Präsenzübung 05
Klassifikation/Entscheidungsbäume

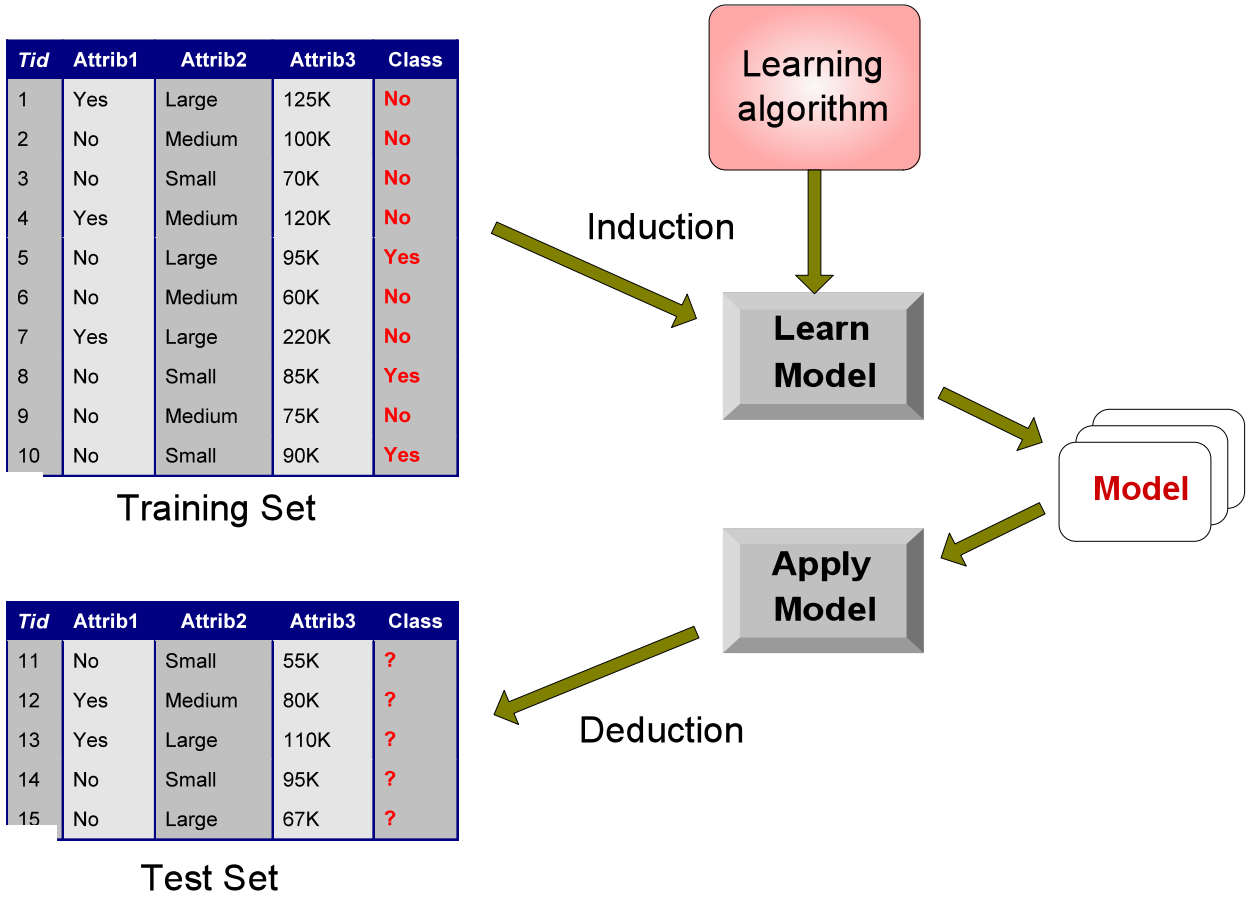
M. Sc. Oliver Pabst

Data Mining SS 15

03.06.2015

Übersicht

Überblick über den Klassifikationsprozess



- Finden eines Modells für die Klassenattribute als Funktion der Werte der anderen Attribute mittels eines Trainingsdatensatzes
- Neue Datensätze sollen so genau wie möglich klassifiziert werden (Vorhersage)
- Bestimmung der Genauigkeit des Modells durch Testdatensatz (Validierung)

