Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Software Engineering I

3. Semester
(20.08.2012 – 09.11.2012)

Kursprojekt: übers Thema und Literatur dazu

Prof. Dr. Dagmar Monett Díaz
Dagmar.Monett-Diaz@hwr-berlin.de

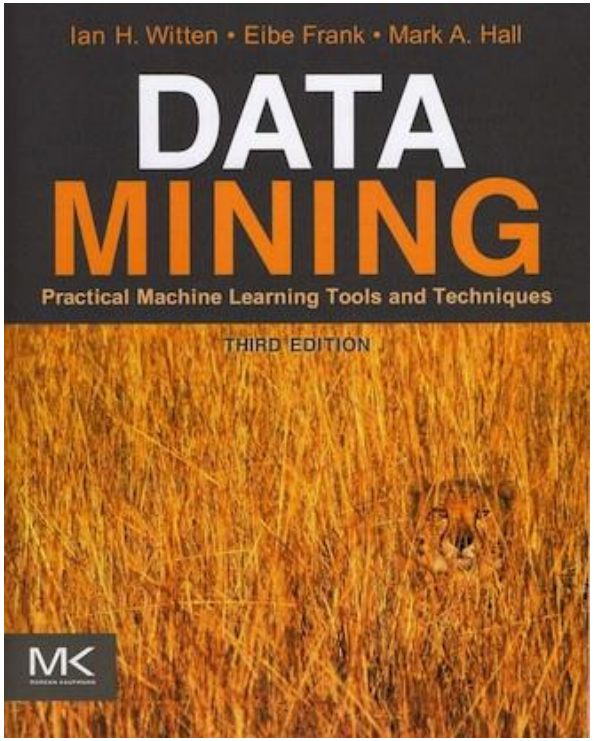
Agenda

- Hauptliteratur zum Thema Data Mining
- Aufgabenstellung
  - Erläuterungen
  - Entscheidungsbäume

Agenda

- Hauptliteratur zum Thema Data Mining
- Aufgabenstellung
  - Erläuterungen
  - Entscheidungsbäume

Hauptliteratur (i)



Data Mining: Practical Machine Learning Tools and Techniques

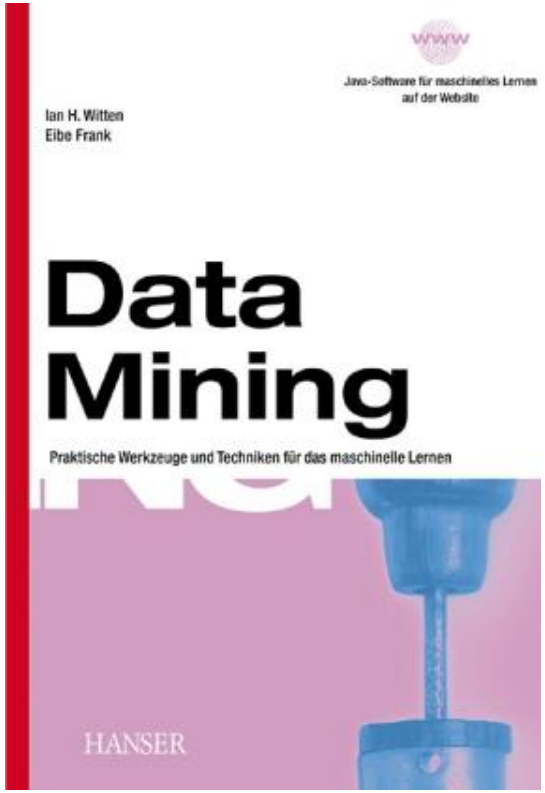
Ian H. Witten, Eibe Frank, Mark A. Hall

Taschenbuch: 629 Seiten, 3. Auflage
Verlag: Morgan Kaufmann, Januar 2011

Sprache: Englisch
ISBN-10: 978-0-12-374856-0

Website zum Buch:
http://www.cs.waikato.ac.nz/ml/weka/book.html

Hauptliteratur (ii)



Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen

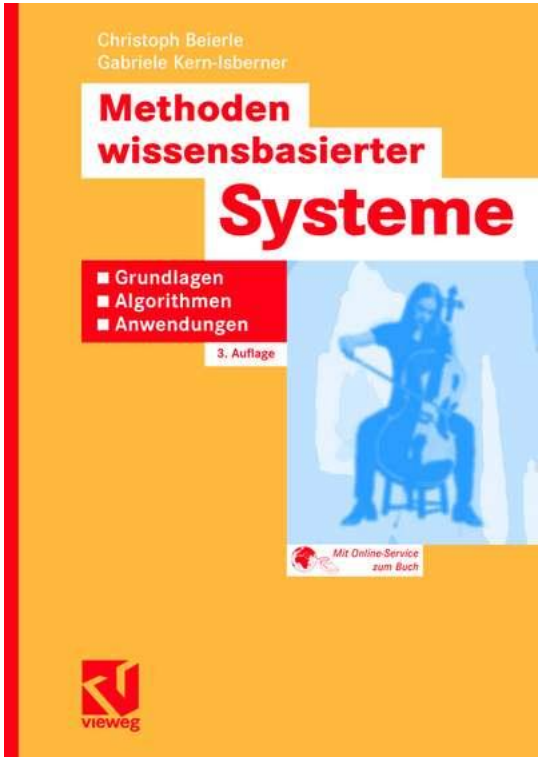
Ian H. Witten, Eibe Frank

Taschenbuch: 406 Seiten, 1. Auflage
Verlag: Hanser Fachbuch, Januar 2001

Sprache: Deutsch
ISBN-10: 978-3446215337

(Deutsche Übersetzung! -siehe auch vorige Folie)

