DHBW Mannheim
Studiengang Wirtschaftsinformatik/Data Science

15. November 2021

Introduction to Data Science
Übungsblatt 1
Lösungsskizze

1. Wissenstreppe [7 Punkte]

Beschreiben Sie die ersten vier Stufen der Wissenstreppe. Geben Sie ein Beispiel an.

Lösung:

- **Zeichen:** Elementare Bausteine zur Beschreibung der Daten → Alphabet / Wörter

  ASCII

- **Daten:** Syntaktisch korrekte, symbolische Darstellungen, bestehend aus einzelnen Datenelementen und Werten (Datentyp: Syntaktische Struktur der Datenelemente)

  T = 16, P = 928, R = CEU

- **Information:** Bringt Daten in einen semantischen Kontext

  T = 16 C°, P = 928 mbar, R = CEU (Mitteleuropa)

- **Wissen:** Systematische Verknüpfung von Information

  T = 16 C°, P = 928 mbar, R = CEU (Mitteleuropa), wahrscheinlich Regen

2. Strukturierte Daten [5 Punkte]

Was sind strukturierte Daten? Wodurch zeichnen sich strukturierte Daten aus?

Lösung:

Daten mit einer explizit vorgegebenen *syntaktischen und semantischen* Struktur

- In der Regel: Darstellung durch Attribute und zugehörige Werte
- Daten können mehrdimensionale Strukturen haben (Vektoren, Matrizen, Relationen, …)

Struktur gibt zwingend die Darstellung, Wertebereiche, Beziehungen, Einschränkungen der einzelnen Datenelemente vor.

3. Data-Science-Prozess – Zentrales Ziel [7 Punkte]

Beschreiben Sie das zentrale Ziel von Data Science?

Lösung:

Wissensgewinnung und Handlungsempfehlungen erstellen:

- Aus bereitgestellten Daten **Wissen und Erkenntnisse** gewinnen

- **Modelle** erstellen, die dieses Wissen beschreiben
- Mit diesen Modellen (neue) Daten-Objekte oder Situationen beschreiben und bewerten sowie geschäfts- oder betriebsrelevante Handlungsempfehlungen ableiten (**actionable insights**)

4. Data-Science-Prozess – Transformation [10 Punkte]

1. Was ist die Aufgabe des Prozessschrittes Transformation?
2. Aus welchen Teilschritten setzt sich die Transformation zusammen?
3. Beschreiben Sie die einzelnen Schritte.

Lösung:

1. Umwandlung der bereinigten Daten zur Verwendung in den Data-Science-Szenarien
2. Teilschritte
    a. **Harmonisierung**: Syntaktische und semantische Abstimmung gefilterter Daten aus unterschiedlichen Quellen sowie deren Anhebung auf eine einheitliche Granularität.
    b. **Integration**: Zusammenführung sowie einheitliche Bereitstellung und Zugriffsmöglichkeiten der harmonisierten Daten .
3. Beschreibung der Teilschritte
    a. Harmonisierung
        i. semantisch und syntaktisch
        ii. Vergleichbarkeit der Daten und Entitäten
        iii. Auflösung von Redundanzen
    b. Integration
        i. Reduktion, Verdichtung, Diskretisierung
        ii. Hierarchiebildung – mehrdimensionale Daten
        iii. Aggregation: Zusammenfassung von Daten – z. B. durch Summen- oder Mittelwertbildung – auf Basis definierter Aggregationspfade (Hierarchiepfade).
        iv. Anreicherung der Daten durch weitere berechnete betriebswirtschaftlich sinnvolle Kennzahlen und abgeleitete Daten
        v. Erzeugung diverser Sichten, Filterung

5. Arten des Lernens [10 Punkte]

a. Nennen Sie vier wichtige Arten des maschinellen Lernens.
b. Beschreiben Sie die Konzepte von zwei Arten des maschinellen Lernens genauer.

Lösung:

a. Arten des maschinellen Lernens
    i. Supervised Learning (Überwachtes Lernen)
    ii. Unsupervised Learning (Unüberwachtes Lernen)
    iii. Self-taught Learning / Semi-supervised Learning (Autodidaktisches Lernen / teilüberwachtes Lernen)
    iv. Reinforcement Learning (Verstärkendes Lernen)
b. Beschreibung von Konzepten von zwei Arten des maschinellen Lernens
    i. Supervised Learning:
        I. Ausgangslage: Datenobjekte haben Eingangs- oder Prädiktorvariablen X (Daten) und Zielvariablen Y. Wert Y soll aus Wert X (mit gewisser Wahrscheinlichkeit) ermittelbar sein.
        II. Ziel: Erlerne Funktion/Modell M, womit für jedes Datenobjekt aus dem Wert X der Zielwert Y (mit gewisser Wahrscheinlichkeit) bestimmt werden kann: Y=M (X)

III. Vorgehensweise: Lernen von Funktion/Modell M mit Trainingsdaten. Lernen endet, wenn ein Modell M mit ausreichender Güte bestimmt wurde. Güte wird mit annotierten/gelabelten Testdaten gemessen

ii. Unsupervised Learning:
   I. Ausgangslage: Datenobjekte haben Eingangsvariablen X aber keine Zielvariablen.
   II. Ziel: Finde Modell M, das (bisher unbekannte, nützliche) Strukturen, Beziehungen in den Daten, etc. beschreibt.
   III. Vorgehensweise: Lernen oder Finden der relevanten Strukturen mit (nicht-annotierten) Trainingsdaten. Lernen ohne Anleitung!

6. Trainingsdatensatz [6 Punkte]

a. Was ist ein Trainingsdatensatz?
b. Was ist ein annotierter/gelabelter Trainingsdatensatz?
c. Für welche Arten von Lernverfahren werden annotierte/gelabelte Trainingsdatensätze verwendet?

Lösung:

a. Ein Trainingsdatensatz wird von einem Lernalgorithmus zum Erlernen eines bestimmten (Wissens-)Modells verwendet.
b. In einem annotierten/gelabelten Trainingsdatensatz werden für alle Datenobjekte die zugehörigen Zielvariablenwerte bereitgestellt.
c. Überwachtes Lernen.