

4. Übung zur Vorlesung “Bioinformatische Methoden in der Genomforschung”

Sebastian Böcker, Martin Hoffmann

Ausgabe: 23.11.2017

Abgabe: 30.11.2017

Aufgabe 1 (5 Punkte)

1. Gegeben seien acht Punkte $p_1 = (2, 10)$, $p_2 = (2, 5)$, $p_3 = (8, 4)$, $p_4 = (5, 8)$, $p_5 = (7, 5)$, $p_6 = (6, 4)$, $p_7 = (1, 2)$, $p_8 = (4, 9)$ auf der Ebene. Wenden Sie den Lloyd Algorithmus an, um diese acht Punkte in drei Clustern C_1 , C_2 , C_3 zuzuordnen. Initialisieren Sie den Algorithmus mit $C_1 = \{p_1\}$, $C_2 = \{p_2, p_7\}$, $C_3 = \{p_3, p_4, p_5, p_6, p_8\}$. Benutzen Sie dabei die Manhattan-Distanz.
2. Ein *stationärer Zustand* ist ein Zustand der berechneten Clustern, aus dem der Lloyd Algorithmus keinen neuen Zustand erreicht. Konstruieren Sie ein möglichst einfaches Beispiel (euklidische Distanz), in dem es mehr als einen stationären Zustand gibt. Konstruieren Sie ein Beispiel mit n Punkte auf der Ebene, die in k Clustern zugeordnet werden soll und für die es mindestens 2^k stationäre Zustände gibt.

Aufgabe 2 (5 Punkte)

Gegeben sei ein zusammenhängender Graph $G = (V, E)$. Für jedes Paar $uv \in \binom{V}{2}$ seien $s(uv)$ die Kosten für das Hinzufügen der Kante uv , wenn $uv \notin E$, bzw. die Kosten für das Löschen der Kante uv wenn $uv \in E$.

Sei $u, v \in V$ mit $uv \notin E$:

$$s(uv) > \sum_{w \in N(u)} s(uw),$$

wobei $N(u)$ die Menge der benachbarten Knoten von u bezeichnet. Zeigen Sie, dass uv keine Kante der optimalen Lösung des CLUSTER EDITING Problems auf G ist.

Aufgabe 3 (5 Punkte)

Erklären Sie den generellen Ablauf des k-means clustering Verfahrens in eigenen Worten (Fließtext). Gehen Sie hierbei auch auf die Laufzeit und die Grundannahmen des Algorithmus ein.