Hauptliteratur (iii)



Methoden wissensbasierter Systeme. Grundlagen, Algorithmen, Anwendungen

Christoph Beierle, Gabriele Kern-Isberner

Taschenbuch: 490 Seiten
Verlag: Vieweg Friedr. + Sohn Ver;
Auflage: 3., erw. A. (April 2006)
Sprache: Deutsch
ISBN-10: 3834800104
ISBN-13: 978-3834805041

(4. Auflage, Juni 2008)

Hauptliteratur (iv)

Data Warehousing und Data Mining: Eine Einführung in entscheidungsunterstützende Systeme

Markus Lusti

Taschenbuch: 2., überarbeitete und erweiterte Auflage
Verlag: Springer, Oktober 2002

Sprache: Deutsch
ISBN-10: 3-540-42677-9

Agenda

- Hauptliteratur zum Thema Data Mining
- Aufgabenstellung
  - Erläuterungen
  - Entscheidungsbäume

Aufgabenstellung

Erstellung eines Programms, das den Algorithmus, einen „optimalen“ Entscheidungsbaum zu finden, demonstriert (★).

- **Zusammenarbeit Vorlesungen** SEW und Datenanalyse!
- **Fachübergreifende** Aufgabe!
- Thema aus anderem Fach **in SEW-I und II intensiviert**!
- **Selbständigkeit** und **Einarbeitung** in einer neuen Thematik!
- **Von Studierenden** vorgeschlagen worden!

★: Siehe auch Aufgabenblatt!

Agenda

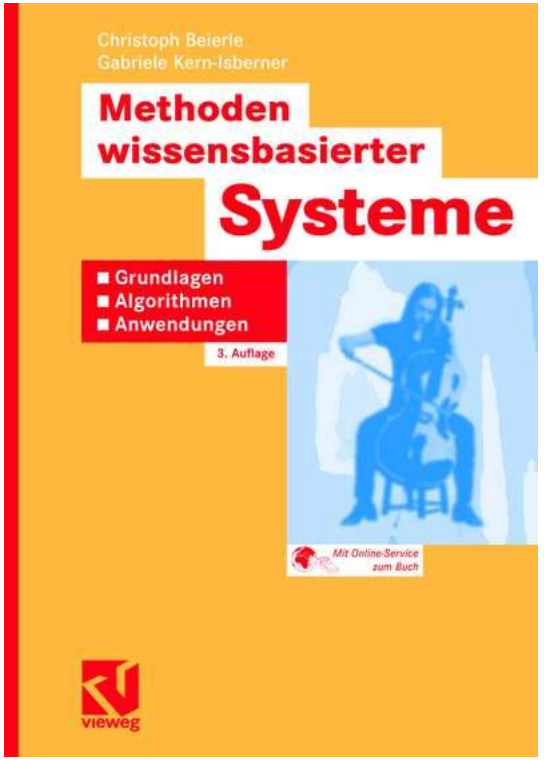
- Hauptliteratur zum Thema Data Mining
- Aufgabenstellung
  - Erläuterungen
  - Entscheidungsbäume

Entscheidungsbäume

In folgenden Folien:

- **Überblick** nach [Beierle & Kern-Isberner, 2006]

  Kapitel 5: Maschinelles Lernen
  5.3: Erlernen von Entscheidungsbäumen
  (Seiten 104 bis 118).



Mehr übers Thema:
- Siehe **weitere Literaturquellen**!
- **Vorlesung Datenanalyse**,
  4. Semester, Prof. Höhne!!

Entscheidungsbäume

- Entscheidungsbäume dienen der **Klassifikation von Objekten**, die durch Mengen von (**Attribut**, **Wert**)-Paare beschrieben sind
- **Entscheidung:** welcher Klasse das betreffende Objekt zuzuordnen ist
- Vereinfachung:
  - binäre Klassifikation, d.h. **Ja/Nein**-Entscheidung (kann leicht verallgemeinert werden)

