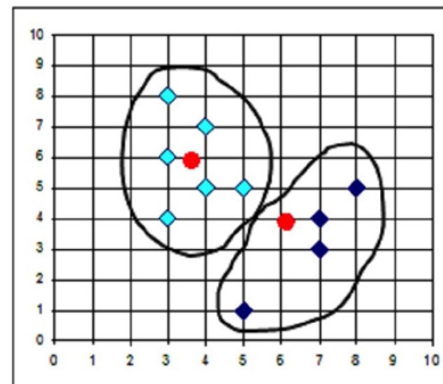
$K=2$

Arbitrarily choose K object as initial cluster center

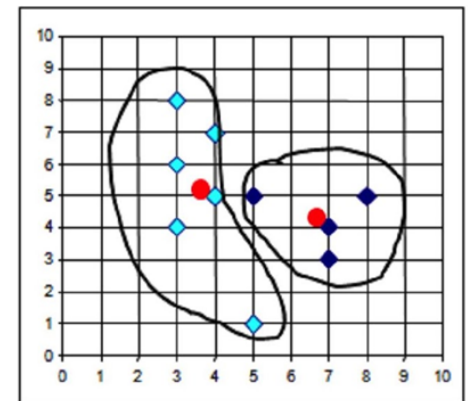
Assign each objects to most similar center



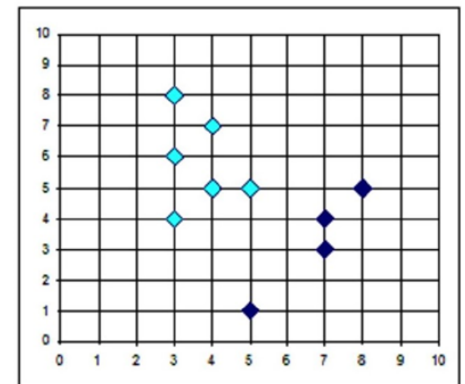
reassign



Update the cluster means



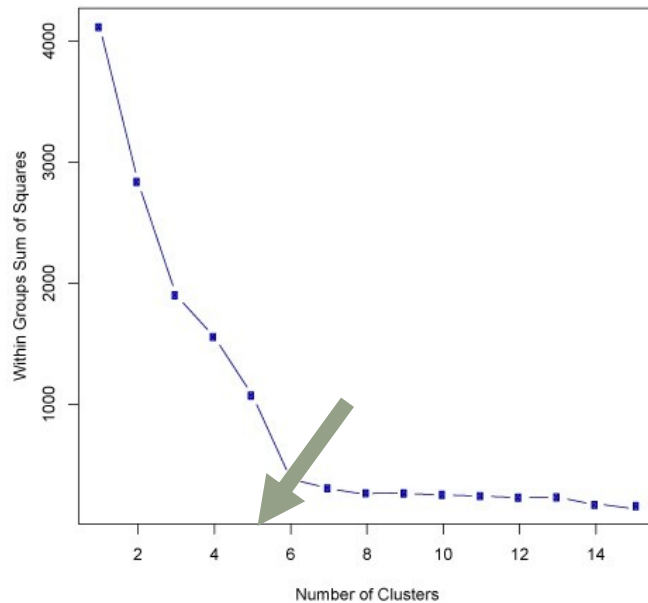
reassign



Update the cluster means

How to decide value of k ?

- ❖ “Elbow” of the k vs. Sum of Square Error (SSE) curve



$$SSE = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

- ❖ Empirical testing and knowledge of nature of data

K-means Clustering: Limitations

- ❖ **Predefined number of clusters:** You must define the number of groups before creating clusters
- ❖ **Distorted by outliers:** Too many outliers can result in non-optimal grouping. These outliers influence the mean value (when calculating centroids)
- ❖ **No hierarchical organization**

