# Was ist Clustering?

Clustering ist die Methode im maschinellen Lernen und genauer gesagt “im überwachten Lernen”, Detenpunkte in Gruppen (Clusters) ohne vorhandenes Wissen zu ordnen. Um das zu schaffen, wird die Ähnlichkeit zwischen die Daten berechnet.

Ziel von Clustering:

1. Ähnliche Datenpunkte (Objekte) sollen im selben Cluster sein.
2. Datenpunkte im verschiedenen Cluster soll möglichst unähnlich sein.

Also Clustering führt zu der Verringerung von der Komplexität.

Die gefundenen Ähnlichkeitsgruppen können graphentheoretisch, hierarchisch, partitionierend oder optimierend sein. Inzwischen gibt es im Clustering eine große Vielfalt an Methoden, die eingesetzt werden können. Das hängt aber vom Anwendungsfall und der Größe der Datenmenge ab.

Was man aber dabei beachten soll, dass die Algorithmen verschiedene Ergebnisse liefern können. Was öfter der Fall ist.

K-Means als Beispiel für partitionierende Verfahren:  
Im partitionierenden Verfahren wird ein Clustering mit k Cluster mit minimalen Kosten gesucht. Und bei K-Means wie der Name bereits sagt, sucht das Algorithmus für jeden Cluster einen zentralen Punkt, bei dem die Varianz zu allen umliegenden Punkten möglichst gering ist.

Das Ganze passiert in einem iterativen Verfahren wie folgendes:

1. Initialisierung: Zufällige Auswahl von K Zentren
2. Zuweisung aller Datenpunkte zum nächstliegenden Zentrum, gemessen an einer Distanzmetrik. Distanzmetrik berechnet hier die Ähnlichkeit.
3. Verschieben der Zentren in den Mittelpunkt aller zugeteilten Datenpunkte
4. Gehe zu 2), außer ein Abbruchkriterium ist erreicht. Der Abbruchkriterium ist, wenn die Zentren sich nicht mehr ändern.

Ensemble Clustering  
Im Zusammenhang mit maschinellem Lernen ist Ensemble im Allgemein definiert als ein maschinelles Lernsystem, das mit einer Reihe von parallel arbeitenden einzelnen Modellen konstruiert wird. Und die Ergebnisse von diesen Modellen werden mit einer Entscheideungsfusionsstratigie kombiniert, um einzige Lösung vom gegebenen Problem zu geben. Die Ensemble Methode war zuerst nur im Bereich des überwachten Lernens benutzt und betrachtet, aber aufgrund seiner erfolgreichen Applikationen in Klassifizierungsaufgaben wurde es versucht, die Ensemble Methode auf unüberwachtes Lernen anzuwenden. Insbesondere auf Clustering Probleme. Die Gründe dafür sind erstens, dass es kein vorhandenes Wissen über die zugrundliegende Struktur. Zweitens es gibt keinen einzelnen Clustering-Algorithmus, der für verschiedene Probleme konsistent gute Leistungen erbringen kann.  
Ensemble Clustering zielt darauf ab, mehrere Clusteringsmodelle zu kombinieren, um ein besseres Ergebnis als die einzelnen Clustering-Algorithmen. Dies ist in vielen Kontexten sehr nützlich:

1. Qualität und Robustheit
2. Knowledge Reuse
3. Distributed Computing

Man kann jetzt Ensemble Cluster als Problem sehen und zwar wie folgendes:

**Gegeben:** Mehrere Clusterings aus verschiedenen Algorithmen von derselben Datenmenge.  
**Gesucht:** Ein kombiniertes Clustering, welches so viele Informationen wie möglich mit den gegebenen Clusterings teilt.

# Objektive Funktion für Ensemble Clustering

Notation: Sei $X=\set{x\_1, x\_2, \ldots, x\_n} eine Menge von Objekten (Datenpunkte). Eine Partitionierung dieser $n Objekten in $k Clusters kann als eine Menge von k Mengen von Objekten $\set{$C\_{$l$} \mid $l$ = 1, \ldots, $k} oder als ein label Verktor \lambda \in \mathbb{N}^{n} dargestellt werden. Ein Clusterer \Phi ist eine Funktion, welche einen label Vektor bei gegebene Objekten liefert.  
Eine Mengen von $r Clusterings (Labelings) \lambda^{(1, \dots, r)} ist zu einem Clustering anhand einer Consensus Funktion kombiniert.