Big Data Anwendungen

Aufgabenblatt 5

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren

- Aufgabenblatt 5 – Stream Mining
  - Aufgabe 1 – H-Tree
  - Aufgabe 2 – CDH-Tree
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 1 (a) - Aufgabenstellung

- Gegeben seien folgende Daten:

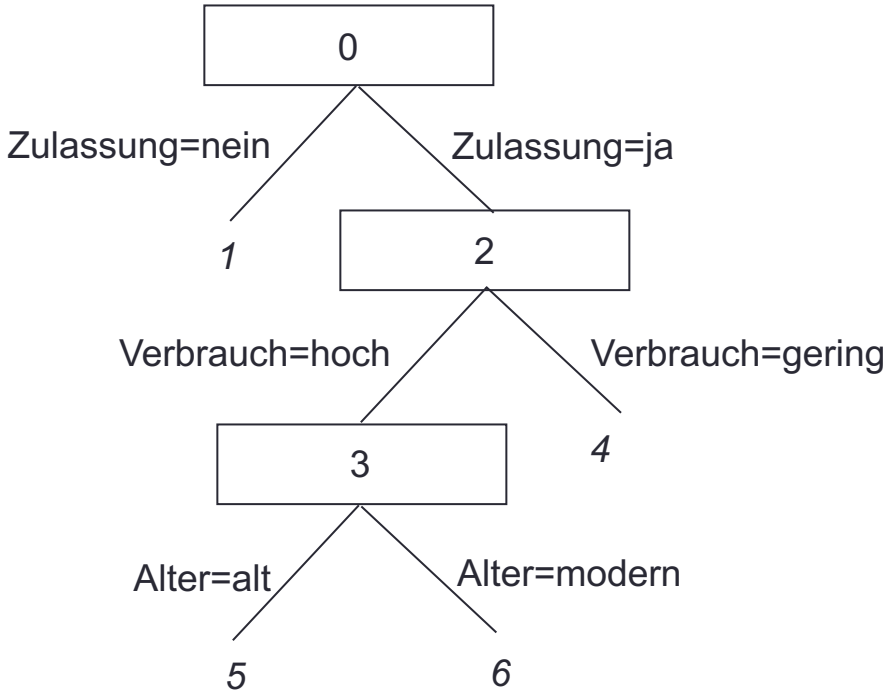
Alter	Familienstand	Geschlecht	Churn
jung	ledig	männlich	0
alt	verheiratet	männlich	0
jung	verheiratet	weiblich	0
alt	ledig	weiblich	0
alt	verheiratet	männlich	1
jung	verheiratet	weiblich	1
alt	ledig	weiblich	0
jung	verheiratet	männlich	1

- 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)

Wiederholung Vorlesung – H-Tree

- Aufbau
  - Knoten entstehen nacheinander
    ... und Entstehung wird nummeriert
  - Split immer dann, wenn Splitkriterium erreicht
  - Splits sind n-är
  - Entscheidungen über „Arrays“
- Beispiel
  - Baum ist entstanden durch
    - Split in 0 ⇒ Knoten 1 und 2
    - Split in 2 ⇒ Knoten 3 und 4
    - Split in 3 ⇒ Knoten 5 und 6
  - Nummer des nächsten Knoten: 7
- Unterschiede zum Entscheidungsknoten
  - Knoten und Blätter haben ID
  - Nächster ID wird gemerkt



Wiederholung Vorlesung – H-Tree

- Idee
  - Speicherung von Arrays an jedem Knoten
  - Arrays sollen Änderungen am Baum erleichtern (enthalten alle Daten für Split und Anwendung)
- Alle Knoten
  - `VerfügbareAttribute`
    Attribute die nicht in Vorgängern für Split genutzt wurde
  - `AnzahlProKlasse`
    Anzahl der Beobachtungen für jede einzelne Klasse
- Innere Knoten
  - `SplitAttribut`
    Gibt genutztes Splitattribut an
- Blattknoten
  - `AnzahlTreffer`
    Zählt die Anzahl der Beobachtungen seit Erstellung
  - `AnzahlProAttributKlasse`
    Anzahl der Beobachtungen für jede einzelne Klasse, Wert Kombination

