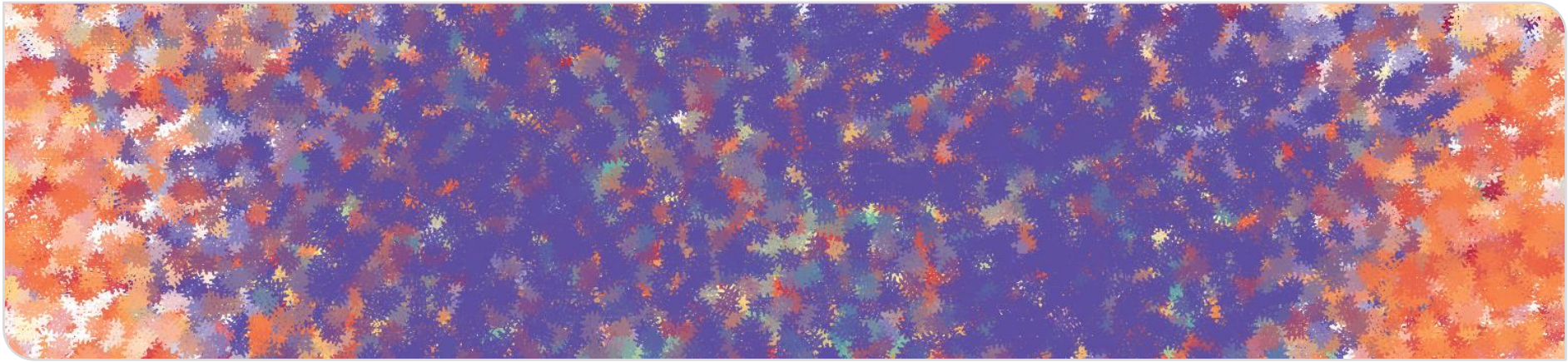
Karlsruher Institut für Technologie

AGD – WS 2020/21 – Übungssitzung 1

Jakob Bach (jakob.bach@kit.edu)



(Vorläufiger) Zeitplan

Woche	Datum*	Übung
1	03.11.2020	
2	10.11.2020	
3	17.11.2020	Abgabe Blatt 1
4	24.11.2020	
5	01.12.2020	Abgabe Blatt 2
6	08.12.2020	Sitzung 1
7	15.12.2020	Abgabe Blatt 3

Woche	Datum*	Übung
8	22.12.2020	
9	12.01.2021	Abgabe Blatt 4
10	19.01.2021	Sitzung 2
11	26.01.2021	Abgabe Blatt 5
12	02.02.2021	Sitzung 3
13	09.02.2021	Abgabe Blatt 6
14	16.02.2021	Sitzung 4

* regulärer Vorlesungstermin am Dienstag; Abgabetermin der Übungsblätter kann ein anderer Tag sein

Agenda

Inhalt: Übungsblätter 1/2, Vorlesungskapitel 1-3

- Umfrage: Vorbereitung auf Sitzung
- Eure Fragen
  - Code-Aufgaben
  - Online-Tests
  - Sonstiges
- Theorie-Aufgaben

Ankündigung: Umfrage zur Übung im ILIAS vom 08.12. bis zum 13.12.

Fragen

- Aufzeichnung?
  - Nein, da ihr zu sehen und zu hören seid → aus Datenschutzgründen keine Aufzeichnung erlaubt
- Letzte Übungsaufgabe (Praktikumsqualifikation)?
  - Kleine Fallstudie, Bearbeitung eines Data-Science-Problems
  - Exploration, Klassifikation und Evaluation
  - Folien zu typischer Data-Science-Pipeline werden noch bereitgestellt
- Basis 2 für Logarithmus bei Entropie des Splits?
  - Andere Basis geht auch, selbe Ergebnisse bei Vergleichen (Absolutwerte der Entropie lediglich mit Konstante multipliziert)
  - Die 2 bezieht sich auf binäre Codierung der Daten

Theorieaufgaben – Aggregation (1)

- Welche Aggregationsfunktionen sind self-maintainable?
  - Self-maintainable bezieht sich auf Einfügen oder Löschen
  - Prinzipiell gibt es sehr viele Aggregationsfunktionen, daher hier nur Beispiele
  - sum(), count() sind self-maintainble beim Einfügen und Löschen
  - min(), max() sind self-maintainable beim Einfügen, aber nicht Löschen
  - mean(), median() sind nicht self-maintainable
    - Bei Median muss man Häufigkeiten aller Werte kennen
    - Bei Mean muss man noch Anzahl oder Summe der Werte kennen

Theorieaufgaben – Aggregation (2)

- Wie ist eine Aggregationsfunktion mathematisch definiert?
  - (Formal nicht in der Vorlesung definiert, insofern hier mehrere Ansätze)
  - Im engeren Sinn: Abbildung von Vektor eines Datentyps auf einen Skalar
  - Im etwas weiteren Sinn: Ergebnis können auch mehrere Werte sein, z.B. Mittelwert und Anzahl (damit Mittelwert über Teilmengen aggregiert werden kann)
  - Im sehr weiten Sinn: einige Datenreduktionstechniken wie Histogramme aggregieren auch

