Übung: KI T-InfT-008 und 010 Datenmengen und Embedded Systems

Cândido Vieira

Zu Unterrricht 10.10.2024

Balthasar-Neumann-Technikum (BNT)

Inhaltsverzeichnis - Übungen

1. Kategorische Variable

- **Kategorische Variable**:

    - Eine **kategorische Variable** ist eine Variable, deren Werte aus einer festen Menge von Kategorien oder Labels bestehen. Diese Werte sind diskret und haben oft keine numerische Bedeutung.
        - Beispiel:
            - Variable: Augenfarbe
                - Kategorien: Blau, Grün, Braun

            - Wenn wir Kategorien numerisch kodieren (z. B. Blau = 1, Grün = 2, Braun = 3), dann sind diese Zahlen willkürlich und bedeuten keine Rangfolge oder Abstände.

1. Kategorische Variable

- **Nominale (nonordinal) Variable:**
  - Eine **nominale Variable** ist eine Unterart von kategorischen Variablen, bei denen die Kategorien keine natürliche Reihenfolge oder Rangfolge haben.

    - Beispiel 1:
      - Variable: Automarke
        - Kategorien: BMW, Audi, Tesla
        - Kodierung: BMW = 1, Audi = 2, Tesla = 3
      - Die Zahlen bedeuten hier keine Reihenfolge. BMW ist nicht "kleiner" oder "größer" als Audi.

1. Kategorische Variable

- **Nominale (nonordinal) Variable:**
    - Beispiel 2:
        - Variable: Wetter
        - Kategorien: Sonnig, Regen, Bewölkt
        - Kodierung: Sonnig = 1, Regen = 2, Bewölkt = 3
    - Auch hier haben die Zahlen keine numerische Bedeutung. Sie dienen nur der Darstellung.

1. Kategorische Variable

- Unterschied zu **ordinalen Variablen:**
  - Eine **ordinale Variable** hat dagegen Kategorien, die eine natürliche Reihenfolge haben.

  - Beispiel:
    - Variable: Bewertung
    - Kategorien: Ausreichend, Gut, Sehr gut
    - Kodierung: Ausreichend =3, Gut = 2, Sehr gut = 1
  - Hier hat die Reihenfolge eine Bedeutung, da "Sehr gut" besser ist als "Gut".

1. Kategorische Variable

- Aufgabe: Identifiziere kategorische und nominale Variablen
  - **Übung 1**: Kodierung
    - Die folgende Tabelle zeigt Variablen mit ihren Kategorien und einer Kodierung. Entscheiden Sie, ob die Kodierung sinnvoll ist und welche Variable nominal ist.

Variable	Kategorien	Kodierung
Augenfarbe	Blau, Grün, Braun	Blau = 1, Grün = 2, Braun = 3
Bewertung	Ausreichend, Gut, Sehr gut	Ausreichend = 3, Gut = 2, Sehr gut = 1
Automarken	BMW, Audi, Tesla	BMW = 1, Audi = 2, Tesla = 3

Frage: Welche der Variablen sind nominal?

Hinweis: Nominale Variablen haben keine Reihenfolge in ihren Kategorien.

1. Kategorische Variable

- Aufgabe: Identifiziere kategorische und nominale Variablen
  - **Übung 2**: Erstellen Sie eigene nominale Variablen
    - Denken Sie sich eine nominale Variable aus und definieren Sie Kategorien.
    - Kodierung: Weisen Sie den Kategorien Zahlen zu und erklären Sie, warum die Zahlen keine numerische Bedeutung haben.

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
  - Es gibt verschiedene Methoden, um **kategorische Variablen** in numerische Werte zu kodieren, die für Machine Learning-Modelle verwendet werden können. Die drei häufigsten sind:
    - **1. One-Hot Encoding**
      - **Beschreibung**: Für jede Kategorie wird eine eigene Spalte erstellt. Jede Spalte enthält entweder eine 1 (wenn die Beobachtung zu dieser Kategorie gehört) oder eine 0 (wenn sie nicht dazugehört).
      - **Anzahl der Spalten**: Eine Spalte pro Kategorie.
      - **Eigenschaften**:
        - Es gibt **keine Redundanz**, weil jede Kategorie eindeutig repräsentiert wird.
        - Es ist geeignet für **nominale Variablen** ohne Reihenfolge.
      - **Beispiel**: Variable: **Farbe** mit den Kategorien: Rot, Blau, Grün
        - One-Hot Encoding erstellt die folgenden Spalten:

Rot	Blau	Grün	
1	0	0	(für Rot)
0	1	0	(für Blau)
0	0	1	(für Grün)

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **1. One-Hot Encoding**
        - **Vorteil**: Einfach und klar. Keine Gefahr von Multikollinearität.
        - **Nachteil**: Kann bei vielen Kategorien sehr viele Spalten erzeugen (hohe Dimensionalität).

