Übung Informationsintegration

Duplikaterkennung

Aufgabe 1 (Verifikation): Gegeben seien Gold-Standard

$$\mathcal{C}_{\text{gold}} = \{\langle 1, 5, 7\rangle, \langle 2, 4\rangle, \langle 8, 10\rangle, \langle 3, 9\rangle, \langle 6\rangle\}$$

und Duplikaterkennungsergebnis

$$\mathcal{C} = \{\langle 1, 5\rangle, \langle 2, 4, 7\rangle, \langle 6, 8\rangle, \langle 3\rangle, \langle 9\rangle, \langle 10\rangle\}$$

- Bestimmen Sie die Menge der Matches fÃ¼r jedes der beiden Clusterings (d.h. $M(\mathcal{C}_{\text{gold}})$ und $M(\mathcal{C})$).
- Bestimmen Sie die True Positives, die False Positives und die False Negatives von $\mathcal{C}$ in Bezug auf $\mathcal{C}_{\text{gold}}$.
- Verwenden Sie diese drei Mengen um den Recall, die Precision und den F_1-score von $\mathcal{C}$ in Bezug auf $\mathcal{C}_{\text{gold}}$ zu berechnen.

Aufgabe 2 (Sorted Neighborhood Method): Gegeben sei folgende Tabelle:

ID	Vorname	Nachname	Jahr
1	Florian	Panse	1983
2	Marko	Meier	1989
3	Fabian	Pause	1981
4	Felix	Mueller	1985
5	Mirko	Maier	1990
6	Stefan	Meier	1979
7	Fabian	Hanse	1980
8	Norbert	Ritter	1968

- Berechnen Sie für jedes Tupel den Schlüsselwert der sich aus den ersten beiden Buchstaben des Vornamens, den ersten beiden Buchstaben des Nachnamens und den letzten zwei Ziffern des Jahres zusammensetzt.
- Benutzten Sie die eben berechneten Schlüsselwerte um die Sorted Neighborhood Method mit einer Fenstergröße $w = 3$ auf den acht Tupeln durchzuführen.
- Berechnen Sie um wieviel Prozent sich der Suchraum reduziert hat.

Aufgabe 3 (Distanzbasierte Entscheidungsmodelle): Gegeben seien die folgenden drei Vergleichsvektoren:

$$\vec{c}_1 = \langle 0.4, 0.7, 0.1 \rangle$$
$$\vec{c}_2 = \langle 0.5, 0.5, 0.5 \rangle$$
$$\vec{c}_3 = \langle 0.1, 0.2, 0.9 \rangle$$

- Berechnen Sie für jeden dieser Vektoren die Vektorraumdistanz $dst(\vec{c}_i, \mathbf{0})$. Verwenden Sie dafür das euklidische Distanzmaß.

Aufgabe 4 (Regelbasierte Entscheidungsmodelle): Gegeben sei ein gelabelter Trainingsdatensatz der aus den folgenden zwölf Vergleichsvektoren besteht:

Vergleichsvektor	Label (Matchingklasse)
$\vec{c}_1 = \langle 0.5, 0.2, 1.0 \rangle$	Match
$\vec{c}_2 = \langle 0.9, 0.8, 0.3 \rangle$	Match
$\vec{c}_3 = \langle 0.5, 0.5, 0.8 \rangle$	Unmatch
$\vec{c}_4 = \langle 0.9, 0.9, 0.4 \rangle$	Match
$\vec{c}_5 = \langle 0.5, 0.9, 0.4 \rangle$	Unmatch
$\vec{c}_6 = \langle 0.5, 0.3, 0.7 \rangle$	Unmatch
$\vec{c}_7 = \langle 0.8, 0.8, 0.9 \rangle$	Match
$\vec{c}_8 = \langle 0.4, 0.6, 0.7 \rangle$	Match
$\vec{c}_9 = \langle 0.7, 0.8, 0.7 \rangle$	Match
$\vec{c}_{10} = \langle 0.6, 1.0, 0.2 \rangle$	Match
$\vec{c}_{11} = \langle 0.6, 0.9, 0.1 \rangle$	Match
$\vec{c}_{12} = \langle 0.7, 0.8, 0.5 \rangle$	Unmatch

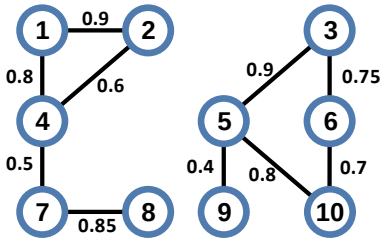
- Versuchen Sie auf Basis dieser Daten ein Profil mit positiven und negativen Regeln zu erstellen, welches diese Vektoren korrekt klassifiziert. Beachten Sie, dass die Reihenfolge der Regeln Auswirkung auf das Klassifikationsergebnis eines Tupelpaares hat.

Aufgabe 5 (Statistische Entscheidungsmodelle): Gegeben Sei ein Entscheidungsmodell, dass ein binäres Matching Pattern mit drei Positionen verwendet, d.h. jedes Matching Pattern ist ein drei-dimensionaler Vektor, wobei jede Position für genau ein Attribut steht.
Gegeben seien zudem folgende Informationen:

- Satz 1: 100% aller Matches und 10% aller Unmatches haben den Wert 1 im ersten Attribut.
- Satz 2: 40% aller Matches und 50% aller Unmatches haben den Wert 1 im zweiten Attribut.
- Satz 3: 70% aller Matches und 2% aller Unmatches haben den Wert 1 im dritten Attribut.
- Satz 4: Auf jedes Match kommen genau 499 Unmatches
- Berechnen Sie auf Basis der vorliegenden Informationen die Wahrscheinlichkeiten $Pr(\vec{p}_1 \mid M)$ und $Pr(\vec{p}_1 \mid U)$ für das Matching Pattern $\vec{p}_1 = \langle 1, 0, 1 \rangle$. Nehmen Sie hierfür an, dass alle Attribute unabhängig sind (d.h. Naive Bayes).

- Bestimmen Sie, ob ein Tupelpaar welches dieses Pattern besitzt als Match oder Unmatch klassifiziert werden würde indem Sie die beiden zugehörigen Koeffizienten berechnen (die Logarithmen können dabei ausser Acht gelassen werden, da sie das Ergebnis nicht beeinflussen).

Aufgabe 6 (Partitioning based on Centers): Gegeben sei folgender Duplicate-Pair Graph:



- Benutzen Sie die Methode *Partitioning based on Centers* um die Konflikte zwischen den einzelnen paarweisen Duplikatsentscheidungen aufzulösen.