UNSUPERVISED LEARNING

Hannah Van Santvliet

November 7, 2020

Contents

1 What is Unsupervised Learning?

As explained by Joschka Braun in his presentation on supervised learning, machine learning can face problems which had been unsolvable or too costly to apply e. g. automatic cancer prediagnostics. The amount of data hidden in medical images is very time consuming to analyse accurately.
Due to the fact, that the majority of data is unlabeled - because normally a big amount of data is expensive to label - unsupervised learning wins more and more importance.

Definition 1.1. Unsupervised learning is a kind of machine learning where a model search for patterns in a dataset with no labels and with minimal human supervision.[Woo]

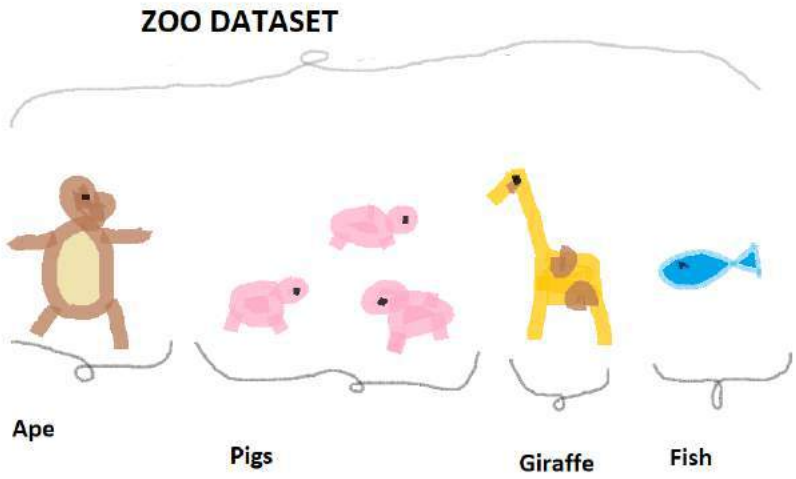


Figure 1: Finding patterns in data can be done quite differently. Sorting animals also could have been done regarding to the animals' requirements: $\{fish\}, \{ape, pigs, giraffe\}$, The fish is the only animal which is not a mammal and needs an aquarium.

To recap the last presentation and additionally evaluate pros and cons, unsupervised learning is going to be contrasted by supervised and semi-supervised learning.

2 Supervised, Unsupervised and Semi-supervised Learning

	Supervised Learning	Unsupervised Learning
training dataset contains:	input variable x and output variable y	input variable x
training dataset contains:	labeled dataset	unlabeled data set
goal:	learns the output y from the input data	learns inherent structure from the input data

	Semi-supervised Learning
training dataset contains:	input variable x and output variable y
training dataset contains:	typically a mixture of a small amount of labeled and a large amount of an unlabeled data set
goal:	learns inherent structure and the output y from the input data

3 Clustering

What have been done with the data consisting of animals had been clustering. Corresponding to some categories like species or living habits subsets of $\{ape, pig1, pig2, pig3, giraffe, fish\}$ had been found. This indicates that there are several ways how to cluster data. It depends what is expected of sorting the data. Is it a register to search for animals? Or a list within a biology class summing up different living habits?
Additionally, it should be mentioned that current application face very big data. An amount of data where counting is not possible anymore. Imagine putting all zoos together and then start counting every animal including ants.

Definition 3.1. Clustering can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. [Dab]

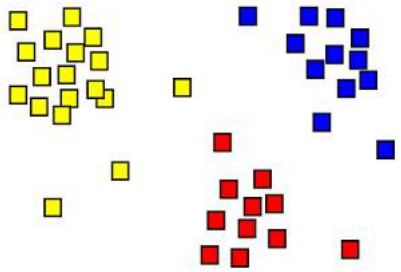


Figure 2: Data points were clustered and marked by colour

There are several possibilities to group those points and so there are many algorithms for clustering e.g. k-means, Hierarchical Clustering, Density-Based Clustering, Principal Component Analysis, Gaussian Mixture Models...

3.1 k-means Algorithm

The algorithm k-means searches for k distinct sets $\{S_1, ..., S_k\} = S$ in a dataset S while n observations $x_1, ..., x_n, k \leq n, k, n \in \mathbb{N}$ and k initial means $m_1, ..., m_k$ are given.
The standard algorithm called Naive k-means works iteratively ($t \to \infty$) and is based on a Assignment and Update step.