Aufgabe 1 (a) – Lösung (H-Tree)

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

- Vorgehen
  - Initialisierung Wurzelknoten
  - Split nach n (hier: 4) Beobachtungen
- Initialisierung
  - VerfügbareAttribute: {Alter, Familienstand, Geschlecht},
  - SplitAttribut: {},
  - Anzahl Treffer: 0
  - AnzahlProAttributKlasse:

	jung	alt	ledig	verheiratet	männlich	weiblich
Churn						
kein Churn						

  - AnzahlProKlasse: {Churn: 0, Kein Churn: 0}

Aufgabe 1 (a) – Lösung (H-Tree)

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

- Vorgehen
  - Initialisierung Wurzelknoten
  - Split nach n (hier: 4) Beobachtungen
- Prüfung auf Split nach 4 Beobachtungen
  - VerfügbareAttribute: {Alter, Familienstand, Geschlecht},
  - SplitAttribut: {},
  - Anzahl Treffer: 4
  - AnzahlProAttributKlasse:

	jung	alt	ledig	verheiratet	männlich	weiblich	Summe
Churn	0	0	0	0	0	0	0
kein Churn	2	2	2	2	2	2	4
Summe	2	2	2	2	2	2	

  - AnzahlProKlasse: {Churn: 0, kein Churn: 4}
- ⇒ kein Split möglich – alle Beobachtungen in Klasse „kein Churn“

Aufgabe 1 (a) – Lösung (H-Tree)

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

- Vorgehen
  - Initialisierung Wurzelknoten
  - Split nach n (hier: 4) Beobachtungen
- Prüfung auf Split nach 8 Beobachtungen
  - VerfügbareAttribute: {Alter, Familienstand, Geschlecht},
  - SplitAttribut: {},
  - Anzahl Treffer: 8
  - AnzahlProAttributKlasse:

	jung	alt	ledig	verheiratet	männlich	weiblich	Summe
Churn	2	1	0	3	2	1	3
kein Churn	2	3	3	2	2	3	5
Summe	4	4	3	5	4	4	

  - AnzahlProKlasse: {Churn: 3, kein Churn: 5}
- ⇒ jetzt ist Split möglich

Aufgabe 1 (a) – Lösung (H-Tree)

- Vorgehen
  - Initialisierung Wurzelknoten
  - Split nach n (hier: 4) Beobachtungen

- Prüfung auf Split nach 8 Beobachtungen
  - AnzahlProAttributKlasse:

	jung	alt	ledig	verheiratet	männlich	weiblich	Summe
Churn	2	1	0	3	2	1	3
kein Churn	2	3	3	2	2	3	5
Summe	4	4	3	5	4	4	

- Entropieberechnung
  - Split Alter/Geschlecht

    $(-2 \cdot \log_2 2 - 1 \cdot \log_2 1 - 2 \cdot \log_2 2 - 3 \cdot \log_2 3 + 4 \cdot \log_2 4 + 4 \cdot \log_2 4)/8 \approx 0{,}91$
  - Split Familienstand

    $(-3 \cdot \log_2 3 - 3 \cdot \log_2 3 - 2 \cdot \log_2 2 + 3 \cdot \log_2 3 + 5 \cdot \log_2 5)/8 \approx 0{,}61$

- ⇒ 1. Split ist Familienstand

Aufgabe 1 (b) - Aufgabenstellung

- 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

- 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)

Wiederholung Vorlesung: Hoeffding Bound

- Bisher: Split immer dann, wenn...
  - ... vorgegebene Anzahl neuer Beobachtungen eingetroffen und...
  - ... Anzahl der Beobachtungen im Knoten nicht eindeutig in einer Klasse

- Problem
  - H Tree soll nicht unendlich lange wachsen
  - H Tree soll durch weiteren Split signifikant besser werden

- Lösung: Hoeffding Bound
  - Split nur dann, wenn Informationsgewinn für Split „deutlich besser“...
    ... als Informationsgewinn aller anderen Splits
  - Dabei soll gelten: Informationsgewinn muss umso „deutlich besser“ sein...
    - ... je größer der Wertebereich des Splitattributs
    - ... je weniger Beobachtungen Grundlage für Split bilden
  - Für Split muss also gelten

$$E(\text{beste Alternative}) - E(\text{potentieller Split}) > E \cdot \sqrt{\ln\left(\frac{1}{\delta}\right)/2n}$$

mit E: Entropie ohne Split, n: Anzahl Beobachtungen, δ: Signifikanzniveau

Aufgabe 1 (b) – Lösung (Hoeffding Bound)