Erzeugung eines Klassifikationsbaumes - Hunt's Algorithmus

- Erzeuge neuen Knoten t
- Bezeichne D_t die Datensätze, die den Knoten t erreichen
- Vorgehensweise:
  - wenn D_t nur Datensätze der Klasse y_t enthält, dann ist t ein Blatt mit dem Label y_t
  - wenn D_t leer ist, dann ist t ein Blatt mit dem Label der Standardklasse y_d
  - wenn D_t Datensätze aus mehr als einer Klasse enthält, finde ein Teilungsattribut

Der Algorithmus wird fortgesetzt rekursiv angewendet, um die Teilbäume zu erzeugen, die schlussendlich zum Klassifikationsbaum führen.

ABER:

Wie bestimmt man das Teilungsattribut (Bedingung/beste Teilung)?

Wie kann die beste Aufteilung bestimmt werden?

- Greedy-Ansatz
  - Knoten mit homogener Klassenverteilung bevorzugt
  - Maximierung des Gains (Minimierung der Kindwerte)
- GINI-Index
  - $GINI(t) = 1 - \sum_j [p(j|t)]^2$
- Entropy
  - $Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$
- Klassifikationsfehler
  - $Error(t) = 1 - \max_j p(j|t)$

Beispiel zur Erzeugung für die Erzeugung eines Modells

Leibniz Universität Hannover

Es soll, analog zur Vorlesung, aus dem folgenden Trainingsdatensatz ein Entscheidungsbaum zum Klassenattribut **Betrug** konstruiert werden.
Für die nominalen Attribute **Einzahlung** und **Bank** sollen alle möglichen binären Aufteilungen (Splits) der Form *Attribut* = *Wert* in Betracht gezogen werden. Für das kontinuierliche Attribut **Betrag** soll jede binäre Aufteilung *Attribut* $>$ *Wert* bzw. *Attribut* $\leq$ *Wert* möglich sein.

Einzahlung	Bank	Betrag	Betrug
Nein	Sparkasse	10.000€	Nein
Ja	Volksbank	5.000€	Ja
Nein	Central Bank of Nigeria	50.000€	Ja
Nein	Sparkasse	1.000€	Nein
Ja	Sparkasse	50€	Nein
Nein	Volksbank	20.000€	Nein
Nein	Sparkasse	2.500€	Nein

Tabelle: Trainingsdatensatz für die Erzeugung eines Entscheidungsbaumes

Bestimmung der ersten Aufteilung

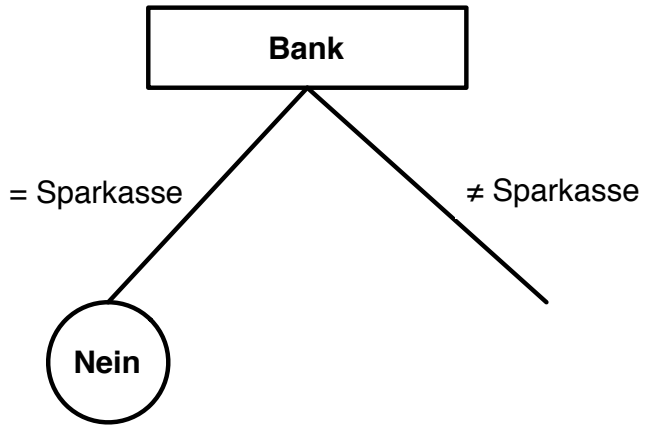
Bedingung	**erfüllt**				**nicht erfüllt**				**$GINI_{split}$**
	$\sum$	Ja	Nein	GINI	$\sum$	Ja	Nein	GINI	
Einzahlung = *'Nein'*	5	2	3	0,48	2	0	2	0	0,34
Bank = *'CBN'*	1	1	0	0	6	1	5	0,27	0,23
Bank = *'Sparkasse'*	4	0	4	0	3	2	1	0,44	0,19
Bank = *'Volksbank'*	2	1	1	0,5	5	1	4	0,32	0,37
Betrag $>$ 50€	6	2	4	0,44	1	0	1	0	0,38
Betrag $>$ 1.000€	5	2	3	0,48	2	0	2	0	0,34
Betrag $>$ 2.500€	4	2	2	0,5	3	0	3	0	0,28
Betrag $>$ 5.000€	3	2	1	0,44	4	1	3	0,375	0,40
Betrag $>$ 10.000€	2	1	1	0,5	5	1	4	0,32	0,37
Betrag $>$ 20.000€	1	1	0	0	6	1	5	0,277	0,24

$\Rightarrow$ Minimaler Wert für Aufteilung bei $0,19$ für Aufteilung nach dem Attribut **Bank**.