Entscheidungsbäume: Struktur

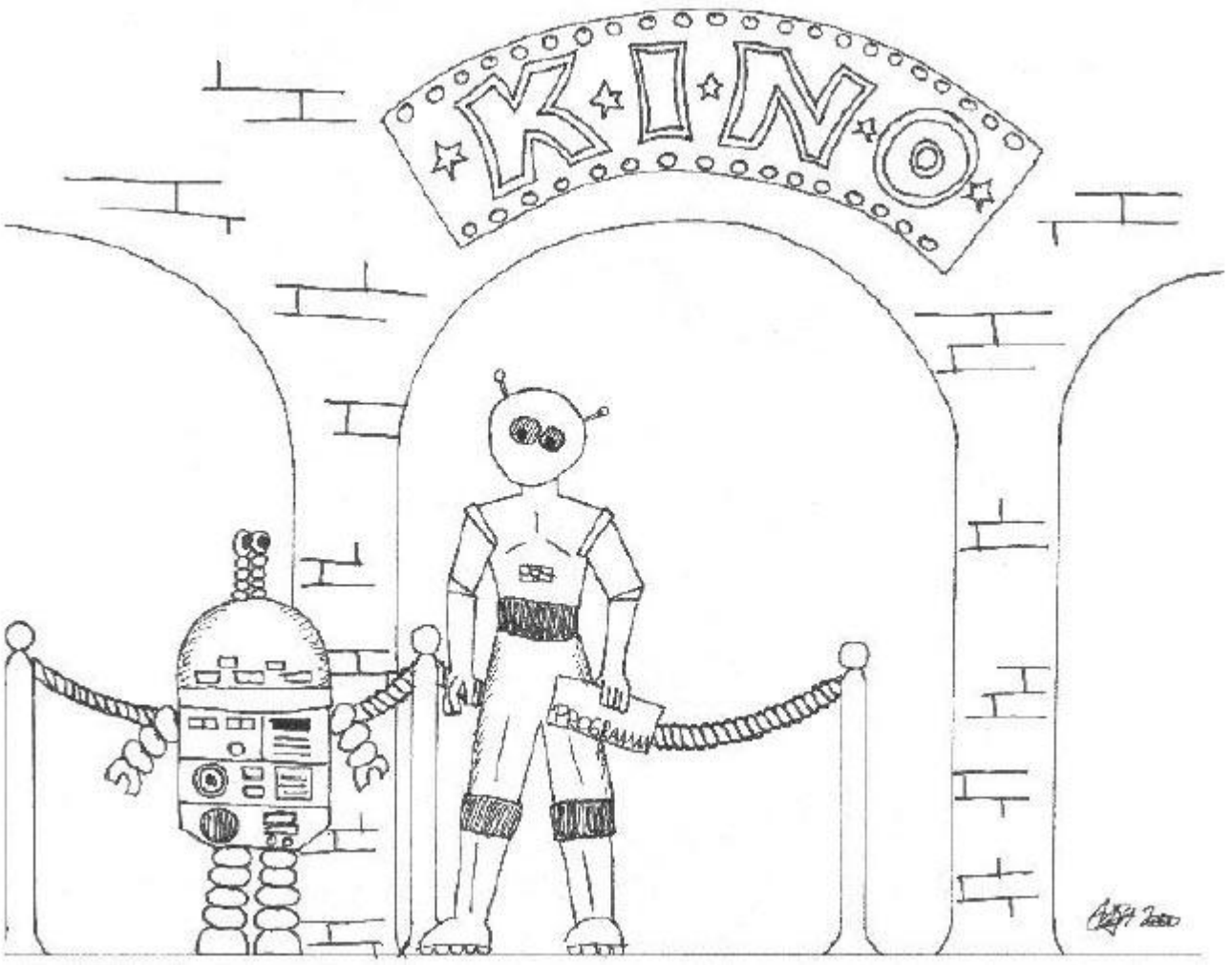
Ein solcher Entscheidungsbaum repräsentiert daher eine **boolesche Funktion**:

- **Wurzel und innere Knoten** des Baumes sind mit **Attributen** markiert und repräsentieren Abfragen, welchen Wert das betrachtete Objekt für das jeweilige Attribut (z.B. *attr*) hat
- Die von einem mit *attr* markierten Knoten **ausgehenden Kanten** sind mit den zu *attr* möglichen **Attributwerten** markiert
- Die **Blätter** sind mit dem Wahrheitswert markiert, der als Ergebnis der Funktion zurückgeliefert werden soll, wenn das Blatt erreicht wird. Sie enthalten die **Klassifikation**.

Objekte werden durch vollständige Pfade durch den Baum klassifiziert

Das Kino-Beispiel

Soll ich ins Kino gehen?



Bilder: [Beierle & Kern-Isberner, 2006]

Das Kino-Beispiel

Entscheidungssituation: *„Kino – ja oder nein?“*

zu klassifizierende **Objekte:** Situationen

relevante **Attribute:**

Ihrer Meinung nach, welche Eigenschaften bzw. ***Attribute*** *könnten eine* ***Situation*** *beschreiben?*

