4. Übung

Schriftliche Aufgaben

Vorname: Tom
Nachname: Tucek
Matrikelnummer: 1325775

1.5 k-Means VS. agglomeratives Clustering

1.5.1 Welches ``Linkage''-Kriterium ist das Beste für jeden der gewählten Datensätze und weshalb?

Für den Datensatz „Aggregation.txt“ (siehe Figure 2) ist das ‚centroid‘-Linkage-Kriterium geeignet, da es die tatsächliche Ground-Truth komplett wiederherstellen kann. Allerdings liefert es eine Warnung (Warning: Non-monotonic cluster tree -- the centroid linkage is probably not appropriate).

Das Clustering für „Aggregation.txt”, welches in der Angabe zu sehen ist, kann mit dem ‚average‘-Linkage-Kriterium erreicht werden, allerdings werden dabei einige Datenpunkte zwischen den Clustern auf der rechten Seite falsch zugeordnet.

Das ‚centroid‘-Linkage-Kriterium verwendet die euklidische Distanz zwischen den Zentroiden der Cluster.

Für den Datensatz „Jain.txt“ (siehe Figure 1) ist kein Linkage-Kriterium perfekt geeignet, da keines die Ground-Truth komplett wiederherstellen kann. Ich habe mich für das ‚weighted‘-Linkage-Kriterium entschieden, da es einen geringen Fehler aufweist und eines der Cluster komplett richtig kategorisiert (also alle tatsächlich zum unteren Cluster gehörigen Daten als solche klassifiziert).

Das ‚weighted‘-Linkage-Kriterium spaltet Cluster rekursiv in kleinere Cluster auf und berechnet danach die Distanz jener kleineren Cluster zu anderen Clustern. Die eigentliche Distanz wird dann aus dem Durchschnitt der Distanzen berechnet.

1.5.2 Welcher Clustering-Algorithmus schneidet bei jedem der gewählten Datensätze besser ab und warum?

Wie in den Grafiken (Figure 2 und Figure 1) sichtbar ist, schneidet bei beiden Datensätzen das agglomerative Clustering besser ab. Beim ersten Datensatz kann die Ground-Truth tatsächlich rekonstruiert werden und beim zweiten Datensatz ist der Fehler halbwegs gering. Das k-Means-Clustering hat in beiden Fällen einen größeren Fehler, da es mehr Daten falsch clustert.

Dies liegt unter anderem an den zufälligen Startpunkten des k-Means-Clustering. Wären diese anders gesetzt, könnten bessere, aber auch noch schlechtere Ergebnisse erreicht werden. Da die euklidische Distanz verwendet wird, werden Cluster kreisförmig gebildet, was das korrekte Clustering des zweiten Datensatzes („Jain.txt“) erschwert, da diese Cluster in nicht kreisförmigen Formen auftreten.