- Bisher (Teilfaugabe (a))
  - Entropie Split Alter/Geschlecht: $\approx 0{,}91$
  - Entropie Split Familienstand: $\approx 0{,}61$

	jung	…	weiblich	Summe
Churn	2	…	1	3
kein Churn	2	…	3	5
Summe	4	…	4	

- Vorgehen
  - Prüfung: $E(\text{beste Alternative}) - E(\text{potentieller Split}) > E \cdot \sqrt{\ln\left(\frac{1}{\delta}\right)/2n}$
- Vorbereitung
  - Entropie vor Split: $E = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} \approx 0{,}95$
  - Anzahl Beobachtungen: $n = 8$
  - Signifikanzniveau: $\delta = 0{,}001$ (vgl. Aufgabe)
- Einsetzen
  - $0{,}91 - 0{,}61 > 0{,}95 \cdot \sqrt{\frac{\ln\left(\frac{1}{0{,}001}\right)}{2 \cdot 8}} \Rightarrow 0{,}30 > 0{,}95 \cdot \sqrt{\frac{\ln(1000)}{2 \cdot 8}} \approx 0{,}63$

⇒ Ungleichung ist nicht erfüllt! Der Split würde nicht durchgeführt!

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren

- Aufgabenblatt 5 – Stream Mining
  - Aufgabe 1 – H-Tree
  - Aufgabe 2 – CDH-Tree
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 2 (a) - Aufgabenstellung

- Beschreiben Sie den Algorithmus für den Aufbau von CDH-Trees.

(5 Punkte)

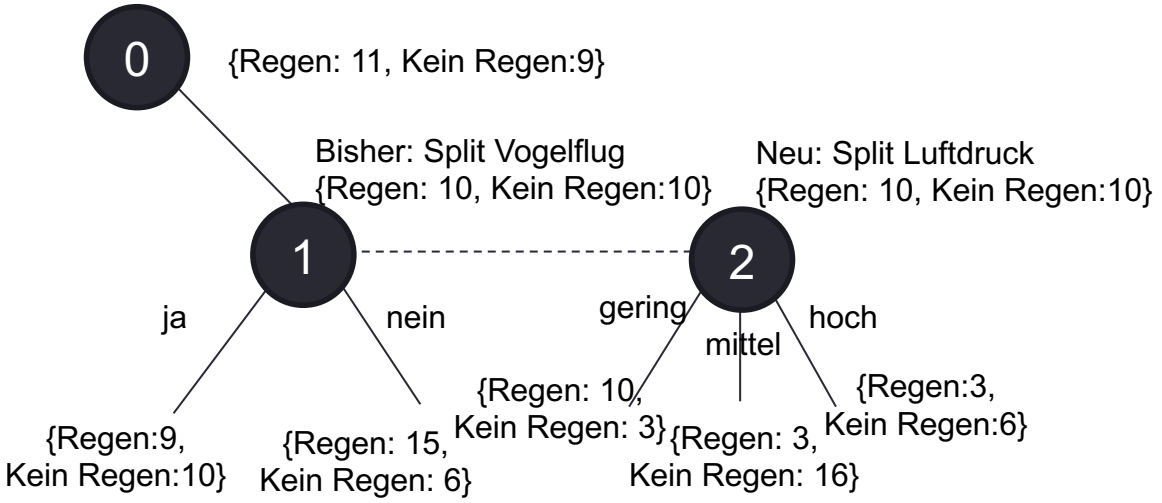
Wiederholung Vorlesung: CDH Tree

- Ziel
  Erweiterung des H Tree, so dass sich das Modell über die Zeit ändern kann