1. Assign each observation x_i to the cluster with the nearest mean:

$$S_i^{(t)} = \{x_p \mid \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\}$$

2. Recalculate means m_i for observations assigned to each cluster:

$$m_i^{(t+1)} = \frac{1}{\|S_i^{(t)}\|} \sum_{x_j \in S_i^{(t)}} x_j$$

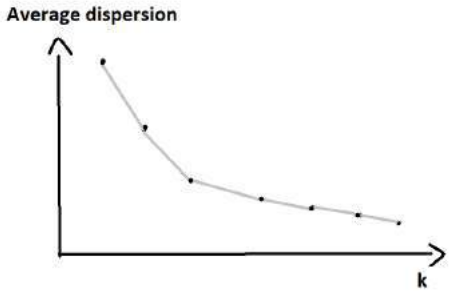
3.2 Example: Finding groups within a dataset based on medical information

Imagining your health insurance would collect data about your weight and height, but an unlucky incidence caused the disappearence of data on your gender. The k-means algorithm could help.
You could order the algorithm to find two (k = 2) groups in the data, assuming that generally, male and female differ in weight and height and put $m_1 = (71.6, 1.66), m_2 = (88.8, 1.80)$ [sev] using data for an average woman/ man in Germany.
After you have given your dataset $x_1, ..., x_n$ where each element has the form $x_i =$ (weight kg,height m), your expected outcomes would be sets S_1, S_2 in which the data is sorted depending on gender.

Problems might occur with an unlucky set choice of initial means as well as an unintuitively k choice, which can lead to very different results. A possibility to avoid extreme or unusual clusters could be running the algorithm for several k and initial means.This can be illustrated by elbow method.

3.3 Elbow Method

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. The average dispersion describes a approximate value between the distance to the centroid of each cluster and the suitable cluster points.



4 Association Rule

With the possibility to go shopping online it became common that personalised advertisement is proposed. Personalised advertisement means that advertisement is different depending on whom is consuming the advertisement.
But how is it possible for e.g. Amazon to propose books you never have searched for? Some examples may seem obvious like proposing an other cookbook if the last search was related to cooking, but how is it possible to formalise this relation, this association rule?

Definition 4.1. Association rule learning is a rule-based machine learning method for discovering relations between variables in large databases. It is intended to identify strong rules found in databases using some measures of interestingness.[PS92]

Before starting to solve the riddle of book taste during online shopping it is necessary to introduce some concepts.

4.1 Support and Confidence

The support is an indicator how often an itemset X appears
in the dataset T:

$$support(X) = \frac{|\{t \in T, X \subseteq t\}|}{|T|}$$

Example:
Let be T := $\{ape, pig$ and ape, pig and $fish, giraffe, fish\}$
Let be $X := \{pig\}$

$$\begin{aligned} support(\{pig\}) &= \frac{|t \in T, \{pig\} \subseteq t|}{|T|} \\ &= \frac{2}{5} \end{aligned}$$

The confidence value of a rule $X \Rightarrow Y$ is the proportion of the transactions that contains X which also contains Y:

$$\begin{aligned} conf(X \Rightarrow Y) &= \frac{support(X \cup Y)}{support(X)} \\ &= \frac{\frac{|\{t \in T, X \cup Y \subseteq t\}|}{|T|}}{\frac{|\{t \in T, X \subseteq t\}|}{|T|}} \\ &= \frac{|\{t \in T, X \cup Y \subseteq t\}|}{|\{t \in T, X \subseteq t\}|} \end{aligned}$$

Example:
Let be T := $\{ape, pig$ and ape, pig and $fish, giraffe, fish\}$
Let be $X := \{pig\}$
Let be $Y := \{ape\}$

$$\begin{aligned} conf(\{pig\} \Rightarrow \{ape\}) &= \frac{|\{t \in T, \{pig, ape\} \subseteq t\}|}{|\{t \in T, \{pig\} \subseteq t\}|} \\ &= \frac{1}{2} \end{aligned}$$

4.2 Apriori Algorithm

Let be T a given transaction dataset and minSupport the support threshold. Let be C_k the candidate set of level k. That means that one element in C_k is a set with k elements. Let be L_k a frequent data set.

Algorithm 1: Pseudo code

```
Data: transaction dataset
Result: frequent sets
initialization:
C_1 = { singletons of T }
L_1 = {t ∈ T : support(t) ≥ minSupport, t ∈ C_1}
for (k = 1; L_k ≠ ∅; k++) do
    candidates generated from L_k:
    C_{k+1} = {t ⊆ r ∈ T : s ∈ L_k ∧ s ⊆ t, |t| = k + 1}
    foreach x ∈ C_{k+1} do
        Scan the dataset to see if each itemset is frequent:
        L_{k+1} = candidates in C_{k+1} with minSupport
    end
end
return all elements of L_k
```

4.3 Example: dataset consisting of books that have been bought

Let be $T = \{I_1, I_2, I_3, I_4\}$ and 50 % the minimum support and 50 % minimum confidence.