Clustering: Applications

Marketing: help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

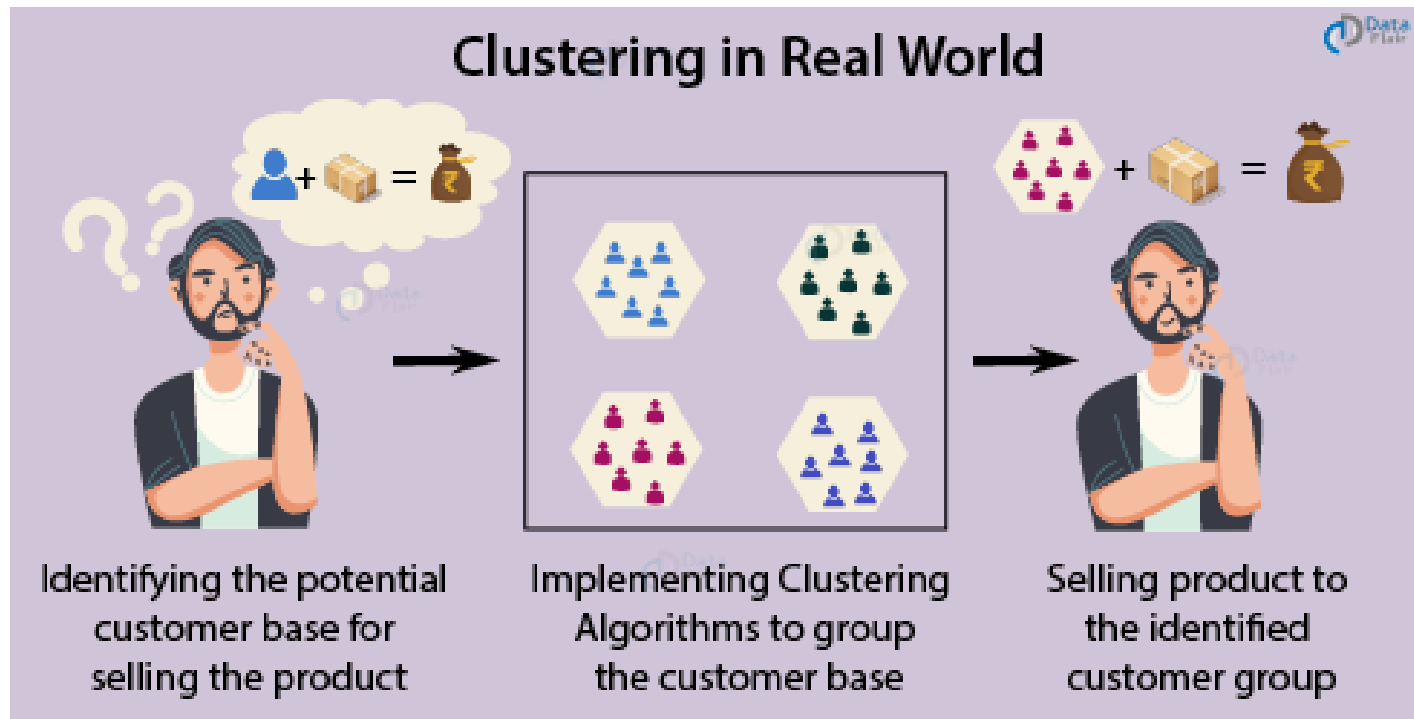
Insurance: Identifying groups of motor insurance policy holders with a high average claim cost

City-planning: identifying groups of houses according to their house type, value, and geographical location

DNA sequences: cluster DNA information based on edit distance

Clustering: Applications

Identifying segments of the customer base to better sell products



Association Rule Mining

- ❖ Shopping centers use **association rules** to place the items next to each other so that users buy more items.
 - **Wal-Mart** studied their data and found that on Friday afternoon young American males who buy diapers also tend to buy beer. So Wal-Mart placed beer next to diapers and the beer-sales went up.
 - **Amazon** uses association mining to recommend you the items based on the current item you are browsing/buying.
 - The **Google** auto-complete, where after you type in a word it searches frequently associated words that user type after that particular word.

