OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
FAKULTÄT FÜR WIRTSCHAFTSWISSENSCHAFT

Big Data Anwendungen

Aufgabenblatt 5 (Stream Mining)

Aufgabe 1 – H-Tree (15 Punkte)

Gegeben seien folgende Daten:

Alter	Familienstand	Geschlecht	Churn
jung	ledig	männlich	0
alt	verheiratet	männlich	0
jung	verheiratet	weiblich	0
alt	ledig	weiblich	0
alt	verheiratet	männlich	1
jung	verheiratet	weiblich	1
alt	ledig	weiblich	0
jung	verheiratet	männlich	1

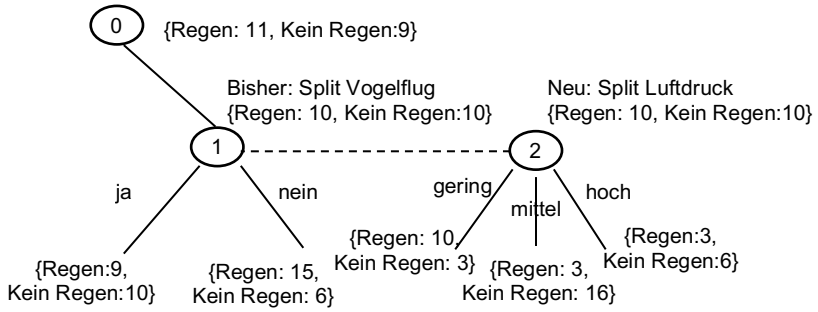
(a) Gehen Sie davon aus, dass die Daten in einem „Data Stream“ sequentiell eintreffen (je weiter oben die Beobachtung in der Tabelle umso früher) und Sie einen H-Tree aufbauen wollen. Ermitteln Sie den 1. Split. Gehen Sie dabei davon aus, dass jeweils nach 4 Beobachtungen über einen Split entschieden wird und das Attribut Churn vorhergesagt werden soll. **(10 Punkte)**

(b) Ermitteln Sie für die gegebenen Daten die Hoeffding Bound und Geben Sie an, ob bei einem Signifikanzniveau von 0,001 ein Split durchgeführt würde. **(5 Punkte)**

Aufgabe 2 – CDH-Tree (20 Punkte)

(a) Beschreiben Sie den Algorithmus für den Aufbau von CDH-Trees. **(5 Punkte)**

(b) Am Ende einer Phase zur Baumerstellung sei folgender Baum entstanden:



Prüfen Sie für, ob der Split „Luftdruck“ den Split „Vogelflug“ ersetzen soll. Gehen Sie dabei von folgenden Testdaten aus: **(10 Punkte)**

Vogelflug	Luftdruck	Regen
ja	hoch	ja
ja	mittel	nein
nein	mittel	ja
ja	gering	ja
ja	mittel	nein
nein	hoch	ja
ja	hoch	nein
nein	gering	ja

(c) Ist es möglich mit einer weiteren Beobachtung für Teilaufgabe (b) die Entscheidung zu ändern? Wenn ja, begründen Sie und geben sie die entsprechende Beobachtung an. **(5 Punkte)**

Aufgabe 3 – Verständnisfragen **(13 Punkte)**

(a) Diskutieren Sie welche Vor- und Nachteile H-Trees gegenüber traditionellen Entscheidungsbäumen besitzen. **(5 Punkte)**

(b) Diskutieren Sie, welche Nachteile der H-Tree bei der Vorhersage von Verkäufen für Saisonware besitzt und wie der CDH-Tree dies adressiert. **(5 Punkte)**

(c) Diskutieren Sie zwei Anwendungsfälle in welchen Daten als Streams auftreten und Stream Mining sinnvoll ist. **(3 Punkte)**