Das Kino-Beispiel

Entscheidungssituation: *„Kino – ja oder nein?“*

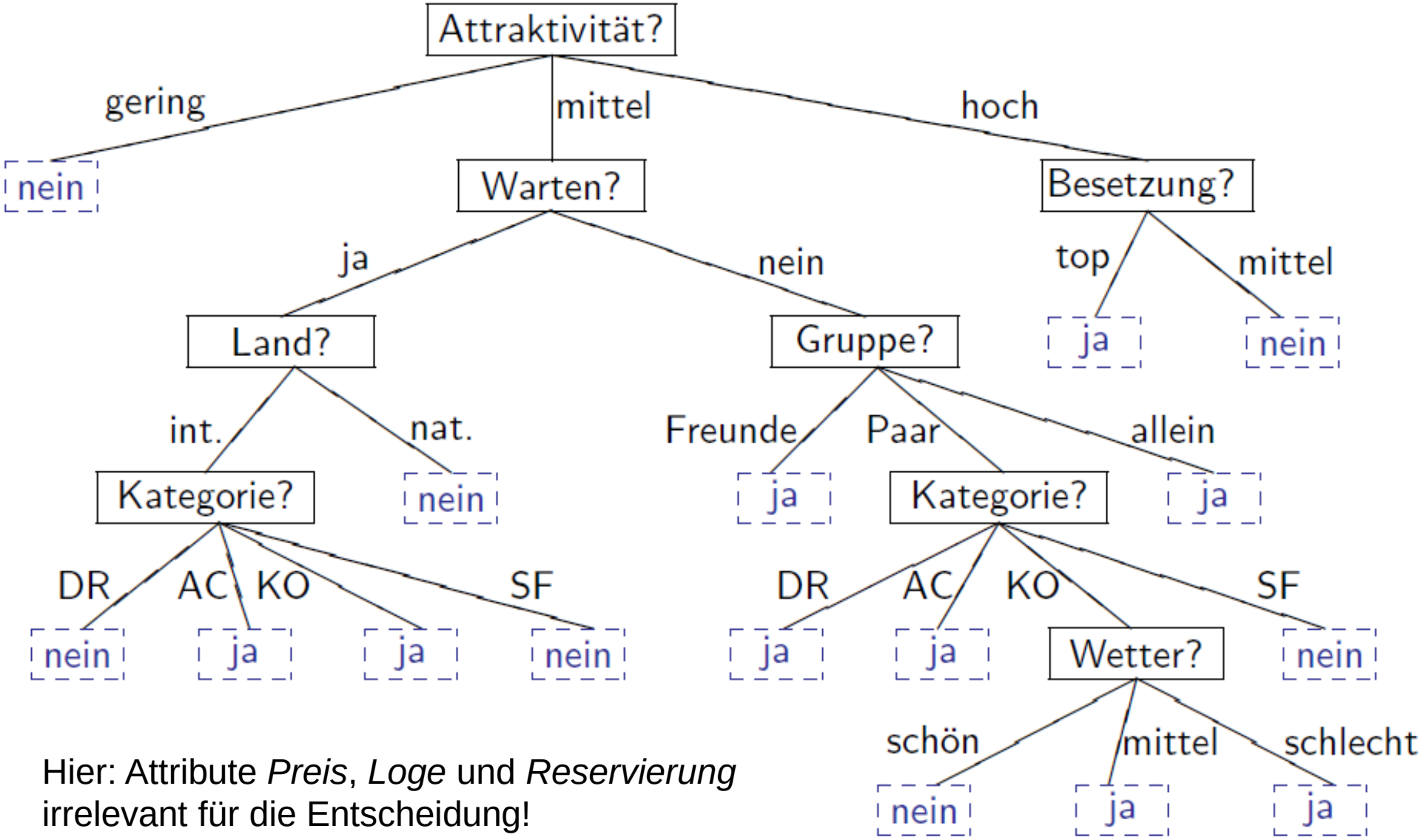
zu klassifizierende **Objekte:** Situationen

relevante **Attribute:**

Attribut	Werte
Attraktivität	*hoch*, *mittel*, *gering*
Preis	normal ($) oder mit Zuschlag ($$)
Loge	verfügbar (*ja*) oder nicht (*nein*)
Wetter	*schön*, *mittel*, *schlecht*
Warten	Wartezeit (*ja*) oder nicht (*nein*)
Besetzung	Cast und Regie sind *top*, *mittel*(mäßig)
Kategorie	*Action (AC)*, *Komödie (KO)*, *Drama (DR)*, *SciFi (SF)*
Reservierung	besteht (*ja*) oder nicht (*nein*)
Land	*nationale (N)* oder *internationale (I)* Produktion
Gruppe	mit *Freunde(n)*, als *Paar*, oder *allein*

Möglicher Entscheidungsbaum

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law



Hier: Attribute *Preis*, *Loge* und *Reservierung* irrelevant für die Entscheidung!

Generieren von Entscheidungsbäumen (i)

- **Lernverfahren:**

  aus einer Menge von Beispielen bzw. Datensätzen (genannt „Trainingsmänge“) einen Entscheidungsbaum zu generieren!

- Ein **Beispiel** besteht dabei aus einer Menge von Attribut/Wert-Paaren zusammen mit der Klassifikation

Beispiele für das Kinoproblem

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Acht positive und sieben negative **Beispiele**:

Beisp.	*Attr.*	*Preis*	*Loge*	*Wetter*	*Warten*	*Bes.*	*Kat.*	*Land*	*Res.*	*Gruppe*	*Kino?*
X_1	hoch	$$	ja	schlecht	ja	top	AC	int.	ja	Freunde	ja
X_2	mittel	$	ja	mittel	nein	mittel	KO	int.	nein	Paar	ja
X_3	mittel	$	nein	mittel	ja	mittel	DR	int.	nein	Freunde	nein
X_4	gering	$	ja	mittel	ja	mittel	SF	int.	nein	allein	nein
X_5	mittel	$	ja	mittel	nein	mittel	DR	int.	nein	Paar	ja
X_6	hoch	$$	ja	schön	nein	top	SF	int.	ja	Freunde	ja
X_7	mittel	$	ja	schlecht	nein	mittel	KO	nat.	nein	Freunde	ja
X_8	mittel	$	nein	schlecht	ja	mittel	AC	int.	nein	Freunde	ja
X_9	gering	$	ja	schön	nein	mittel	KO	nat.	nein	Freunde	nein
X_{10}	mittel	$	ja	schön	nein	mittel	KO	int.	nein	Paar	nein
X_{11}	hoch	$	ja	mittel	ja	top	DR	int.	nein	Paar	ja
X_{12}	mittel	$	nein	schlecht	ja	mittel	AC	nat.	nein	allein	nein
X_{13}	hoch	$$	ja	mittel	ja	mittel	SF	int.	nein	allein	nein
X_{14}	mittel	$	ja	schön	ja	top	DR	int.	ja	Freunde	nein
X_{15}	mittel	$	ja	schlecht	nein	mittel	AC	int.	nein	Paar	ja

Generieren von Entscheidungsbäumen (ii)

- **Trivialer Ansatz:**

  Man konstruiert einen Baum derart, dass für jedes Beispiel ein entsprechender Pfad von der Wurzel zu einem Knoten besteht.

- **Problem damit:**

  Wir können keine sinnvolle Generalisierung auf andere Fälle erwarten!

