Chapter Three

Unsupervised Learning

Prepared by: Tsehay A. (B.Sc., and M.Sc., in Computer Science)

Department of Computer Science

2015 E.C

Contents

- Introduction
- Clustering approaches
- K-Means clustering
- *K nearest neighbors*
- Hierarchical clustering
- Evaluation of clustering models
- Association rule learning
- FP-G algorithm
- *Reinforcement learning*

Unsupervised learning

- Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets without labeled responses.
- In unsupervised learning algorithms, classification or categorization is not included in the observations.
- There are no output values and so there is no estimation of functions. Since the examples given to the learner are unlabeled, the accuracy of the structure that is output by the algorithm cannot be evaluated.
- The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns.

Cont.

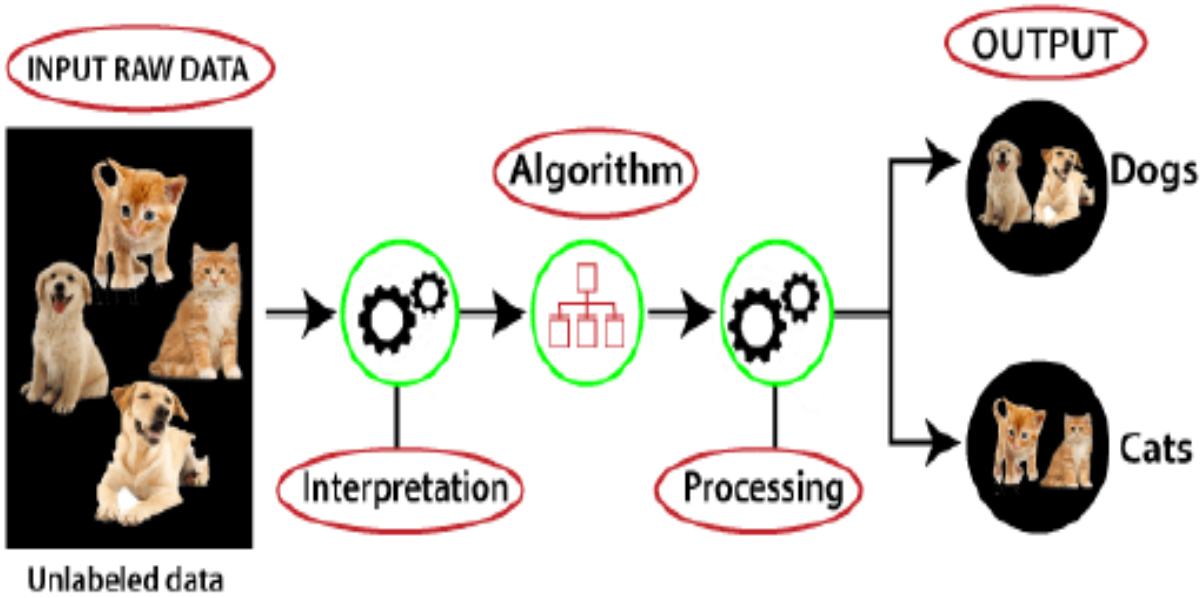


Fig. 3.1. Unsupervised learning

Cont.

- Consider the following data regarding patients entering a clinic.
- The data consists of the gender and age of the patients and each patient is labeled as “healthy” or “sick”.
- Based on this data, can we infer anything regarding the patients entering the clinic?

Table 3.1. Dataset for unsupervised learning

Gender	Age
Male	48
Male	67
Female	53
Male	49
Female	34

Clustering approaches

- **Clustering**: Clustering is a method of grouping objects into clusters.
- The objects with the most similarities remain in a group and have less or no similarities with the objects of another group.
- Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

Clustering approaches continued’

- *Hierarchical* versus *Partitioned*

- The most commonly discussed distinction among different types of clustering is whether the set of clusters is nested or unnested

- Partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

- If we permit clusters to have subclusters, then we obtain a hierarchical clustering, which is a set of nested clusters that are organized as a tree.

Why clustering?

- Labeling a large set of sample patterns can be costly.
- The contents of the database may not be known.
- Clustering can be used for finding features that will later be useful for categorization.
- It may help to gain insight into the nature of the data.
- It may lead to the discovery of distinct subclasses or similarities among patterns.

K-Means clustering

- The most commonly used clustering algorithm is *k-means*.
- It allows grouping the data according to the existing similarities among them in k clusters, given as input to the algorithm.
- Imagine that there are five objects (say 5 people) each described by two features (height and weight).
- *Step 1*: determine the number of *K* clusters.
- Let us group them into *k=2* clusters.

Cont.