Resultierender Klassifikationsbaum

Leibniz Universität Hannover

Nach der ersten Aufteilung nach dem Attribut **Bank** ergibt sich der folgende Klassifikationsbaum:



Da alle Knoten, die für das Attribut **Bank** die Ausprägung *'Sparkasse'* besitzen, der gleichen Klasse (**Betrug** = *'Nein'*) angehören, wird als linkes Kind ein Blatt mit dem Label *'Nein'* erzeugt.

Bestimmung der zweiten Aufteilung

Leibniz Universität Hannover

Das Attribut **Bank** wurde bereits zur Aufteilung verwendet; Datensätze, die für das Attribut **Bank** die Ausprägung **Sparkasse** besitzen, also bereits klassifiziert wurden, erreichen diesen neuen Knoten nicht.

Einzahlung	Bank	Betrag	Betrug
Ja	Volksbank	5.000€	Ja
Nein	Central Bank of Nigeria	50.000€	Ja
Nein	Volksbank	20.000€	Nein

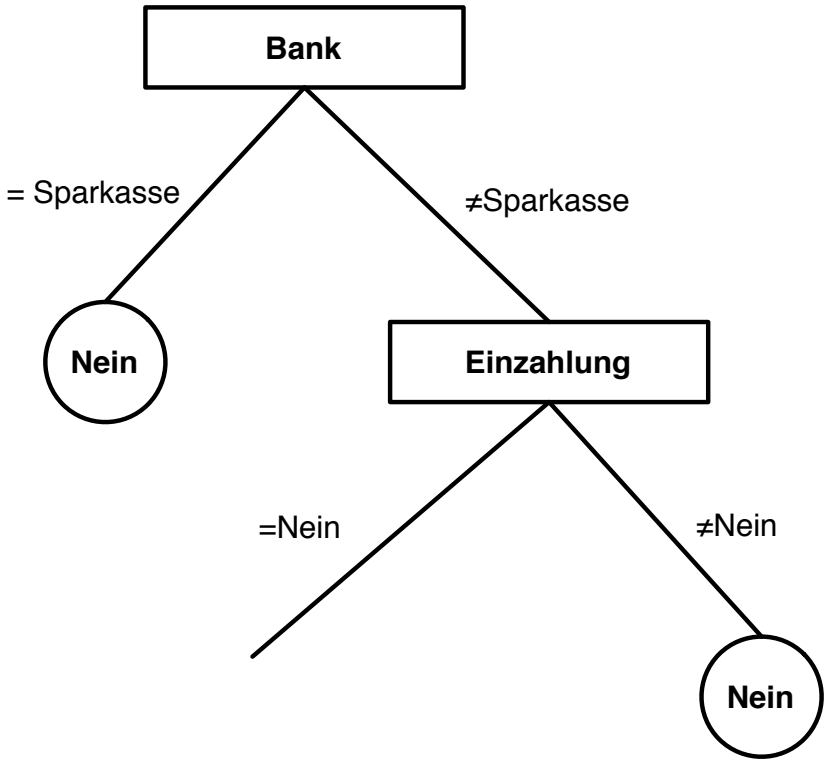
Im Folgenden werden nun die Split-Werte für die beiden verbliebenen potentiellen Teilungsattribute **Einzahlung** und **Betrag** berechnet.

Bedingung	**erfüllt**				**nicht erfüllt**				**$GINI_{split}$**
	$\sum$	Ja	Nein	GINI	$\sum$	Ja	Nein	GINI	
Einzahlung = *'Nein'*	2	1	1	0,5	1	1	0	0	0,66
Betrag > 5.000€	2	1	1	0,5	1	1	0	0	0,66
Betrag > 20.000€	1	1	0	0	2	1	1	0,5	0,66

⇒Da der Split-Wert für alle möglichen Bedingungen gleich ist, kann eine beliebige Bedingung gewählt werden (hier gewählt: Einzahlung = *'Nein'*).