- Dann, **Lernaufgabe:**

  Erzeuge Entscheidungsbaum aus Trainingsmenge, so dass

  - Beispiele der Trainingsmenge korrekt klassifiziert werden und
  - sich der Entscheidungsbaum auch für andere Beispiele generalisieren lässt

→ **Induktives Lernen**

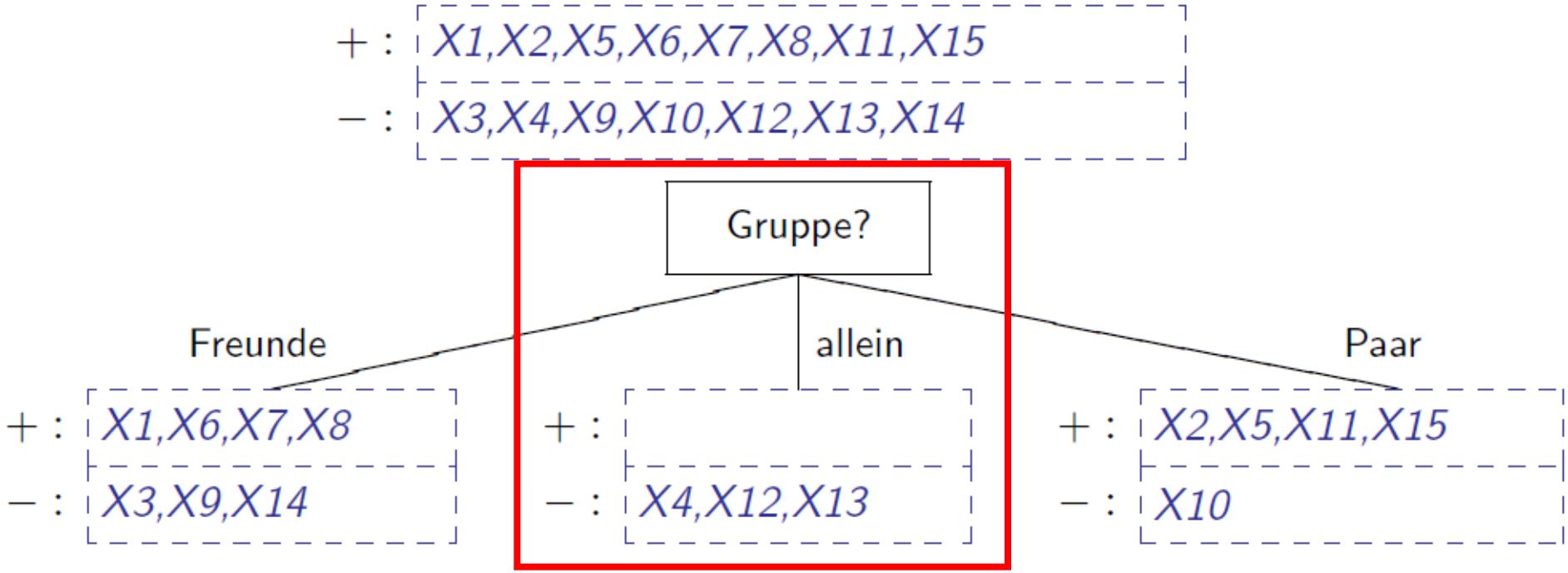
TDIDT Algorithmus

1. **Wähle** ein **Attribut** *attr* für den nächsten Knoten.
2. **Für jeden Wert** von *attr* **erzeuge** einen **Nachfolgeknoten**; markiere die zugehörige Kante mit diesem Wert.
3. **Verteile** die **aktuelle Trainingsmenge** auf die Nachfolgeknoten, entsprechend den jeweiligen Werten von *attr*.
4. Wende TDIDT auf die neuen Blattknoten an (**Rekursion**)

Das Kino-Beispiel (Forts.)

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Test für das Attribut *Gruppe*:



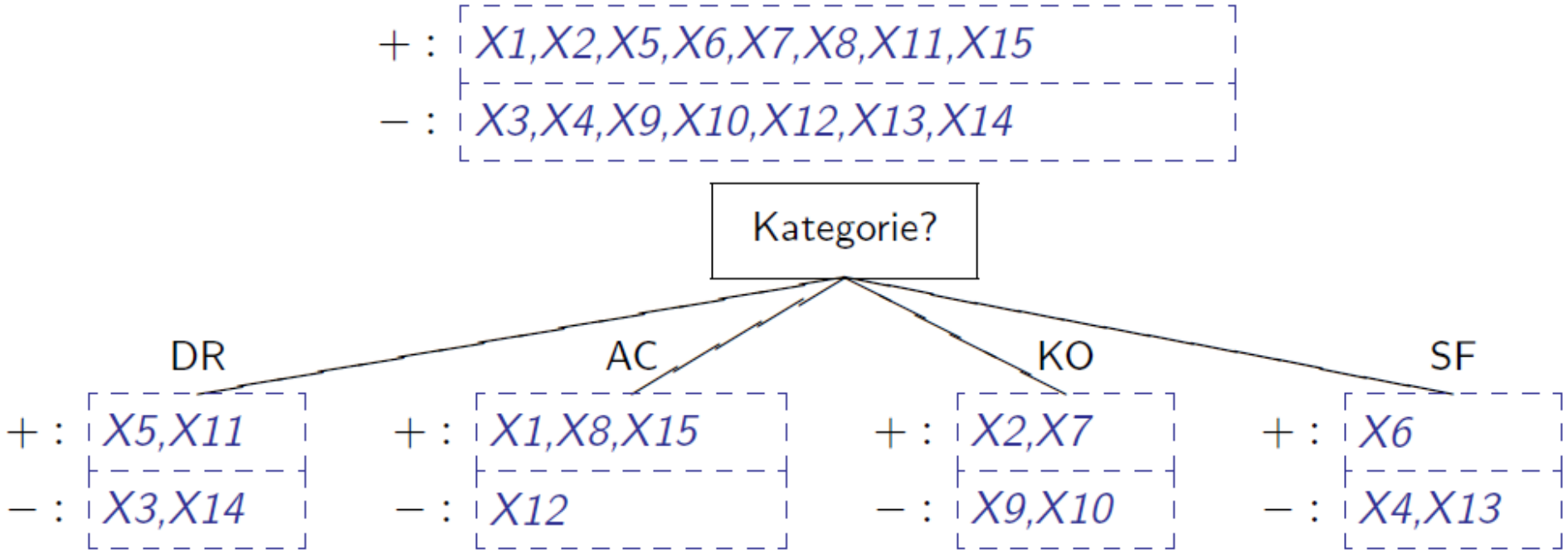
Beim Wert *Gruppe* = *allein* werden **alle verfügbaren (drei) Beispiele** vollständig klassifiziert.