Table 3.2. Dataset for K-means clustering

Person	Height (cm)	Weight (Kg)
Person 1	167	55
Person 2	120	32
Person 3	113	33
Person 4	175	76
Person 5	108	25

- *Step 2*: Initialize the value of the centroids of clusters. For instance, let's choose Person 2 and Person 3 as the two centroids *c1* and *c2* so that *c1=(120,32)* and *c2=(113,33)*.
- *Step 3*: Compute the Euclidean distance between each of the two centroids and each point in the data.

Cont.

Table 3.3. The distance of data points to centroids

Person	Distance of object from c1	Distance of object from c2
Person 1	52.3	58.3
Person 2	0	7.1
Person 3	7.1	0
Person 4	70.4	75.4
Person 5	13.9	9.4

- At this point, we will assign each object to the cluster it is closer to (taking the minimum between the two computed distances for each object). We can then arrange the points as follows:

Person 1→Cluster 1

Person 2→Cluster 1

Person 3→Cluster 2

Person 4→Cluster 1

Person 5→Cluster 2

Cont.

- *Step 4*, redefine the centroids by calculating the mean of the members of each of the two clusters.
- So *c'1* = ((167+120+175)/3, (55+32+76)/3) = (*154, 54.3*) and *c'2* = ((113+108)/2, (33+25)/2) = (*110.5, 29*)
- Then, we calculate the distances again and re-assign the points to the new centroids.
- Repeat this process until the centroids don't move anymore (or the difference between them is under a certain small threshold).

Cont.

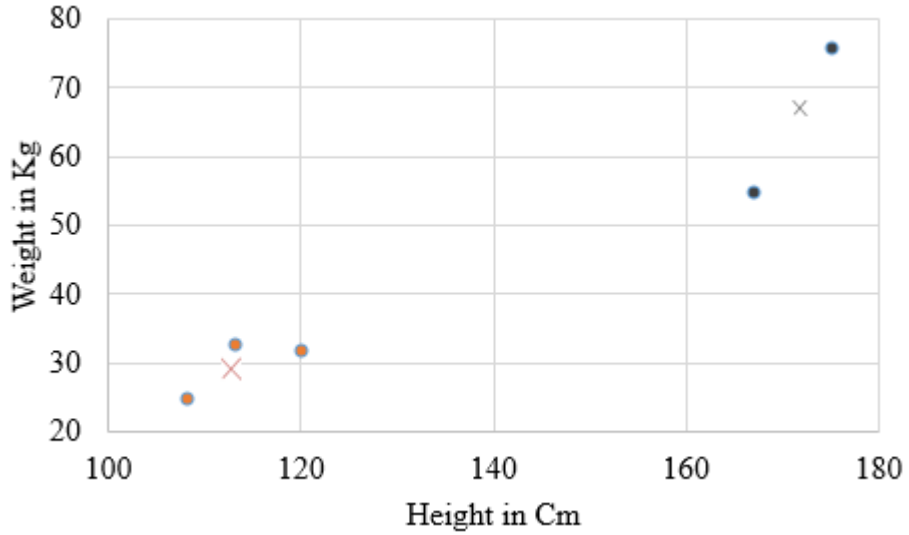


Fig. 3.2. Cluster of the person

- The two different clusters in red and black.
- The crosses indicate the position of the respective centroids.

Exercise

- Customer segmentation: Goal: To make 3 marketing strategies
- Features age of customer in years
- Engagement with the page (in days per week)
- Identify the three customer groups or clusters for implementing the strategies for the customer information given in Figure 3.3.

Fig. 3.3. Customer dataset

Evaluation of clustering models

- Clustering is the most common form of unsupervised learning.
- In clustering, a set of features for observation are used to create clusters with similar observations.
- Clustering model is evaluated based on some similarity or dissimilarity measure such as the *distance between cluster points*.
- If the clustering algorithm separates dissimilar observations apart and similar observations together, then it has performed well.

Cont.

- The popular evaluation metrics for clustering algorithms is *Silhouette* coefficient.

Silhouette Coefficient, $s = \frac{b - a}{max(a, b)}$

a: The mean distance between a sample and all other points in the same cluster.

b: The mean distance between a sample and all other points in the next nearest cluster.

- The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.
- The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. The score is higher when clusters are dense and well separated.

Association rule learning

- The association rule is a learning technique that helps identify the dependencies between two data items.
- Association rule learning can be viewed as a two-step process

1. Find all frequent item sets
    a) FP-Growth method
    b) Apriori method

2. Generate strong association rules from frequent item sets: by definition, these rules must satisfy minimum support and confidence.

Definition: Frequent Itemset

- **Itemset:** A collection of one or more items. Example: {Milk, Bread, Diaper}
- k-itemset: An itemset that contains k items.
- **Support count (σ):** Frequency of occurrence of an Itemset: E.g. σ({Milk, Bread,Diaper}) = 2
- **Support:** Fraction of transactions that contain an itemset. E.g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset:** An itemset whose support is greater than or equal to a *minsup* threshold.