Theorieaufgaben – Median und Mittelwert

- Was sagt uns das Verhältnis von Median und arithmetischem Mittel über die Verteilung der Daten?
  - Aussage über Streuung der Daten und Outlier → wie schief die Daten sind
  - einseitige Outlier: arithmetisches Mittel verschoben, Median dennoch gleich
  - In AGD2-Vorlesung wird „Schiefe“ (Skewness) auch formal eingeführt, definiert als drittes zentrales Moment (Varianz ist zweites zentrales Moment): https://de.wikipedia.org/wiki/Schiefe_(Statistik)

Theorieaufgaben – Entropie (1)

- Was sind klassische Punkte für Entropie-Splits?
  - Eine Interpretation der Frage: Wo soll man splitten?
    - dort, wo Split-Entropie minimal ist
  - Eine andere Interpretation der Frage: Wo soll man Split-Entropie berechnen?
    - Numerische / ordinale Attribute (1): für alle (im Datensatz vorhandenen) Werte des Attributs
    - Numerische / ordinale Attribute (2): für alle Punkte in der Mitte zweier aufeinanderfolgender Werte des Attributs (nachdem die Werte sortiert wurden)
    - Nominale Attribute (1): alle Aufteilungen der Attributwerte in zwei Mengen
    - Nominale Attribute (2): alle Aufteilungen der k Attributwerte in eine k-1-elementige Menge und eine einelementige Menge (also immer nur einen Attributwert absplitten)
    - Nominale Attribute (3): Attributwerte in eine beliebige Reihenfolge bringen und nur Splitpunkte bzgl. dieser Reihenfolge untersuchen

Theorieaufgaben – Entropie (2)

- In welchem Kontext kann man das Konzept der Split-Entropie anwenden?
  - Typischerweise bei Entscheidungsbäumen
  - Generell auch für Diskretisierung von Attributen, ohne dass danach ein Entscheidungsbaum gelernt werden muss

Theorieaufgaben – PCA

- Warum erfasst die erste Principal Component immer die höchste Standardabweichung / Varianz aus den Ursprungsdaten?
  - PCs erhält man als Eigenvektoren der Kovarianzmatrix der Ursprungsdaten
  - Höhe der erfassten Varianz ist proportional zu zugehörigen Eigenwerten
  - Eigenwerte werden (für die Bestimmung der Reihenfolge der PCs) der Größe nach sortiert, sodass erste PC die meiste Varianz erfasst

Theorieaufgaben – Datenreduktion

- Was für grundlegende Arten der Datenreduktion gibt es und wie unterscheiden Sie sich?
  - Dimensionality reduction: Anzahl der Attribute reduzieren (bei Datensatz in Tabellenform: Anzahl der Spalten reduzieren), z. B. Feature Selection, PCA
  - Numerosity reduction: Anzahl der Datenobjekte reduzieren (bei Datensatz in Tabellenform: Anzahl der Zeilen reduzieren), z. B. Sampling, Clustering
  - Diskretisierung: Anzahl der Attributwerte reduzieren, z. B. entropiebasierte Splits, Chi2-Merge

Theorieaufgaben – NULL-Werte

- Was macht man bei der Analyse mit NULL-Werten im Datenbestand?
  - Rausschmeißen (d.h. aus Attribut entfernen bzw. alle Datenobjekte entfernen, in denen mindestens ein Attributwert NULL ist)
    - Kann problematisch werden bei Datensätzen mit wenigen Datenobjekten und/oder vielen Attributen
  - Ersetzen
    - Numerische Attribute: z. B. durch Durchschnitt, Median
    - Kategorische Attribute: z. B. durch häufigsten Wert
  - Bei kategorischen Attributen: als separate Kategorie betrachten

Theorieaufgaben – Statistische Tests

- Welche statistischen Tests gibt es?
  - (Hier nur die aus der Vorlesung)
  - Chi-Quadrat-Test: Unabhängigkeit von Verteilungen zweier Zufallsvariablen als Nullhypothese
  - Kolmogorov-Smirnov-Test: identische Verteilungsfunktionen zweier Zufallsvariablen als Nullhypothese
  - Wilcoxon-Mann-Whitey-Test (Mann-Whitney-U-Test): identischer Median der Verteilungen zweier Zufallsvariablen als Nullhypothese
  - Je nach Quelle / Software auch andere Nullhypothesen der Tests, siehe z. B. Lösung zu Aufgabe 3 des 1. Übungsblatts

Theorieaufgaben – Chi-Quadrat-Test

- Warum kann ein Chi-Quadrat-Test nicht sinnvoll auf kontinuierlichen Daten angewandt werden?
  - Test basiert auf Häufigkeiten der Attributwerte
  - Wenn die Werte kontinuierlich sind: sehr viele Zeilen & Spalten in Kontingenztabelle, kein sinnvoller Vergleich von erwarteten Einträgen und tatsächlich vorliegenden Einträgen möglich (andererseits steigt aber auch Anzahl der Freiheitsgrade und somit wird es schwieriger, niedrigen p-Wert zu erhalten; Test selbst hat also eine Art „Gegenmittel“)
  - Selbes Problem kann auch bei kategorischen Attributen mit vielen Kategorien auftreten