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **2. Dummy Coding**
        - **Beschreibung**: Ähnlich wie One-Hot Encoding, aber es wird **eine Kategorie als Referenz** gewählt und nicht kodiert. Die anderen Kategorien werden in binäre Spalten umgewandelt.
        - **Anzahl der Spalten**: Eine Spalte weniger als die Anzahl der Kategorien.
        - **Eigenschaften**:
            - **Reduziert die Anzahl der Spalten**, um Multikollinearität zu vermeiden.
            - Die Referenzkategorie dient als Basis für den Vergleich (z. B. in linearen Modellen).

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **2. Dummy Coding**
        - **Beispiel**: Variable: **Farbe** mit den Kategorien: Rot, Blau, Grün (Referenzkategorie: Grün)
            - Dummy Coding erstellt die folgenden Spalten:

Rot	Blau	
1	0	(für Rot)
0	1	(für Blau)
0	0	Grün - (für Referenzkategorie)

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **2. Dummy Coding**
        - **Vorteil**: Effizienter als One-Hot Encoding, insbesondere für lineare Modelle.
        - **Nachteil**: Die Interpretation kann schwieriger sein, da die Referenzkategorie nicht explizit dargestellt wird.

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **3. Effect Coding**
        - **Beschreibung**: Ähnlich wie Dummy Coding, aber die Referenzkategorie wird mit **-1** kodiert, anstatt sie wegzulassen.
        - **Anzahl der Spalten**: Eine Spalte weniger als die Anzahl der Kategorien.
        - **Eigenschaften**:
            - Verwendet sowohl positive (1) als auch negative (−1) Werte, um die Balance zwischen den Kategorien zu gewährleisten.
            - Wird häufig in statistischen Analysen verwendet, da sie **keine Abhängigkeit von einer spezifischen Referenzkategorie** hat.

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **3. Effect Coding**
        - **Beispiel**: Variable: **Farbe** mit den Kategorien: Rot, Blau, Grün (Referenzkategorie: Grün)
            - Effect Coding erstellt die folgenden Spalten:

Rot	Blau	
1	0	(für Rot)
0	1	(für Blau)
-1	-1	(für Grün - Referenzkategorie)

2. Kodierungsmethoden

- **Unterschiede zwischen den Kodierungsmethoden für kategorische Variablen:**
    - **3. Effect Coding**
        - **Vorteil**: Unabhängig von der Referenzkategorie. Verwendet alle Informationen, einschließlich der Referenzkategorie.
        - **Nachteil**: Komplexere Interpretation der Werte in Modellen.

2. Kodierungsmethoden

- **Vergleich der Methoden:**

Eigenschaft	One-Hot Encoding	Dummy Coding	Effect Coding
Anzahl der Spalten	Anzahl Kategorien	Kategorien - 1	Kategorien - 1
Referenzkategorie	Nein	Ja	Ja (-1 kodiert)
Einfachheit	Sehr einfach	Mittel	Komplex
Gefahr von Multikollinearität	Nein	Nein	Nein
Interpretation	Leicht	Mittel	Schwieriger

2. Kodierungsmethoden

Wann welche Methode verwenden?

1. **One-Hot Encoding**:
   a. Verwenden Sie diese Methode, wenn Sie Modelle wie Entscheidungsbäume, Random Forests oder Neuronale Netze verwenden. Sie profitieren von klaren und unabhängigen binären Spalten.
   b. Geeignet für nominale Variablen (z. B. Farbe, Geschlecht).