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Frequent Pattern Growth (Fp-G) Method

- Fp-G is used to find frequent item set in databases with different transaction.

- The steps involved in association rule learning with FP-G are as follows:

1. Compute the support count of each item in a transaction.
2. Arrange the items in the transaction in descending order of support count.
3. Construct a frequent pattern tree, Fp-tree.
4. Construct conditional databases.
5. Generate frequent item sets.
6. Stop

Example

- Generate frequent item sets using FP-G for transactions given in Table 3.4. Consider support =2.

Table 3.4. List of hypothetical transactions

Transactions	List of Items
T 1	I1, I2, I5
T 2	I2, I4
T 3	I2, I3
T 4	I1, I2, I4
T 5	I1, I3
T 6	I2, I3
T7	I1, I3
T 8	I1, I2, I3, I5
T 9	I1, I2, I3

Example

1. Compute the support count of each item set.

Table 3.5. Support count of items

Item set	Support count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

2. Arrange the items in descending order of support

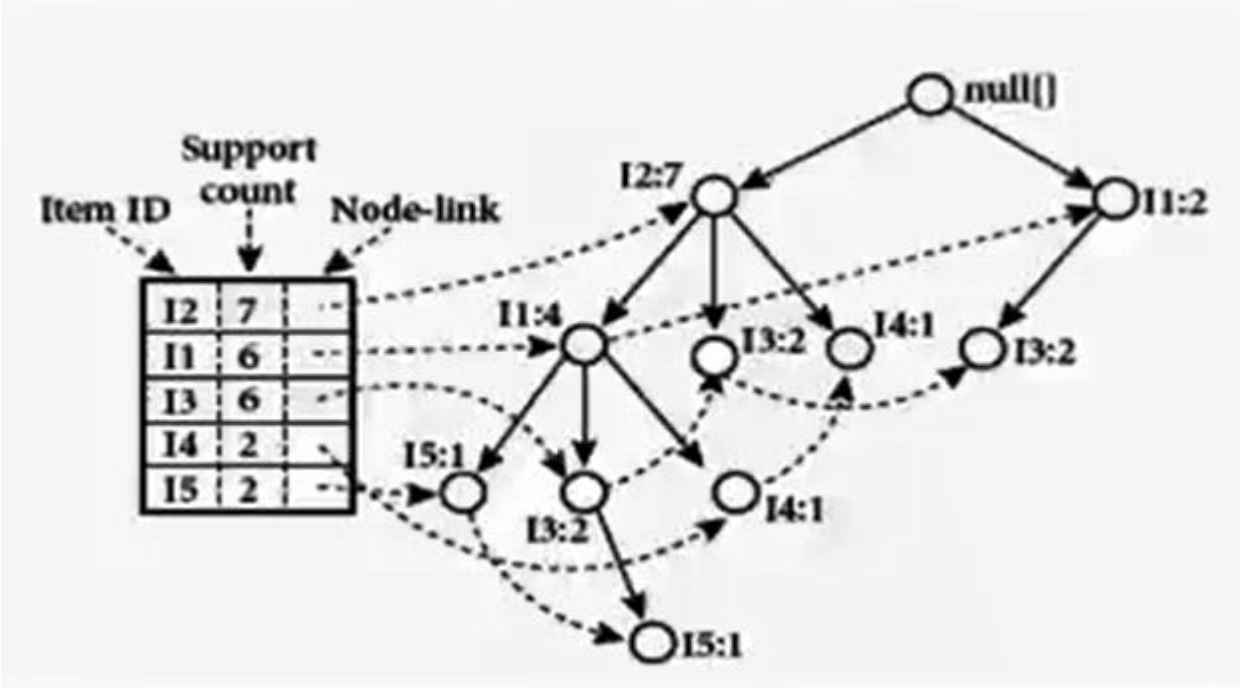
Table 3.6. Item set in descending order of support

Item set	Support count
I2	7
I1	6
I3	6
I4	2
I5	2

Cont.

3. Construct the FP-tree. Considering the root node as null.

- Fp-tree is a data structure that stores quantitative information about frequent patterns in a database.



Cont.

- Generate frequent pattern: Join the conditional FP-tree with the item.
- Conditional FP-tree: conditional pattern base that satisfies the support.
- Conditional pattern Base: the path between, root node to an item.

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}

Chapter review questions

a) Define unsupervised learning.

b) Compare unsupervised learning techniques with supervised learning.

c) What is clustering.

d) Compare and contrast the Apriori algorithm with FP-G.

e) Define K-means clustering.

f) What are the challenges of K-means clustering?

g) State types of clustering.

h) Define association rule mining.

i) Why is it difficult to evaluate clustering methods in comparison to supervised methods.

j) Mention a class of problems that can be solved by K-means clustering.