Prof. Dr. F. Kammer
M. Huber
Fachbereich MNI
TH Mittelhessen

Grundlagen der DS (INF2207) (WS 24/25)

24.10.2024 Übung 2 2

Die folgenden Aufgaben lösen Sie in Python.

Aufgabe 1 (Apriori 1). Verstehen Sie das Beispiel unter der Adresse: https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/ und führen Sie es selbst bei sich aus. Sie sollen in der Lage sein, den Code zu erklären.

Aufgabe 2 (Apriori 2). Laden Sie den Datensatz *shopping.csv* in Ihr Notebook/Umgebung und machen Sie sich mit dem Datensatz vertraut. Wie oft wird ein Produkt relativ zur Anzahl aller Transaktionen gekauft? Welche Kombinationen von Produkten werden am häufigsten zusammen gekauft (filtern Sie Warenkörbe der Größe eins heraus)? Verwenden Sie hierfür die Bibliothek mlxtend. *Hinweis:* Konvertieren Sie das DataFrame in eine Liste von (Warenkorbs)-Listen.

Aufgabe 3 (K-Means Datenvorbereitung).
Laden Sie den Datensatz *vehicles.csv* aus Moodle herunter. Die Bedeutung einzelner Spalten können Sie der nachfolgenden Adresse entnehmen – wundern Sie sich nicht, dass die heruntergeladene .csv-Datei weniger Spalten enthält: https://www.fueleconomy.gov/feg/ws/index.shtml#vehicle.
Filtern Sie anschließend die Daten, indem Sie:

(a) Nur Fahrzeuge behalten, die Front-, Heck- oder Allradantrieb haben.
(b) Nur Fahrzeuge behalten, die eine der folgenden Energiequellen verwenden: Diesel, Benzin, Erdgas, Elektro.
(c) Nur Fahrzeuge behalten, die kein stufenloses Getriebe verwenden.

Gruppieren Sie anschließend den Datensatz nach Jahr, Hersteller und Modell. Extrahieren Sie als neues Merkmal die Anzahl der Gänge, die ein Auto besitzt, sowie die Art des Getriebes: manuell oder automatisch.
Säubern Sie die Daten abschließend, indem Sie für jede relevante Spalte statistische Kennzahlen ermitteln. Nutzen Sie hierfür vorgegebene Pandas-Funktionen sowie eigene Berechnungen. Achten Sie auch auf NaN-Werte und fehlerhafte Werte und ersetzen Sie diese sinnvoll.

Hinweis: Achten Sie beispielsweise darauf, dass Elektrofahrzeuge einen sinnvollen Kraftstoffverbrauch besitzen.

Aufgabe 4 (K-Means).

(a) Erweitern Sie ihre Lösung aus der vorherigen Aufgabe. Diskretisieren Sie die CO2-Spalte Ihres Datensatzes, beispielsweise mithilfe eines Binnings (z.B. Kein Ausstoß, Geringer Ausstoß, ..., Hoher Ausstoß). Ordnen Sie jedem Ihrer Binnings eine Farbe zu und erstellen Sie einen 3D-Scatterplot mit Variablen, die Sie für geeignet halten, um Ihre vorher gebildeten Kategorien wiederzuerkennen. Achtung: Für viele Autos ist der CO2-Ausstoß nicht bekannt. Färben Sie jeden Punkt Ihres Scatterplots entsprechend der Kategorie ein. Erkennen Sie ein Muster?

(b) Wenden Sie den K-Means-Algorithmus aus der scikit-learn Bibliothek auf Ihren Datensatz an. Visualisieren Sie die gefundenen Cluster. Ist das Ergebnis sinnvoll genug, um Autos, bei denen der CO2-Ausstoß nicht bekannt ist, grob einordnen zu können?
Eine Möglichkeit dies zu beanworten wäre die Accuracy zu berechnen. Dies kann zum Beispiel mit der Funktion accuracy_score aus der sckit-learn-Bibliothek erfolgen.

Studentische Präsentationen: 31.10.2024