ICW - Dichte-basierte Clusteringverfahren für große Datenmengen

Problem: 1-2 Mio. Datensätzen, bestehend aus 100-200 numerischen Featuren (Vektoren). Sie möchten wissen, ob es in diesen Datensätzen eine Struktur gibt, ob Cluster von Datensätzen existieren.

Sie haben sich für ein dichtebasiertes Clustering Verfahren entschieden (warum ist dies sinnvoll?), nehmen eine Bibliothek (z.b. Sklearn), starten das clustern auf ihrem Laptop/PC. Nach einer Stunde rechnet das Programm immer nur noch, auch nach sechs Stunden. Sie sehen sich das Programm nochmal an, können aber keinen Fehler finden. Da es mittlerweile schon n Abend ist, lassen sie das Programm weiter Rechnen, am nächsten Morgen sind sie gespannt, ob das Programm fertig ist aber es läuft noch. Sie müssen zur Uni und lassen das Programm weiter rechnen. Sie fangen an zu überlegen, was passiert da? Ist das ein Fehler in der Implementierung? Gibt es ein Problem in den Daten? Gibt es ein Problem mit dem Verfahren? Sie fangen an mehr über das Verfahren zu lesen. Nach 5 Tagen ist ihr Rechner immer noch nicht fertig und sie brechen das Programm schließlich ab, mit dem schlechten Gefühl: “vielleicht wäre der Prozess ja Inder nächsten Stunde terminiert?”

Rahmenbedingung von ICW

Lt. Modulbeschreibung:

- eine schriftliche modulbegleitend geprüfte Studienleistung (Hausarbeit oder Projektarbeit)
- Lernergebnis/Kompetenzen:
  - Die Studierenden lernen selbständig ein Fachgebiet zu erarbeiten, entweder durch Erarbeitung eines Selbststudienprogramms, durch Durchführung eines Forschungsprojekts oder durch die Erstellung eines Produktes.
  - Sie bauen ihre Kompetenzen im Bereich des selbstgesteuerten Lernens mit Praxisrelevanz aus.

Inhalte

ICW 1 Einarbeitung, Abschätzung und Suche nach Lösungsmöglichkeiten

SoSe20

Vorgehen:

- Lesen Sie die Papiere zu DBSCAN, HDBSCAN, OPTICS
  - Wie funktionieren diese Verfahren?
  - Wie ist ihre Komplexität?
- Schätzen Sie ab wie lange ihr Prozess gerechnet hätte
  - Generieren Sie künstliche Daten
    - zufällig, normalverteilt mit 10 Cluster
  - Ermitteln Sie wie lange das Verfahren auf einem Rechner, einem Kern für eine Teilmenge der Daten benötigt

        - 1000, 10000, 50000 Datensätzen
    - rechnen Sie die Laufzeit hoch
- Überlegen Sie/Finden Sie heraus, ob es Optimierungsmöglichkeiten gibt
    - welche sind das?
    - wie funktionieren die?
    - was bringen Sie?
- Wie können Sie ermitteln, ob in die Datensätze eine Struktur besitzen?
    - würde der Ansatz auch funktionieren, wenn es sehr viele Cluster gibt, die nur aus 2-5 Datensätzen bestehen?
    - wie müssten diese im Datenraum verteilt sein?

Erwartetes Ergebnis (Noten relevant)

- Kurzer Bericht (10-15 Seiten) über Beschreibung der Verfahren, bestimmte Rechenaufwände und Optimierungsmöglichkeiten

ICW 2 Exploration alternativer Clustering-Ansätze

WS 20/21

Noch weiter zu präzisieren

- Recherchieren Sie nach alternativen, parallelisierbaren Ansätzen
    - Beschreiben Sie diese
    - Welche Komplexität besitzen diese Verfahren?
    - Kann mit diesen Verfahren ihr Problem gelöst werden? Innerhalb von 5 Tagen

Erwartetes Ergebnis

Noch weiter zu präzisieren