Bei den Werten *Freunde* und *Paar* sind weitere Tests notwendig.

Das Kino-Beispiel (Forts.)

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Test für das Attribut *Kategorie*:

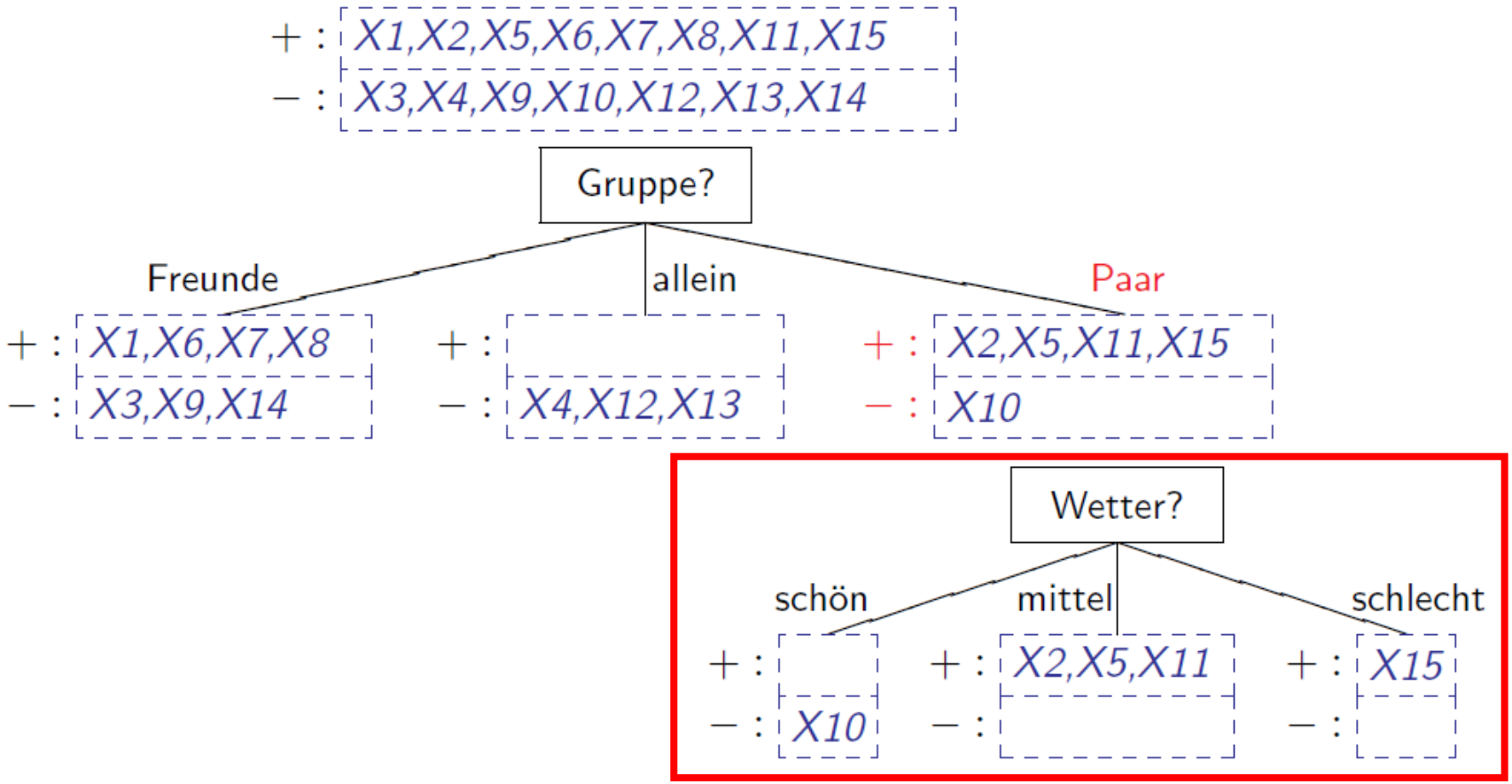


Das Attribut *Kategorie* kann **kein einziges Trainingsbeispiel** mit nur einem Test klassifizieren.

Gruppe ist also als erstes Attribut besser geeignet als *Kategorie* (klassifiziert **mehr** Beispiele).