Transaction	item set
I_1	A, B, C
I_2	A, C
I_3	A, D
I_4	B, E, F

C_1	item	support
	$\{A\}$	3/4
	$\{B\}$	2/4
	$\{C\}$	2/4
	$\{D\}$	1/4
	$\{E\}$	1/4
	$\{F\}$	1/4
L_1	$\{A\}$	3/4
	$\{B\}$	2/4
	$\{C\}$	2/4

C_2	item	support
	$\{A, B\}$	1/4
	$\{B, C\}$	1/4
	$\{A, C\}$	2/4
L_2	$\{A, C\}$	2/4

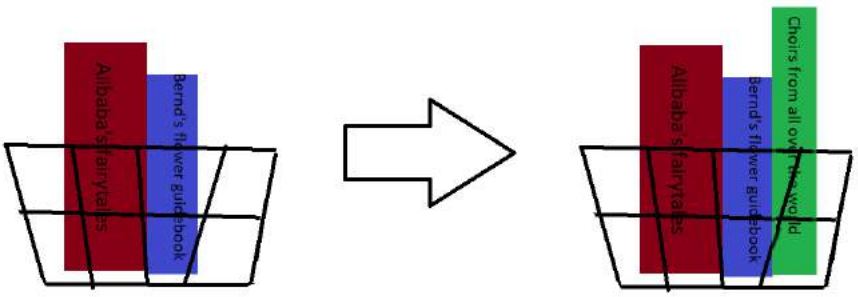
$\Rightarrow \{A, C\}$ is a frequent itemset.

Subsequently, two association rules $A \to C$ and $C \to A$ are possible.

$conf(A \Rightarrow C) = \frac{|\{t \in T, A \cup C \subseteq t\}|}{|\{t \in T, C \subseteq t\}|}$
$= \frac{2}{2}$
$\Rightarrow$ 100% confidence
$conf(C \Rightarrow A) = \frac{|\{t \in T, A \cup C \subseteq t\}|}{|\{t \in T, A \subseteq t\}|}$
$= \frac{2}{3}$
$\Rightarrow$ 66.67% confidence

Therefore, both rules are aboth the minimal confidence of 50%.

If A and B are standing for books, it is likely (the likeliness is described by the confidence), that if a bunch of books containing book A is bought that book B is bought as well.



The tables show that this algorithm might be exact but has to save plenty of data before finding the frequent datasets. Additionally, the iteration searching for appropriate subsets can last very long if the support is set low.
A common improvement of the apriori algorithm is the FP-growth algorithm which has no need of a candidate generation and saves therefore time and complexity.

References

[Dab] Dabbura, Imad: *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.* https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. – (zuletzt besucht: 30. 10. 2020)

[Gar] Garg, Anisha: *Complete guide to Association Rules (1/2).* https://towardsdatascience.com/association-rules-2-aa9a77241654. – (zuletzt besucht: 30. 10. 2020)

[PS92] Piatetsky-Shapiro, Gregory: *Discovery, analysis, and presentation of strong rules.* Cambridge : AAAI/MIT Press, 1992

[Rom] Roman, Vistor: *Unsupervised Machine Learning: Clustering Analysis.* https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e. – (zuletzt besucht: 30. 10. 2020)

[sev] several: *Complete guide to Association Rules (1/2).* https://www.laenderdaten.info/durchschnittliche-koerpergroessen.php. – (zuletzt besucht: 6. 11. 2020)

[Wika] Wikipedia1: *Unsupervised Learning.* https://en.wikipedia.org/wiki/Unsupervised_learning. – –(*zuletztbesucht* : 18. 6. 2020)

[Wikb] Wikipedia3: *k-means Clustering.* https://en.wikipedia.org/wiki/K-means_clusterin. – –(*zuletztbesucht* : 30. 10. 2020)

[Wikc] Wikipedia4: *Cluster Analysis.* https://en.wikipedia.org/wiki/Cluster_analysis. – –(*zuletztbesucht* : 30. 10. 2020)

[Woo] Wood, Thomas: *Unsupervised Learning.* https://deepai.org/machine-learning-glossary-and-terms/unsupervised-learning. – (zuletzt besucht: 30. 10. 2020)