DHBW Mannheim
Studiengang Wirtschaftsinformatik/Data Science

5. Dezember 2021

Introduction to Data Science
Übungsblatt 3

1. Zusammenhänge von Merkmalen [6 Punkte]

Alle zu analysierenden Datenobjekte O_j eines Datensatzes DS = {O_j |j=1, ... ,n} seien durch numerische Variablen X_j und Y_j beschrieben.

1. Welche Größe gibt einen statistischen Hinweis darauf, dass die beiden Merkmale X und Y sich in die gleiche Richtung bzw. entgegengesetzte Richtungen ändern?
2. Geben Sie diese Größe explizit unter Zuhilfenahme des Datensatzes DS an.
3. Wann ändern sich die beiden Merkmale X und Y in die gleiche Richtung und wann in die entgegengesetzte Richtung?

Lösung:

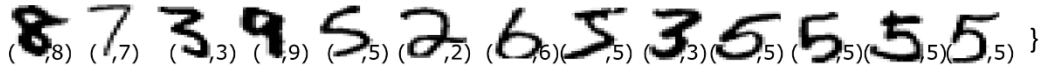
1. Die Größe ist die Kovarianz. Große Kovarianz deutet auf einen Zusammenhang hin.
2. Definition mit Mittelwerten $\overline{x}$ und $\overline{y}$:

$$\mathrm{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

3. Ein positiver Wert der Kovarianz zeigt an, dass sich die beiden Merkmale in die gleiche Richtung ändern. Ein negativer Wert zeigt an, dass sich die beiden Merkmale in entgegengesetzte Richtungen ändern.

2. Gütemaße binärer Klassifikatoren – Score und Schwellenwert [8 Punkte]

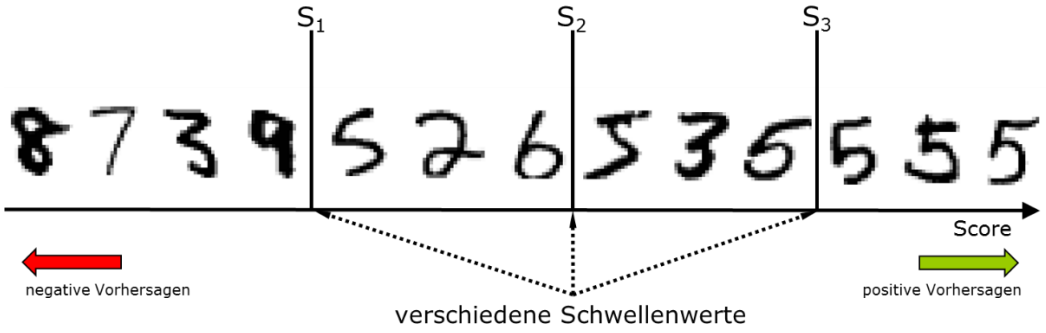
Sei **K** ein binärer, schwellenwertbasierter Klassifikator für handschriftliche Ziffern, der erkennen soll, ob eine Ziffer eine 5 oder keine 5 ist. Sei

TE = {  }

ein Testdatensatz bestehend aus 13 annotierten handschriftlichen Bildern von Ziffern.

Der Klassifikator berechnet für jedes Bild einen Score, der schematisch in der folgenden Grafik dargestellt ist. Der Klassifikator ermittelt aus dem Score und einem Schwellenwert S die Klassifikation (Score > S: positive, Score ≤ S: negative).

Berechnen Sie jeweils für die Schwellenwerte S_1, S_2 und S_3 anhand der folgenden Grafik die Gütemaße Relevanz und Sensitivität von **K** auf dem gegebenen Testdatensatz.



Lösung:

Relevanz $\frac{tp}{tp+fp}$:	$\frac{6}{9} = 0,\bar{6} \approx 67\%$	$\frac{5}{6} = 0,8\bar{3} \approx 83\%$	$\frac{3}{3} = 1 = 100\%$
Sensitivität $\frac{tp}{tp+fn}$:	$\frac{6}{6} = 1 = 100\%$	$\frac{5}{6} = 0,8\bar{3} \approx 83\%$	$\frac{3}{6} = 0,5 = 50\%$

3. Spam [3 Punkte]

Würden Sie die Aufgabe, Spam-E-Mails zu erkennen als überwachte oder unüberwachte Lernaufgabe einstufen? Begründen Sie Ihre Antwort.

Lösung:

Spam-E-Mail-Erkennung ist eine typische überwachte Lernaufgabe. Dem Lernalgorithmus werden viele E-Mails und deren Labels – Spam oder Nicht-Spam – zum Lernen bereitgestellt.

4. Frequent Item Sets – A-Priori-Algorithmus [15 Punkte]

1. Machen Sie sich die Apriori-Eigenschaft in Warenkörben klar:

   Jede Untermenge X einer häufigen Artikel-Menge Y ist häufig.

   Und im Umkehrschluss:

   Wenn die Artikel-Menge X innerhalb einer Artikel-Menge Y nicht häufig ist, dann ist auch die Artikel-Menge Y nicht häufig.

2. Mit dem Apriori-Algorithmus werden alle häufigen Artikelmengen in einer Transaktionsdatenbank (= Warenkorbdatenbank) bestimmt.

   a. Machen Sie sich zuerst das Apriori-Ausschneideprinzip unter Zuhilfenahme von Aufgabe (1) klar:

      Wenn eine Artikel-Menge X nicht häufig ist, verwerfe alle Obermengen von X

   b. Machen Sie sich nun damit den Apriori-Algorithmus klar. Dabei sei im Folgenden X eine **k-Artikel-Menge**, wenn X = $\{x_1, \ldots, x_k\}$, also wenn X genau k Artikel enthält.

      Apriori-Algorithmus:

      i. Scanne die Transaktionsdatenbank, um alle häufigen 1-Artikel-Mengen zu bestimmen. Nehme die häufigen 1-Artikel-Mengen in der Menge L_1 auf.

      ii. Setze nun k = 1 und $L_k = L_1$

iii. Erzeuge aus L_k alle (k+1)-Artikel-Mengen, die nur häufige Artikel-Mengen aus L_k , L_{k-1} , … , L_1 enthalten und nehme diese (k+1)-Artikel-Mengen in C_{k+1} auf.

Dabei spielt das Apriori-Ausschneideprinzip eine entscheidende Rolle!

iv. Test jeden Kandidaten in C_{k+1} auf seine Häufigkeit.

v. Nehme die so ermittelten häufigen (k+1)-Artikel-Mengen in der Menge L_{k+1} auf.

vi. Beende den Prozess, falls L_{k+1} die leere Menge ist.

vii. Ansonsten setze k := k + 1 und fahre mit Schritt (iii) fort.

3. Das Sortiment eines Supermarktes bestehe aus den fünf Artikeln A, B, C, D, E. In diesem Supermarkt wurden an einem Tag die folgenden vier Warenkörbe in der Transaktionsdatenbank registriert:

Warenkorbnummer (TID)	Artikel
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Verwenden Sie den Apriori-Algorithmus, um alle häufigen Artikel-Mengen in der Transaktionsdatenbank zu bestimmen. Support für die Häufigkeitsbestimmung sei **S = 0,5**.

Lösung:

1. siehe Folien und Diskussion in der Vorlesung
2. siehe Folien und Diskussion in der Vorlesung
3. Der Apriori-Algorithmus liefert folgende häufigen Artikelmengen L_1 , L_2 , L_3 :