Übung Clustering I+II (23. + 30.10.2019)

Aufgabe 1: Zeichnen Sie eine 2-dimensionale Matrix (Maßstab 1cm pro Bein bzw. Dezimeter) und ordnen Sie folgende Tiere ein:

Name	Beine	Ca. Länge [dm]
1.Schimpanse	2	12
2.Biene	6	0.1
3.Vogelspinne	8	1
4.Hund	4	8
5.Katze	4	5
6.Gans	2	5

A) Erstellen Sie durch hierarchisches Clustern per Hand (Manhattan-Abstand, Single Linkage) einen „Stammbaum“ der Tiere.
B) Teilen Sie per Hand die Tiere mit K-Means in 2 Gruppen ein. Die zufälligen Mittelpunkte der Cluster liegen in diesem Fall bei (2/2) und (6/10).
C) Ist die Gruppeneinteilung mit K-Means immer gleich, ganz egal wie die zufälligen Startwerte gelegt werden? Wie viele verschiedene Einteilungen sind bei diesem Beispiel möglich?

Aufgabe 2: In der Datei „Milch.csv“ sind Daten zur Zusammensetzung der Muttermilch verschiedener Tierarten enthalten.
A) Führen Sie eine Clusteranalyse mit und ohne Standardisierung durch. Sind die Unterschiede signifikant? Welchen Einfluss hat eine Änderung der Metrik (Euklidisch/Manhattan/Canberra)?
B) In einer zweiten Datei „qMilch.csv“ sind Daten zu zwei unbekannten Tierarten A und B enthalten. Versuchen Sie diese über ihre Ähnlichkeit zu den anderen Daten einer Tierart zuzuordnen.

Aufgabe 3: Die Datei „Gebiss.csv“ enthält Daten zu den Gebissformen verschiedener Tierarten.
A) Ist eine Standardisierung notwendig? Welchen Einfluss hat eine Änderung der Metrik (Euklidisch/Manhattan) und Methode (Single, Complete, Average, Ward)?
B) Ermitteln Sie eine optimale Zahl für die Clusterbildung.
C) Sind die gebildeten Cluster (biologisch) sinnvoll?

Aufgabe 4: Laden Sie die Datei „Lebenserwartung.csv“.
A) Analysieren Sie die Daten mit einem partitionierenden Verfahren.
B) Clustern Sie in Abhängigkeit von Alter und Geschlecht. Wie verändert sich die Zahl und Zuordnung der Klassen bei der Betrachtung unterschiedlicher Merkmale? Was schlussfolgern Sie daraus?

Aufgabe 5: Laden Sie die Expressionsdaten von 128 Patienten mit Akuter lymphoblastischer Leukämie (ALL). Der Datensatz beinhaltet 12625 Gene, zusätzliche Informationen: Geschlecht: weiblich (F), männlich (M); Tumorstadium: 1, 2, 3, 4; ALL-Subtyp: B-cell ALL (B), T-cell ALL (T).

A) Analysieren Sie die Daten mit der k-nearest-neighbour-Methode. Teilen Sie den Datensatz in Trainings- (60 B-Zell, 20 T-Zell) und Testdaten (35 B-Zell, 13 T-Zell) auf. Für k=1: Wie gut können die Subtypen in den Testdaten vorhergesagt werden? Wie z.B. ab k=19? Ändern Sie auch die Aufteilung der Trainings- und Testdaten und vergleichen Sie die Vorhersagewerte. Bitte jeweils begründen.
B) Analysieren Sie die Daten mit der k-means-Methode. Verwenden Sie folgende Einstellungen: 50 Gene, 2/3/10 Cluster, 10 Wiederholungen. Wie ist die Clusterverteilung? Wiederholen Sie die Analyse für eine Aufteilung in die B-/T-Subtypen und Tumorstadien (2/3/10 Cluster). Vergleichen Sie die Clusterverteilung. Welche Einteilung würden Sie bevorzugen (bitte begründen)?