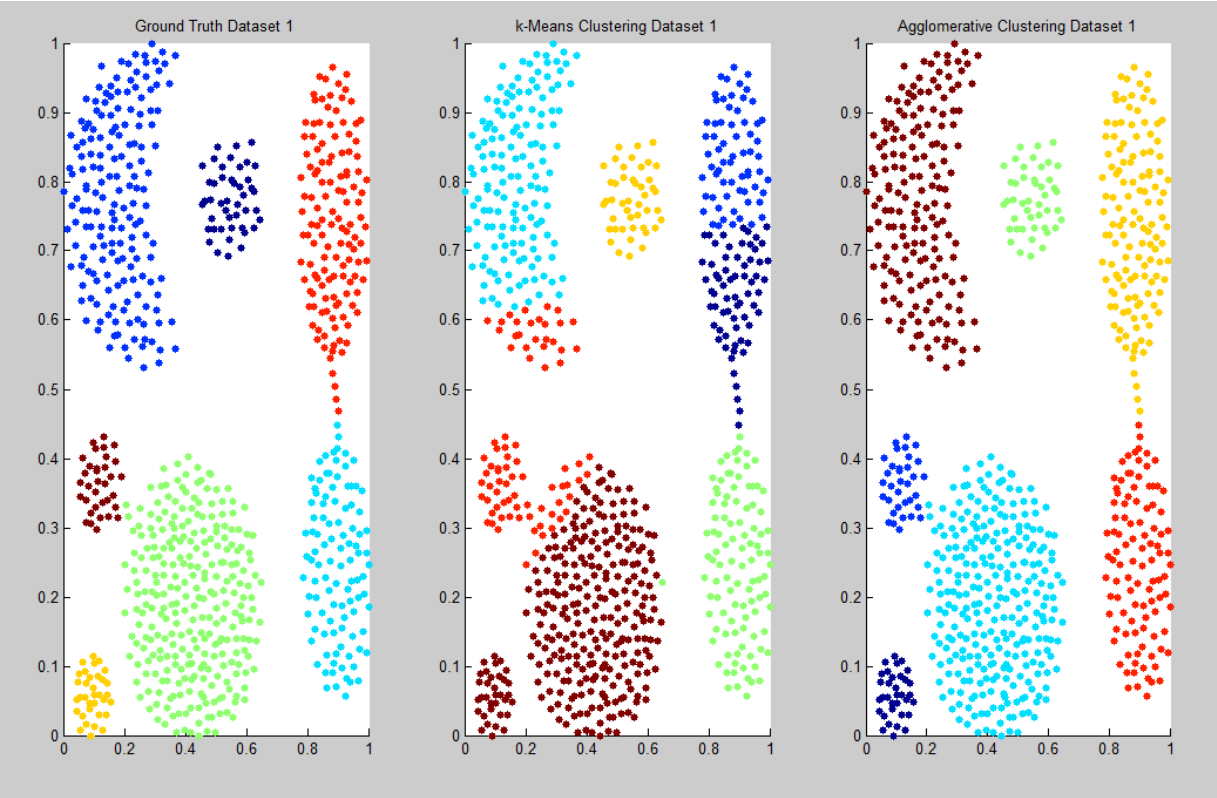


Figure 2: Datensatz 1 - Aggregation.txt

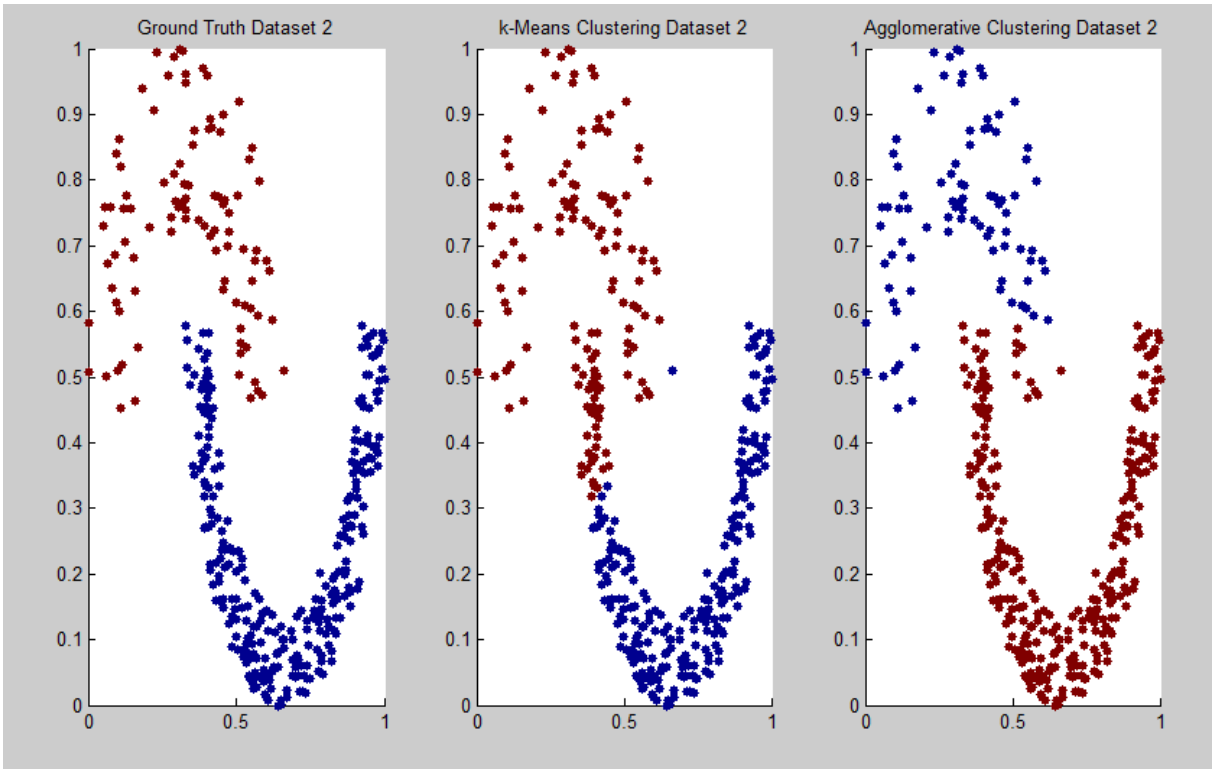
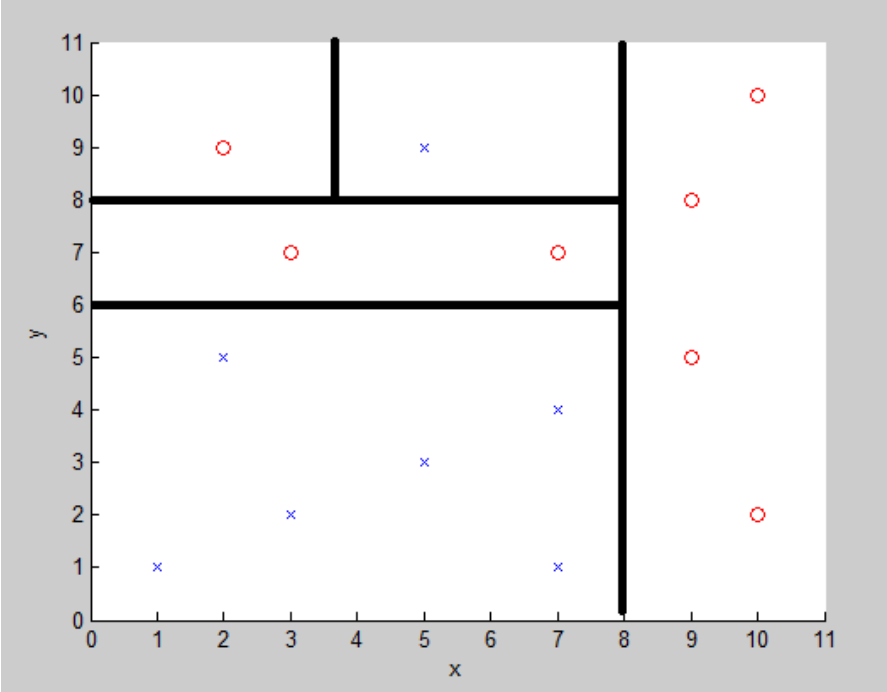


