Klausur Data Mining - Sommersemester 2010

Lehrstuhl Wirtschaftsinformatik II

Präambel

In dieser Klausur können maximal 60 Punkte erreicht werden. Erläuterungen sind kurz zu fassen. Bei Berechnungen genügt es nicht, das Ergebnis zu nennen, der Lösungsweg soll erkennbar sein.

1 Grundlagen 12

a) Definiere den Begriff *Klassifikation* im Kontext von Data Mining. (3)

b) Definiere den Begriff *Clustering* im Kontext von Data Mining. (3)

c) Welche drei Eigenschaften muss ein Attribut besitzen, um als Zielattribut bei der Klassifikation genutzt werden zu können? (3)

d) Nenne die sechs Phasen des CRISP-DM Prozesses. (3)

2 Klassifikation 22

2.1 Naive Bayes 5

Gegeben ist folgender Datensatz bestehend aus 10 Instanzen mit drei Attributen (A, B, C) und einem Zielattribut (Class):

Instanz	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

a) Bestimme alle bedingten Wahrscheinlichkeiten für die drei Attribute, bei gegebenem Zielattribut. Trage die Ergebnisse in eine Tabelle ein. (3)

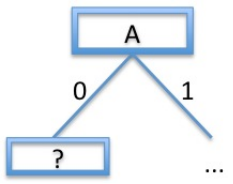


Abbildung 1: Entscheidungsbaum für Aufgabe 2.2 b.

b) Wie würde nach Naive Bayes die Instanz (A=0, B=1, C=0) klassifiziert werden? (2)

2.2 Entscheidungsbäume 8

a) Wie groß ist die Entropie des Zielattributes des Datensatzes aus Aufgabe 2.1? (1)

b) Ein Entscheidungsbaum Algorithmus hat sich bereits für das Attribut A im Wurzelknoten entschieden. Welches Attribut muss nach dem Information Gain für den Knoten unter dem linken Ast des Baumes verwendet werden (vergleiche Abbildung 1)? (5)

c) Welches Problem kann die Benutzung des Information Gain als Gütefunktion mit sich bringen? Wie kann es gelöst werden? (2)

2.3 Evaluation 9

Ein Algorithmus erzeugt für die Daten aus Aufgabe 2.1 einen Klassifikator, welcher die Trainingsdaten wie folgt klassifiziert:

Instanz	1	2	3	4	5	6	7	8	9	10
Klassifizierung	-	-	-	+	+	-	-	-	+	+

a) Stelle für das oben genannte Beispiel die Confusion Matrix auf. (2)

b) Wie groß ist die Accuracy des Klassifikators? (1)

c) Wie groß ist Precision und Recall bezüglich der positiven Klasse? (2)

d) Erweitere die Confusion Matrix um eine weitere Klasse k. Wie berechnet sich der Recall bezüglich dieser Klasse? (2)

e) Die Accuracy hat bei gewissen Eigenschaften der Daten Probleme, ein gutes Qualitätsmaß zu sein. Nenne zwei Eigenschaften der Daten bei denen das der Fall ist. (2)

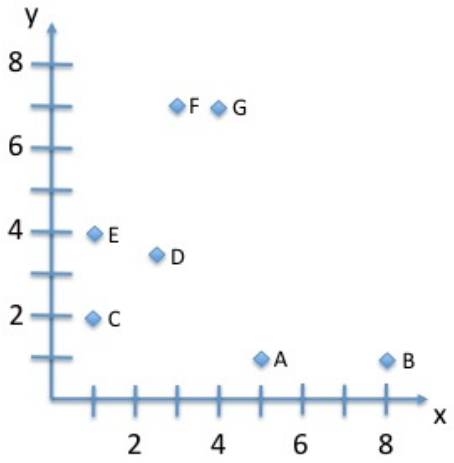


Abbildung 2: Datenpunkte für Aufgabe 3.1 b.

3 Clustering 19

3.1 Algorithmen 11

a) Beschreibe kurz die Funktionsweise des K-Means Algorithmus. (5)

b) Gegeben sind sieben Datenpunkte im zweidimensionalen Raum (vergleiche Abbildung 2):

Instanz	A	B	C	D	E	F	G
x	5	8	1	2,5	1	3	4
y	1	1	2	3,5	4	7	7

Erzeuge (d.h. skizziere) das Dendrogramm unter Benutzung der MIN Inter-Cluster Similarity (Single Linkage), bei euklidischer Distanz. (6)

3.2 Evaluation 8

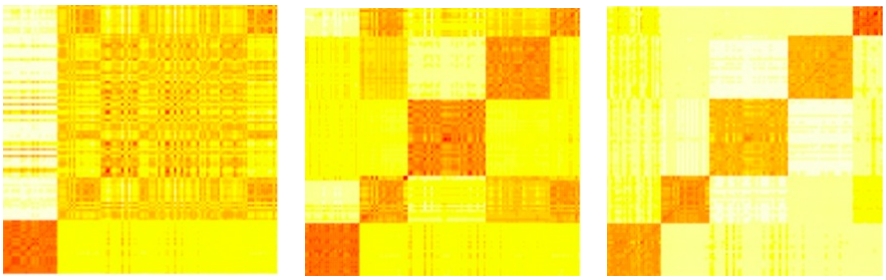
a) Zwei Clustering Algorithmen erzeugen für die Datenpunkte aus Aufgabe 3.1 b jeweils ein Clustering mit jeweils zwei Clustern:

	Cluster 1	**Cluster 2**
Clustering 1	{A, B}	{C, D, E, F, G}
Clustering 2	{A, B, C}	{D, E, F, G}

Wie groß ist der Jaccard-Coefficient zwischen diesen beiden Clusterings? (4)

b) Drei verschiedene Clustering Algorithmen haben jeweils ein Clustering erzeugt mit jeweils drei etwa gleich großen Clustern. Abbildung 3 zeigt die

Distanz-Matrizen der Datenpunkte für jedes Clustering, wobei die Datenpunkte nach ihrer Cluster-Zugehörigkeit sortiert sind. Treffe für jedes Clustering eine Aussage bezüglich der Qualität auf Grundlage der entsprechenden Distanzmatrix. (3)



(a) (b) (c)

Abbildung 3: Distanz Matrizen für Aufgabe 3.2 b

c) Mit welchem Wert für den Parameter k sollte K-Means initialisiert werden, wenn als einzige Grundlage für diese Entscheidung Abbildung 3 c zur Verfügung steht? (1)

4 Assoziationsregeln 7

1. Erkläre kurz den Apriori-Algorithmus. (3)
2. Grenze die Begriffe Frequent Itemset, Maximal Frequent Itemset und Closed Itemset von einander ab. (4)