- Inhaltliche Änderungen
  - Alle Arrays werden auch an inneren Knoten aktualisiert
  - Speicherung der Daten über „Sliding Windows“ (FIFO-Methode)
    - Beobachtungen werden in betroffenen Knoten gesammelt
    - Übersteigt die Anzahl der Beobachtungen im Knoten einen Wert n...
      ... Löschen der zuerst eingefügten Beobachtungen
  - Für innere Knoten regelmäßige Prüfung, ob Splits noch analog...
    ... Gegebenenfalls: Umstrukturierung des entsprechenden Teilbaums
    - Phase 1: Aufbau
      - Dafür Identifikation „verdächtiger“ Knoten
        (Alternativer Split gemäß Hoeffding Bound besser als aktueller Split)
      - Für jeden „verdächtigen“ Knoten Aufbau eines parallelen Teilbaums
    - Phase 2: Auswahl
      - Berechnung der Vorhersagegüte aller Teilbäume...
        ... Ersetzen des Ursprungsbaums durch besten Teilbaum

Aufgabe 2 (b) - Aufgabenstellung

- Am Ende einer Phase zur Baumerstellung sei folgender Baum entstanden:



Testdaten:

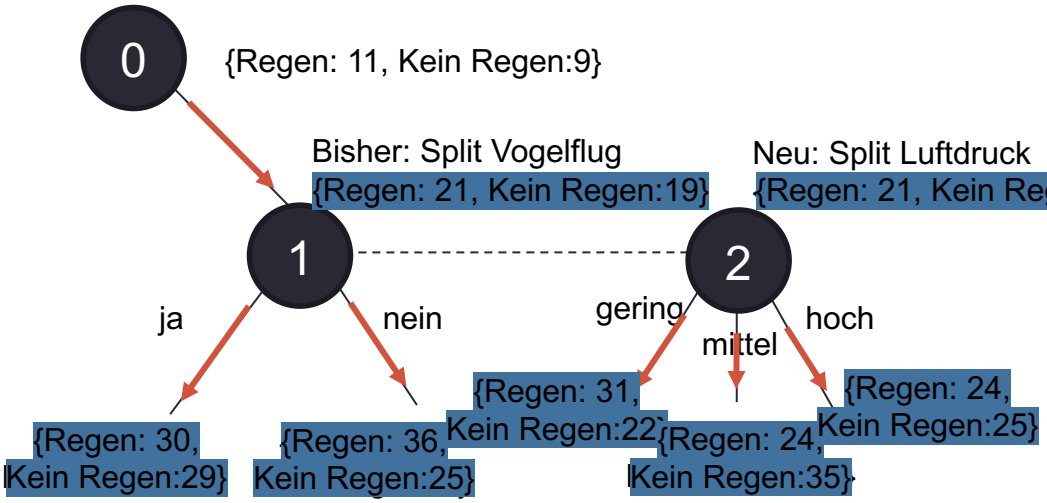
Vogelflug	Luftdruck	Regen
ja	hoch	ja
ja	mittel	nein
nein	mittel	ja
ja	gering	ja
ja	mittel	nein
nein	hoch	ja
ja	hoch	nein
nein	gering	ja

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

Aufgabe 2 (b) – Lösung (Prüfung Teilbaum)

- Vorgehen
  - Ermittlung der Vorhersage in den Blattknoten
  - Prüfen der Vorhersagen mit Testdaten

- Ermittlung der Vorhersagen in den Blattknoten



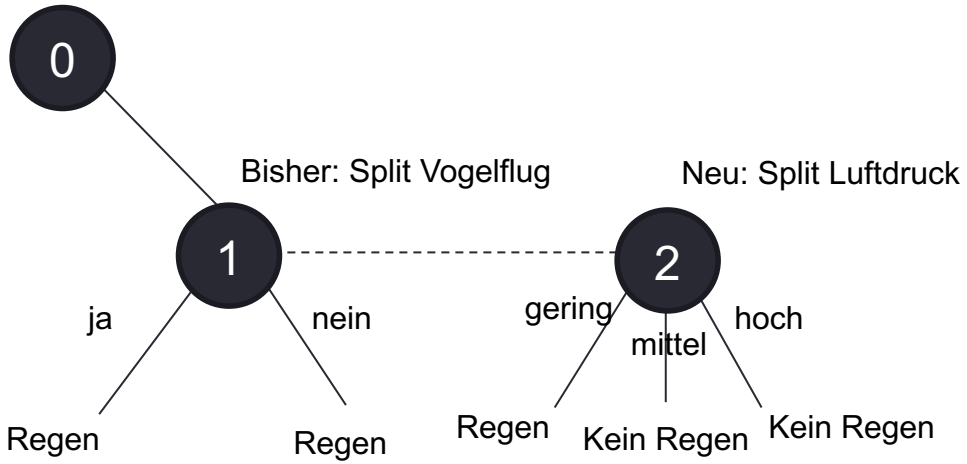
Testdaten:

