Dr. Konrad Völkel
Heinrich-Heine-Universität Düsseldorf
Wintersemester 2024/25

Data Science 2
Blatt 10, Abgabe am 08.01.2025 um 12:00

Bitte geben Sie ein Notebook und die daraus erstellte PDF-Datei ab, sowie gegebenenfalls weitere Codefragmente (`.py`-Dateien). Wenn Sie zum Testen in Aufgabe 2 einen kleinen (!) Teildatensatz erstellt haben, können Sie diesen auch als `.csv` beilegen, und nicht nur den Code, um diesen zu generieren (damit die Auswertung schneller geht).

Aufgabe 1. *(40%) Implementieren Sie BFR-Clustering und vergleichen Sie Ihre Implementierung auf dem Iris-Flower-Datensatz mit $k = 3$ mit dem Fitten eines gemischten Gaussschen Modells mit Scikit-Learn, jeweils mit voller Kovarianzmatrix, diagonaler Kovarianzmatrix (wie bei BFR) und Identitätsmatrix als Kovarianzmatrix (= soft k-Means). Visualisieren Sie diesen Vergleich der 4 Clusteringmethoden geeignet und beurteilen Sie den Tradeoff zwischen Ergebnisqualität und Rechenaufwand (Zeit).*

Sie sollen in dieser Implementierung noch keinerlei Parallelisierung verwenden, also nur mit einem Kern/Prozess/Thread arbeiten.

Aufgabe 2. *(30%)*

1. *Sie haben bereits eine explorative Datenanalyse auf dem New York Yellow Taxi Cab Datensatz 2023 (wahlweise 2009–2024) durchgeführt. Welche Features des Datensatzes könnten als Vektoren zum Clustering verwendet werden?*
2. *Überlegen Sie sich, wie die Ergebnisse eines Clusterings für eine Aufgabe genutzt werden können (etwa um eine Frage zu beantworten, die ein Taxiunternehmen haben könnte).*
3. *Führen Sie den Vergleich der 4 Clusterings mit Visualisierung aus der vorigen Aufgabe auf einer hinreichend kleinen Teilstichprobe des Taxi Datensatzes mit diesen Features durch.*

Datenquelle: www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Aufgabe 3. *(30%) Schätzen Sie ab, wie lange Ihr Programm in Aufgabe 1 benötigen würde, um den Gesamt-Taxidatensatz zu clustern.*

Überlegen Sie sich einen Weg, wie Parallelisierung genutzt werden könnte, um BFR-Clustering (oder eine Approximation/Variante davon) schneller zu berechnen. Überlegen Sie sich dabei etwas, dass Sie selbst auch implementieren könnten. Beschreiben Sie Ihr Vorgehen in kurzen, knappen Sätzen oder Stichpunkten.

Tipp: es kann hilfreich sein, zunächst über Parallelisierung einer einzelnen k-Means-Iteration nachzudenken.

Tipp 2: Die finale parallelisierte Implementierung von BFR wird Gegenstand des folgenden Übungsblatts sein, Sie können also durchaus auch schon damit anfangen.

Frohe Feiertage und Guten Rutsch!