Das Kino-Beispiel (Forts.)

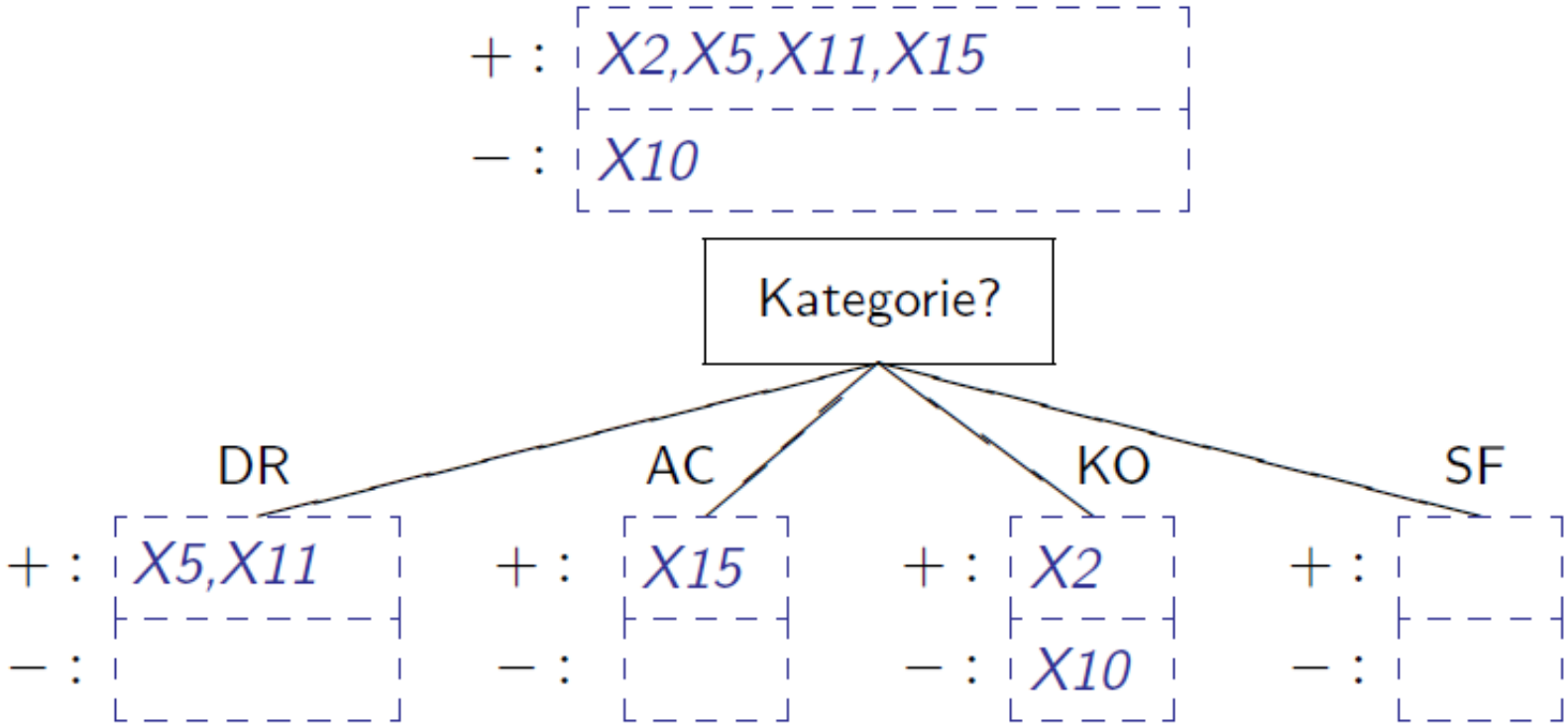
Gruppe als Erstes und dann *Wetter*, für *Gruppe* = *Paar*:



Wetter **klassifiziert alle übrige gebliebene Beispiele** der Menge *Gruppe* = *Paar* vollständig.

Das Kino-Beispiel (Forts.)

Gruppe als Erstes und dann *Kategorie*, für *Gruppe* = *Paar*:



Kategorie kann zwei Beispiele (*X2* und *X10*) **nicht eindeutig** klassifizieren.

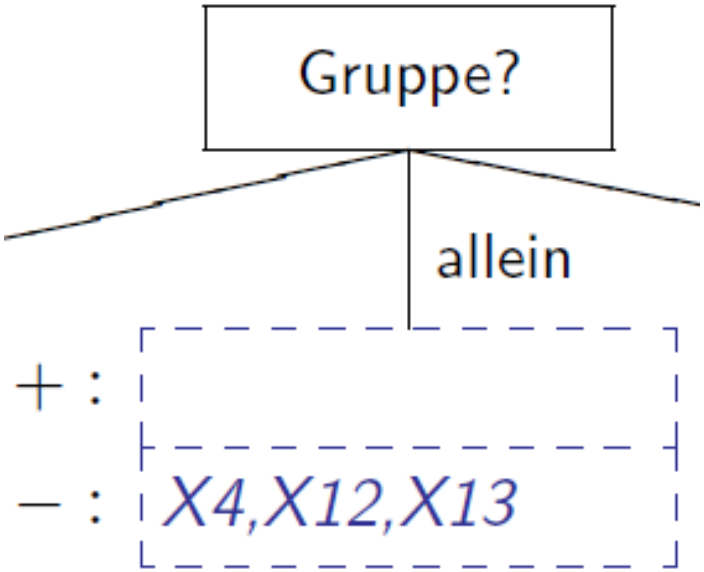
D.h. *Wetter* als zweites Attribut an dieser Stelle besser geeignet.