Association Rule Mining

- ❖ **Apriori algorithm** is an association rule mining algorithm used in data mining. It is used to find the frequent itemset among the given number of transactions.
- ❖ Association rules find all sets of items (itemset) that have **support greater than the minimum support**.
- ❖ Then, the large itemsets are used to generate the desired rules that have confidence greater than the minimum confidence

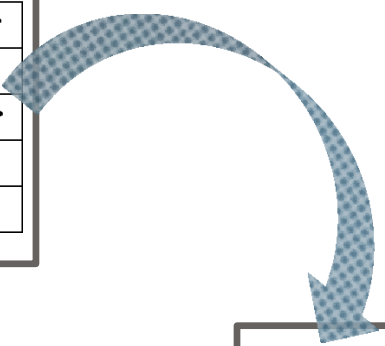
Association Rule Mining

- ❖ An associative rule has 2 parts:
 - ❖ Antecedent (if) and
 - ❖ Consequent (then)
- ❖ An antecedent is something that is found in data
- ❖ A consequent is an item that is found in combination with the antecedent.

“If a customer buys bread, he is 70% likely to buy milk”

Association Rule Mining

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk



$X \Rightarrow Y$	s	α
Bread \Rightarrow PeanutButter	60%	75%
PeanutButter \Rightarrow Bread	60%	100%
Beer \Rightarrow Bread	20%	50%
PeanutButter \Rightarrow Jelly	20%	33.3%
Jelly \Rightarrow PeanutButter	20%	100%
Jelly \Rightarrow Milk	0%	0%

Association Rule Mining

Rule: $X \Rightarrow Y$

$Support = \frac{freq(X,Y)}{N}$

$Confidence = \frac{freq(X,Y)}{freq(X)}$

$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- ❖ **Support:** How popularity an itemset, as measured by the proportion of transactions in which an itemset appears.
 - ❖ E.g. Cereal implies milk; milk implies cereal. Cereal and milk might appear together in 40% of the transactions i.e. support = 40%
- ❖ **Confidence:** How likely item Y is purchased when item X is purchased, expressed as $\{X \rightarrow Y\}$. This is measured by the proportion of transactions with item X, in which item Y also appears.
 - ❖ E.g. Cereal might appear in 50 transactions; 40 of the 50 also include milk. Cereal implies milk with 80% confidence.

Association Rule Mining

- ❖ **Lift:** This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent.
 - ❖ $\text{Lift} > 1$: Item Y is likely to be bought if item X is bought,
 - ❖ $\text{Lift} < 1$: Item Y is unlikely to be bought if item X is bought.
- ❖ E.g. Convenience store customers who buy orange juice also buy milk with a 75% confidence. The combination of milk and orange juice has a support of 30%. Convenience store customers in general buy milk 90% of the time. Assuming that 40% of the customers buy orange juice,

$$\text{Lift} = 0.3 / (0.4 * 0.9) = \mathbf{0.83 \text{ (no real cross-selling opportunity)}}$$

ARM: Drawbacks

1. **Computational intensive** – if itemset size is large
⇒ requires repeated scanning during itemset generation
2. **Can mine misleading patterns** – if there are one-time (rare) items, could be down to chance ⇒ does not have a historical context of items

ARM: Applications

- ❖ **Market Basket Analysis**: Given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.
- ❖ **Telecommunication**: Each customer is a transaction containing the set of phone calls
- ❖ **Credit Cards/ Banking Services**: Each card/account is a transaction containing the set of customer's payments
- ❖ **Medical Treatments**: Each patient is represented as a transaction containing the ordered set of diseases
- ❖ **Basketball-Game Analysis**: Each game is represented as a transaction containing the ordered set of ball passes

Reading Material

- Andrea Trevino, “[Introduction to K-means Clustering](#)”, Datascience.com, 2016.
- George Seif, “The 5 Clustering Algorithms Data Scientists Need To Know”, Towards Data Science, 2018.
- Chris Moffitt, “Introduction to Market Basket Analysis in Python”, Practical Business Python, 2017.