2. **Dummy Coding**:
   a. Diese Methode ist vorteilhaft für lineare Modelle (z. B. Lineare Regression, Logistische Regression), da sie effizienter ist und Multikollinearität reduziert.
   b. Besonders nützlich, wenn Sie Vergleiche mit einer Referenzkategorie interpretieren möchten.

3. **Effect Coding**:
   a. Verwenden Sie diese Methode, wenn Ihre Analyse **statistisch fokussiert** ist und Sie keine Abhängigkeit von einer Referenzkategorie haben möchten.
   b. Wird häufig in ANOVA und anderen experimentellen Designs verwendet.

2. Kodierungsmethoden

Übung 1: Vergleich der Kodierungsmethoden

Anleitung:

Sie haben Daten zu Mietpreisen in drei deutschen Städten: Berlin, München und Hamburg. Die Zielvariable ist der Mietpreis, und die unabhängige Variable ist die Stadt.

Gegebene Daten:

Stadt	Mietpreis (€)
Berlin	1500
Berlin	1600
München	2000
München	2100
Hamburg	1300
Hamburg	1400

2. Kodierungsmethoden

Aufgaben:

1. **One-Hot Encoding**
   - Kodieren Sie die Städte mit One-Hot-Encoding.
   - Wie viele Spalten werden benötigt?
   - Warum hat One-Hot-Encoding keine Referenzkategorie?
   - Zeichnen Sie ein Diagramm, das zeigt, wie die Spalten für jede Stadt gefüllt werden (z. B. 1 für die Stadt und 0 für die anderen).
2. **Dummy Coding**
   - Wählen Sie **Hamburg** als Referenzkategorie.
   - Kodieren Sie die Städte mit Dummy Coding.
   - Wie viele Spalten werden benötigt?
   - Was bedeutet ein Koeffizient für Berlin, wenn das Modell Hamburg als Basis verwendet?
3. **Effect Coding**
   - Kodieren Sie die Städte mit Effect Coding.
   - Hamburg soll die Referenz sein.
   - Warum wird in dieser Methode -1 verwendet? Wie wird die Referenzkategorie berücksichtigt?

2. Kodierungsmethoden

Übung 2: Wählen Sie die geeignete Kodierungsmethode

Aufgabe:

Wählen Sie die geeignete Kodierungsmethode basierend auf den folgenden Szenarien:

1. Sie möchten eine einfache Darstellung der Kategorien ohne Bezug zu einer Referenzkategorie.
2. Sie möchten Hamburg als Referenzkategorie verwenden, um die Mietpreise in Berlin und München mit Hamburg zu vergleichen.
3. Sie möchten, dass alle Kategorien gleichwertig in das Modell eingehen, und die Referenzkategorie soll durch -1 dargestellt werden.

3. Normalisierung

Übung 1: Min-Max-Skalierung für den Bereich [0,1]

Gegeben ist der Vektor x:

$$x=[2,4,6,3,10]$$

1. Formulieren Sie die **Min-Max-Skalierungsformel**, die die Werte von x in den Bereich **0 bis 1** transformiert.
2. Verwenden Sie Ihre Formel, um die transformierten Werte für alle Elemente im Vektor x zu berechnen. Schreiben Sie die Ergebnisse in einer übersichtlichen Tabelle.

3. Normalisierung

Übung 2: Min-Max-Skalierung für den Bereich [a,b]

Gegeben ist der Vektor x:

$$x=[2,4,6,3,10]$$

1. Wenden Sie die **Min-Max-Skalierungsformel** an, um die Werte von x in den **Bereich [−1,1]** zu transformieren.
2. Formulieren Sie die allgemeine **Min-Max-Skalierungsformel**, mit der Werte von x in einen beliebigen **Bereich [a, b]** transformiert werden können.
3. Berechnen Sie die transformierten Werte für den gesamten Vektor x und tragen Sie die Ergebnisse in eine Tabelle ein. (**Bereich: a bis b**)

3. Normalisierung

Übung 3: Umkehrung der Min-Max-Skalierung

1. Verwenden Sie die transformierten Werte aus **Übung 1** (Bereich: 0 bis 1).
2. Formulieren Sie die Formel, mit der Sie aus den transformierten Werten wieder die ursprünglichen Werte berechnen können.
3. Wenden Sie Ihre Formel an, um die ursprünglichen Werte für alle transformierten Werte zu berechnen.