



Klausur im Fach

Big Data Anwendungen

Sommersemester 2017

Angaben zur Klausur

Prüfer: Dr. Stephan Schosser

Datum: 25. Juli 2017

Prüfungsnummer: 21807

Persönliche Angaben (in Druckbuchstaben ausfüllen)

Nachname: _____ Vorname: _____

Matrikelnummer: _____ Fakultät: _____

Bewertung (wird vom Prüfer ausgefüllt)

Aufgabe	1	2	3	Gesamtpunkte	Note
Punkte					

Zugelassene Hilfsmittel

- Nicht-programmierbarer Taschenrechner
ohne Kommunikations- oder Datenverarbeitungsfunktion (lt. Aushang des Prüfungsamtes)

Hinweise zur Klausur

- Die Bearbeitungszeit für diese Klausur beträgt 60 Minuten.
- Die Klausur besteht aus 3 Aufgaben, von denen 3 Aufgaben zu bearbeiten sind.
- Die Klausur umfasst 2 Seiten.
- Die Heftung dieser Unterlagen darf nicht gelöst werden.

Hinweise zur Bearbeitung

- Bitte tragen Sie oben auf diesem Deckblatt zuerst Ihre persönlichen Daten ein.
- Bitte prüfen Sie die Vollständigkeit der Klausur.
- Sie sind dafür verantwortlich, dass das Aufsichtspersonal Ihre Klausur erhält.
- Viel Erfolg beim Lösen der Klausuraufgaben!

Aufgabe 1 (Clustering)**(20 Punkte)**

Gegeben seien folgende Daten:

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

- (a) Geben Sie an welches Attribut Sie für ein kMeans Clustering ohne Transformation nicht nutzen können. Schlagen Sie eine Strategie vor, wie Sie dieses Attribut umwandeln können um es doch zu nutzen. **(3 Punkte)**
- (b) Sehen Sie sich die Daten an. Gehen Sie davon aus, dass Sie die Daten in 3 Cluster unterteilen wollen. Geben Sie an, welche Zuordnung der Datenpunkte zu Clustern Sie erwarten, wenn Sie die Euklidische Distanz als Distanzmaß wählen. Erörtern Sie kurz, wie sie zu einem weniger vorhersehbaren Ergebnis gelangen können. **(3 Punkte)**
- (c) Wenden Sie den kMeans Algorithmus an um die Daten auf Basis der 3 numerischen Attribute in 3 Cluster zu zerlegen. Nutzen Sie dabei die Euklidische Distanz und lösen Sie das Problem rechnerisch. **(12 Punkte)**
- (d) Erläutern Sie kurz was der Unterschied zwischen supervised und unsupervised learning Verfahren ist. **(2 Punkte)**

Aufgabe 2 (Klassifikation)**(20 Punkte)**

Gegeben seien folgende Trainingsdaten:

Empfehlung	Form	Farbe	Material
Kaufen	Formumspielend	rot	Leder
Kaufen	Körperbetont	gelb	Baumwolle
Kaufen	Körperbetont	rot	Leder
Nicht kaufen	Formumspielend	gelb	Leder
Nicht kaufen	Körperbetont	rot	Baumwolle
Nicht kaufen	Körperbetont	gelb	Leder

- (a) Leiten Sie das Attribut Empfehlung mit Hilfe eines Entscheidungsbaums ab. Nutzen Sie hierfür als Splitkriterium den Informationsgewinn und entwickeln Sie solange neue Knoten bis entweder kein Splitattribut mehr verfügbar ist oder in einem Knoten alle Daten in der gleichen Klasse sind. **(10 Punkte)**
- (b) Sagen Sie mit Hilfe des Entscheidungsbaums aus (a) die Empfehlung für folgende Beobachtung voraus (Formumspielend, gelb, Baumwolle). **(3 Punkte)**
- (c) Nennen Sie zwei andere Klassifikationsverfahren neben den Entscheidungsbäumen und diskutieren Sie kurz die Vorteile der Verfahren gegenüber Entscheidungsbäumen. **(4 Punkte)**
- (d) Erläutern Sie wie Entscheidungsbäume zur Feature Selection genutzt werden können. **(3 Punkte)**

Aufgabe 3 (Sonstiges)**(20 Punkte)**

- (e) Erläutern Sie anhand der Datenerzeugung, Datenverarbeitung und Datennutzung wie Big Data Probleme entstehen und grenzen Sie dies von traditionellem Data Mining ab. **(6 Punkte)**
- (f) Erläutern Sie kurz, wie sie mit Hilfe eines sozialen Netzwerks entscheiden, welche Mitglieder für Werbung besonders relevant sind. **(4 Punkte)**
- (g) Geben Sie an, wie Assoziation Rules helfen können um Prospekte für Supermärkte gewinnbringend aufzubauen. **(5 Punkte)**
- (h) Nennen Sie die zwei unterschiedlichen Ansätze für Recommender Systeme und diskutieren Sie deren Vor- und Nachteile. **(5 Punkte)**