Resultierender Klassifikationsbaum

Nach der Aufteilung nach dem Attribut **Bank** gefolgt vom Attribut **Einzahlung** ergibt sich folgender Klassifikationsbaum:



Für diese Aufteilung gilt, dass alle Datensätze, die für das Attribut **Einzahlung** die Ausprägung $\neq$ *'Nein'* haben, der selben **Betrugs**-Klasse angehören; folglich wird als Kind ein Blatt mit dem Label *'Nein'* erzeugt.

Berechung der dritten Aufteilung

Leibniz Universität Hannover

Wie zuvor werden nur noch die Datensätze aufgeführt, die den neuen Knoten erreichen.

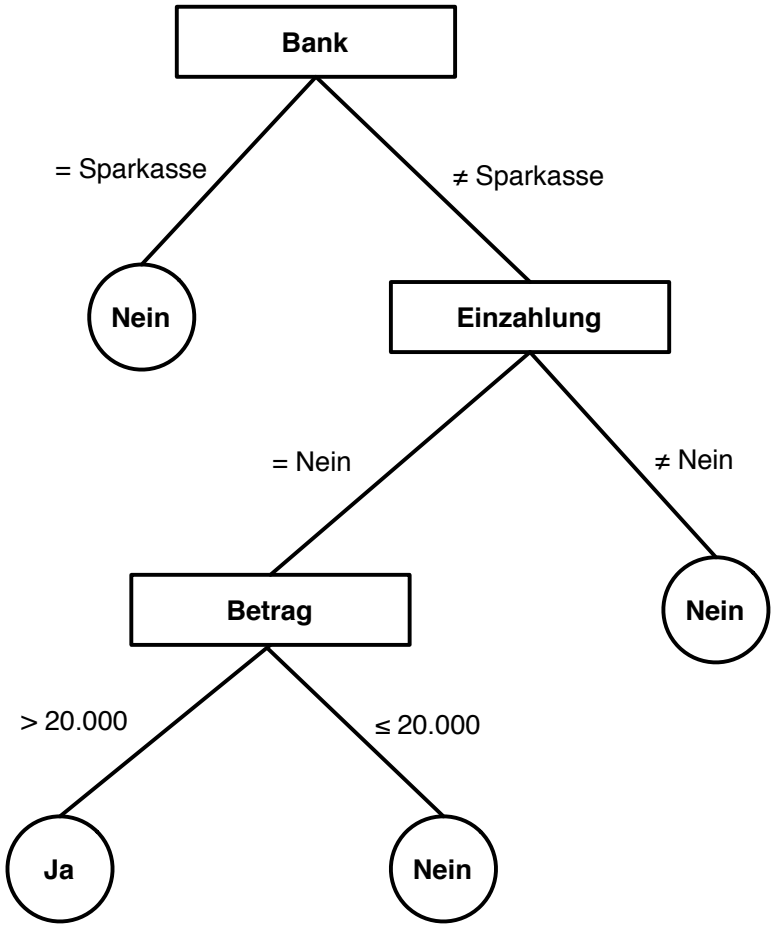
Einzahlung	Bank	Betrag	Betrug
Nein	Central Bank of Nigeria	50.000€	Ja
Nein	Volksbank	20.000€	Nein

Da nur noch zwei Datensätze verbleiben, gibt es nur eine Bedingung, anhand derer nach dem Attribut **Betrag** eine Aufteilung durchgeführt werden kann.

Bedingung	**erfüllt**				**nicht erfüllt**				**$GINI_{split}$**
	$\sum$	Ja	Nein	GINI	$\sum$	Ja	Nein	GINI	
Betrag $>$ 20.000€	1	1	0	0	1	0	1	0	0

Somit ergibt sich die finale Aufteilung durch das Attribut **Betrag** durch die Bedingung '*Betrag* $>$ 20.000€'.

Resultierender Klassifikationsbaum



Die beiden verbliebenen Datensätze gehören jeweils einer verschiedenen Klasse für das Attribut **Betrug** an. Somit gehören in der Folge an den neuen Knoten alle Datensätze der gleiche Klasse an, so dass Knoten mit dem jeweiligen Klassenlaben erzeugt werden.

Danke für die Aufmerksamkeit! Fragen?

Leibniz Universität Hannover

1. Überblick Klassifikation

2. Beispielrechnung mit dem GINI-Index