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

Aufgabe 2 (b) – Lösung (Prüfung Teilbaum)

- Vorgehen
  - Ermittlung der Vorhersage in den Blattknoten
  - Prüfen der Vorhersagen mit Testdaten

- Ergebnis der Vorhersagen in den Blattknoten



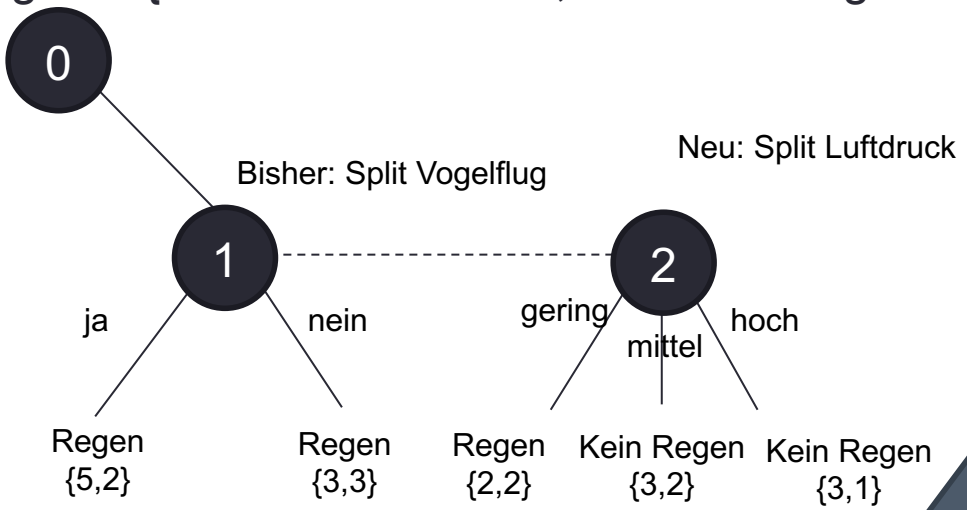
Testdaten:

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

Aufgabe 2 (b) – Lösung (Prüfung Teilbaum)

- Vorgehen
  - Ermittlung der Vorhersage in den Blattknoten
  - Prüfen der Vorhersagen mit Testdaten

- Prüfung der Vorhersagen in den Blattknoten:
  Angabe: {Anzahl klassifiziert, Anzahl richtig klassifiziert)



Testdaten:

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

Knoten 2 („Luftdruck“) ersetzt den bisherigen Knoten 1 („Vogelflug“) nicht, da nicht besser.

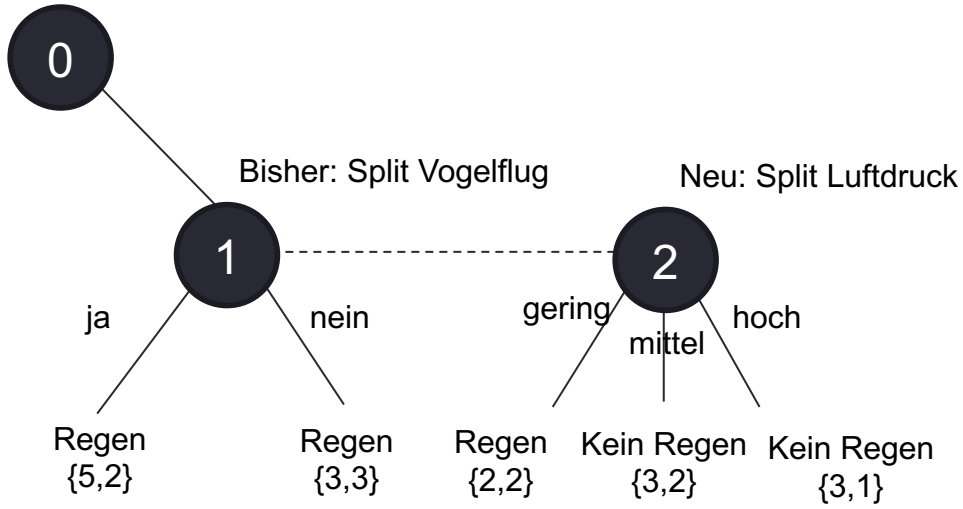
- Bewertung der Splits
  Split Vogelflug: {8, 5}; Split Luftdruck: {8, 5}