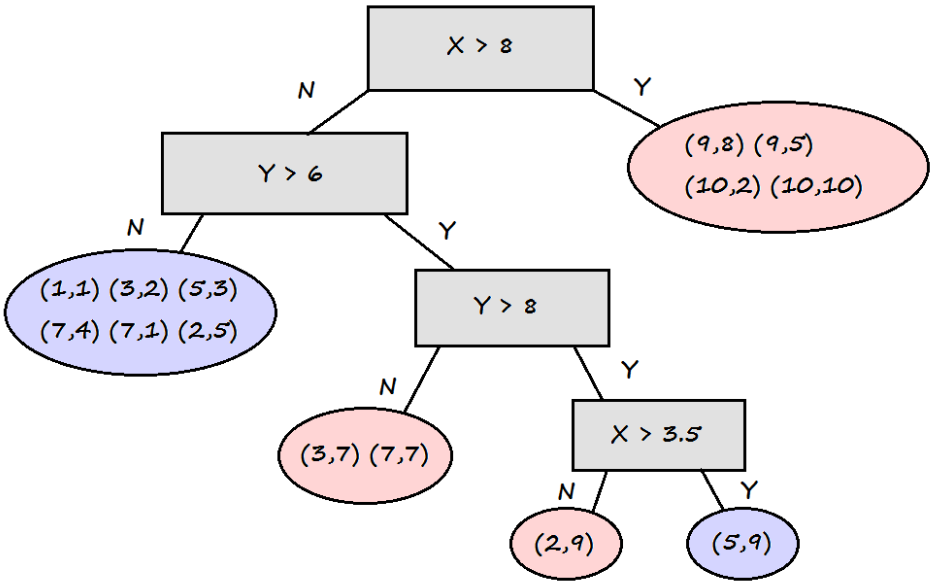
Figure 1: Datensatz 2 - Jain.txt

2.2 Entscheidungsbaum und -regionen zeichnen

Ausgabe des Programms:

Node: sepearation in x. Threshold:8
Node: sepearation in y. Threshold:6
Leaf:(1,1)(3,2)(5,3)(7,4)(7,1)(2,5)
Node: sepearation in y. Threshold:8
Leaf:(3,7)(7,7)
Node: sepearation in x. Threshold:3.5
Leaf:(2,9)
Leaf:(5,9)
Leaf:(9, 8)(9, 5)(10, 2)(10,10)





3.1 Confusion-Matrix

Predicted

Actual	**Positive**	**Negative**	**Total**
Positive	80%	25%	10
Negative	20%	75%	8
Total	10	8	18

3.2 Kennzahlen

Precision Hund = 25/34 = 73.5%
Precision Katze = 20/26 = 76.9%
Precision Vogel = 27/32 = 84.4%
Precision Hase = 25/31 = 80.6%
Precision Igel = 26/29 = 89.7%
Precision Fisch = 25/28 = 89.3%

Recall Hund = 25/30 = 83.3%
Recall Katze = 20/30 = 66.7%
Recall Vogel = 27/30 = 90.0%
Recall Hase = 25/30 = 83.3%
Recall Igel = 26/30 = 86.7%
Recall Fisch = 25/30 = 83.3%

Overall Accuracy = 1/180 * (25+20+27+25+26+25) = 82.2%