1 Clusteranalyse

Nicht-signierte Daten sind Daten, denen es an einer eindeutigen Klassifizierung fehlt, die aber trotzdem in Cluster eingeteilt werden können. Dieses Einteilen in Clustern wird bei Produktempfehlungen oder Kundendaten genutzt um Vorschläge zu machen, was nun empfohlen werden soll. Algorithmen sind dabei der k-Means-Algorithmus und der k++-Means-Algorithmus. Dabei ist zwischen harten und weichen Clusteringalgorithmen zu unterscheiden. Harte Clusteralgorithmen trennen die Klassenzugehörigkeit strikt binär, während weiche Algorithmen Mehrklassifizierungen durch probabalistische Vorhersagen erlauben. Die Hyperparameter der Clusteringalgorithmen kann durch ein Silhouettendiagramm oder das Ellenbogenkriterium bestimmt werden. Solche Cluster können ebenfalls als hierarchische Bäume organisiert werden.

2 k-Means-Algorithmus

Beim k-Means-Algorithmus werden zuerst k-zufällige Zentroiden ausgewählt. Ein Zentroid ist dabei ein Datenpunkt, der als Zentrum für das Cluster dient. Danach gruppieren wir die Daten durch eine Abstandsfunktion zum jeweiligen Cluster des nächsten Zentroids. Der Abstand ist dabei der euklidische Abstand.

$$d(x,y)^2 = \sum_{j=1}^{m}(x_j - y_j)^2 = ||x-y||^2 \tag{1}$$

Das Ziel des k-Means-Algorithmus kann also als Optimierungsproblem formuliert werden. Für jedes Cluster soll also die Varianz, also der Abstand zum Clusterzentrum durch Auswahl der Punkte, minimiert werden.

$$d(x,y)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\omega^{(i,j)}(x^{(i)} - \mu^{(j)})^2 \tag{2}$$

Der Vorfaktor ω gibt binär zurück ob ein Element zu einer bestimmten Klasse gehört oder nicht. Somit wird also die Varianz, zum Beispiel mit der Optimierungsmethode der kleinsten Quadrate, gelöst.

3 $k++$-Means-Algorithmus

Dieser Algorithmus funktioniert ähnlich wie der vorherhige Algorithmus, trifft allerdings eine andere Vorauswahl der Zentroiden. Dabei wird als erstes Element ein Zentroid gewählt. Danach wird durch eine Wahrscheinlichkeitsfunktion P festgestellt,

$$P = \frac{d(\mu^p, M)^2}{\sum_{i=1}^{n}(x^{(i)}, M)^2} \tag{3}$$

welcher Datenpunkt den höchsten Abstand zu den anderen Zentroiden hat. Dieser wird dann der Menge hinzugefügt und diese Schritte werden fortgeführt bis man k-Zentroiden ausgewählt hat. Danach wird der k-Mean-Algorithmus ausgeführt.

4 Fuzzy-C-Means-Algorithmus

Zuerst werden allen Punkten eine zufällige Klassenzugehörigkeit zugewiesen. Dabei wird das binäre System der Zuordnung durch ein probabalistisches System ausgetauscht. Das heißt es wird jetzt eine Wahrscheinlichkeit angegeben, dass ein Punkt zu einem Cluster gehört.
Die zu optimierende Zielfunktion lautet dann

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k}\omega^{m(i,j)}(x^{(i)} - \mu^{(j)})^2 \tag{4}$$

Dabei ist ω^m ein reeller Gewichtungskoeffizient, der die Wahrscheinlichkeit angibt zu einem Clusters j zu gehören. m ist dabei Hyperparameter, der kleine Wahrscheinlichkeiten noch weiter minimiert. Umso größer dieser Koeffizient umso kleiner werden kleine Wahrscheinlichkeiten. Nun gilt für diesen Koeffizient:

$$\omega^{m(i,j)} = [\sum_{p=1}^{k}(\frac{||x^{(i)} - \mu^{(j)}||}{||x^{(i)} - \mu^{(p)}||})^{\frac{2}{m-1}}]^{-1} \tag{5}$$

Das Clusterzentrum μ berechnet sich hier allerdings aus den Mittelwerten der Punkte der Punkte eines Clusters.

$$\mu^{(j)} = \frac{\sum_{i=1}^{n} w^{m(i,j)}x^{(i)}}{\sum_{i=1}^{n} w^{m(i,j)}} \tag{6}$$

Diese Berechnungen sind für einen Computer sehr rechenaufwändig. Diese Schritte werden so lange durchgeführt bis sich die Gewichtungswerte nicht mehr ändert.

5 Cluster als hierarchischer Baum

Es gibt verschiedene Arten der Konstruktion, einerseits die Zusammenführung aus Einzelclustern und das Aufspalten aus einem großen Einzelcluster. Diese Entscheidungen werden in einem Dendrogramm repräsentiert. Diese Konstruktion eines Dendrogammes wird daurch erreicht, dass jeder Datenpunkt in einem eigenen Cluster sitzt. Diese werden dann in jeder Iteration mit den ähnlichsten Clustern zu einem neuen Cluster zusammengefügt. Die Ähnlichkeit wird wie oben durch die euklidische Distanz berechnet.

6 Bereiche hoher Dichte ermitteln.

Bei diesem Algorithmus werden verschiedene Objekte mit verschiedenen Labels klassifiziert. Dabei ist ein Objekt ein Kernobjekt, wenn es in einem Kreis, dessen Mitte das Objekt ist, mit dem Radius r mindestens k-Objekte gibt. Ein Randobjekt ist dabei die entgegengesetzte Bedingung eines Kernobjektes und der Anwesenheit in dem Radius eines anderen Kernobjektes. Dabei werden Cluster mit den Kernobjekten zusammengesetzt, die im Radius einens anderen Kernobjekts sind. Im Cluster sind dann alle Objekte, die Randobjekte dieser Kernobjekte. Objekte, die keine dieser Bedingungen erfüllen sind Rauschobjekte, die herausgefiltert werden können. Der Vorteil ist die Bildung eines Clusters, was nicht unbedingt sphärisch sein muss.

7 Bewertungen der Hyperparameter

Plottet man für den Hyperparameter k die endgültige minimierte Varianz, so lässt sich graphisch anhand eines Punktes, bei dem die Steigung sehr abnimmt den optimalen Hyperparameter zu bestimmen.
Eine weitere Methode ist die Berechnung des Koeffizienten $s^{(i)}$.

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\left(b^{(i)}, a^{(i)}\right)} \tag{7}$$

Dabei ist $a^{(i)}$ der Mittelwert zwischen den Distanzen des Objektes i und den anderen Objekten im Cluster, sowie $b^{(i)}$ als der Mittelwert der Distanz

des Objektes zum nächsten Cluster. Ist die Distanz zum nächsten Cluster groß und die Geschlossenheit im eigenen Cluster klein, so bähert sich der Koeffizient 1. Ist es allerdings andersrum, so nähert sich der Koeffizienz -1.