Aufgabe 2 (c) - Aufgabenstellung

- 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 2 (c) – Lösung (Prüfung Teilbaum)

- Bisher



Testdaten:

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

- Idee:
  - Beide Splits haben gleiche Vorhersagegüte
  - Eine Beobachtung kann Split Luftdruck besser als Vogelflug machen, …
    … wenn diese bei Luftdruck richtig, aber bei Vogelflug falsch
  - Beispiele:
    - (Vogelflug: ja, Luftdruck: hoch, Regen: nein)
    - (Vogelflug: nein, Luftdruck: mittel, Regen: nein)

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren

- Aufgabenblatt 5 – Stream Mining
  - Aufgabe 1 – H-Tree
  - Aufgabe 2 – CDH-Tree
  - Aufgabe 3 – Verständnisfragen

- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 3 (a) - Aufgabenstellung

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

Wiederholung Vorlesung: H Tree und CDH Tree

- Vorteile der Verfahren gegenüber Entscheidungsbäumen
  - Erlauben kontinuierliches Trainieren und Testen des Modells
  - Anwendung des Modells schon während des Trainierens möglich
  - Verfahren sind performant, da...
    - ... relevante Zwischenergebnisse vorgehalten werden
    - ... Ermittlung abhängig von absoluten Zahlen
  - Kontinuierliche Anpassung (CDH Tree)...
    ... kann auch wandelnde Welt abbilden

- Nachteile
  - Evaluation im „klassischen Sinn“ schwer / kaum möglich
  - „Ausprobieren“ verschiedener Ansätze und Vergleich kaum möglich
  - Management muss Einsatz vertrauen
  - Wie wahrscheinlich ist ein „Use Case“ ohne historische Daten?
    ... bzw. ist Sammeln von Daten und damit höhere Flexibilität nicht besser?

Aufgabe 3 (b) - Aufgabenstellung

- Diskutieren Sie, welche Nachteile der H-Tree bei der Vorhersage von Verkäufen für Saisonware besitzt und wie der CDH-Tree dies adressiert.

(5 Punkte)

Aufgabe 3 (b) – Lösung (H-Tree vs. CDH-Tree)

- Eigenschaften H-Tree
  - Vergisst keine bereits erstellten Knoten
  - Lernt genau ein Mal aus jeder eingehenden Beobachtung
  - Kein „Vergessen“ alter Beobachtungen

- Konsequenz
  - Änderungen auf der Datengrundlage können nicht abgebildet werden
  - Beispiel: keine saisonale Schwankungen, Abnutzungserscheinungen

- Lösung CDH-Tree
  - Durch Ersetzen von Teilbäumen Anpassung des Modells
  - Kontinuierliche Verbesserung des Modells

Aufgabe 3 (c) - Aufgabenstellung

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

Aufgabe 3 (c) – Lösung (Anwendungsfälle)

- Sensornetze (H-Tree)
  - Exemplarische Analyse: Fällt System aus
  - Beschaffenheit der Daten
    - Datenmenge zu groß, um permanent vorzuhalten (bei entsprechender Größe)
    - Vorhersagemodell oft nicht relevant
    - System sollte „autonom“ funktionieren – Keine Evaluierung möglich

- Analyse von Webtraffic (CDH-Tree)
  - Exemplarische Analyse: Vorhersage direktes Verlassen der Webseite
  - Beschaffenheit der Daten
    - Datenmenge zu groß, um vorzuhalten
    - Nutzerverhalten und damit Vorhersage ändert sich kontinuierlich
    - Wertebereich der Eingaben (Browser, Betriebsysteme, …) ändern sich kontinuierlich