Fälle für die rekursiven Lernprobleminstanzen (i)

An den (aktuellen) Blattknoten können **vier verschiedene Fälle** auftreten:

1. Alle Beispiele haben die gleiche Klassifikation C

   → **Blatt mit Klassifikation C**



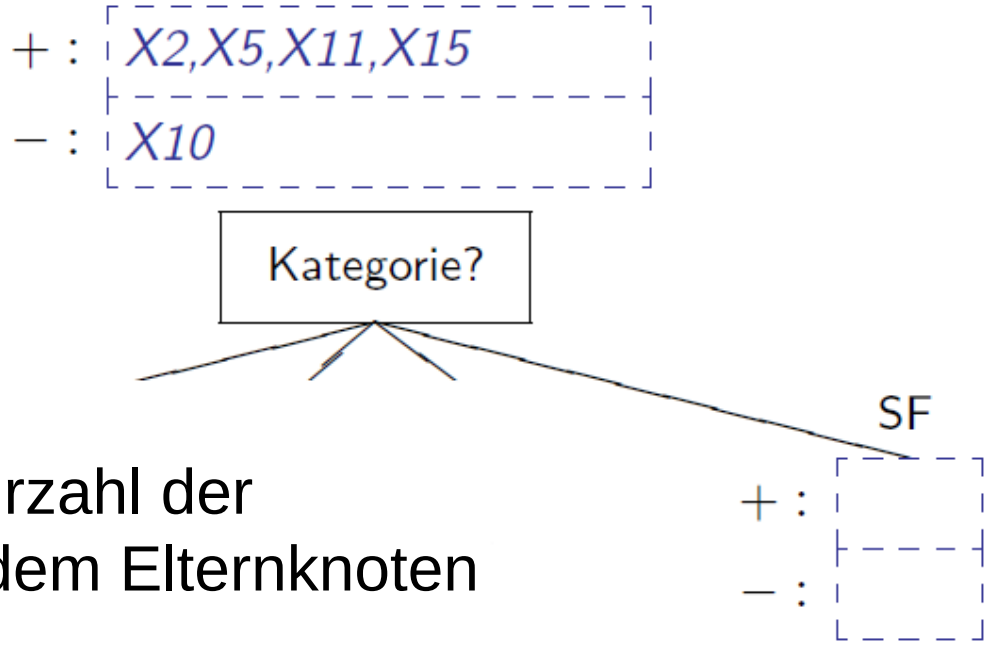
Fälle für die rekursiven Lernprobleminstanzen (ii)

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

An den (aktuellen) Blattknoten können **vier verschiedene Fälle** auftreten:

2. Beispielmenge ist leer → **Blatt mit Default-Klassifikation**

- Kein Beispiel mit entsprechenden Attribut-Wert-Kombination vorhanden



- Bsp. für Default-Klassifikation: Wert der Mehrzahl der klassifizierten Beispiele an dem Elternknoten

- Wenn gleich: positive Klassifikation liefern

Fälle für die rekursiven Lernprobleminstanzen (iii)

An den (aktuellen) Blattknoten können **vier verschiedene Fälle** auftreten:

3. Es gibt noch positive und negative Beispiele, aber es sind keine Attribute mehr übrig → **Inkonsistenz** (es gibt Beispiele mit genau denselben Attributwerten, aber unterschiedlicher Klassifikation)

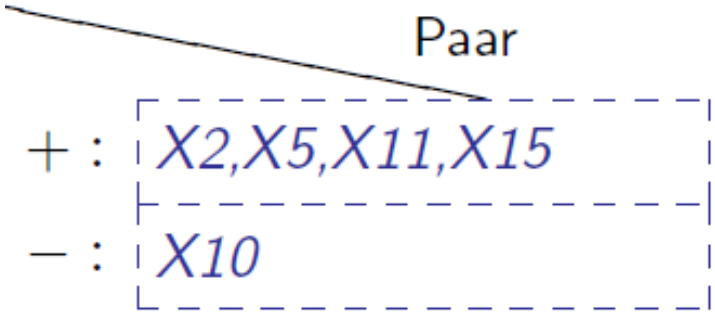
   - Könnte bedeuten: einige Beispiele sind falsch
   - aber auch: zusätzliche Attribute müssen eingeführt werden um Beispiele zu unterscheiden und damit die Situation vollständiger beschreiben zu können.
   - Annahme für TDIDT: Abbruch mit Fehlermeldung

Fälle für die rekursiven Lernprobleminstanzen (iv)

An den (aktuellen) Blattknoten können **vier verschiedene Fälle** auftreten:

4. es gibt noch positive und negative Beispiele, die aktuelle Menge der Attribute ist nicht leer → **nächster Rekursionsschritt**

   - Wähle bestes Attribut gemäß seiner „Wichtigkeit“ aus



Wichtigkeit von Attributen

Zentrales Problem:

- Wie findet man das (jeweils nächste) **beste Attribut**, um den Entscheidungsbaum aufzubauen?
- Die **Wichtigkeit** eines Attributes ist jedoch ein **relativer Begriff**. Sie hängt stark von der aktuellen Beispielmenge ab!

Welches Attribut a soll als nächstes gewählt werden?

- Wähle dasjenige Attribut, das **am wichtigsten** ist, d.h. das
  - soviel Beispiele wie möglich klassifiziert ($\rightarrow$ **Kardinalitätskriterium**);
  - die **meiste Information** enthält

Vollständiger Lernalgorithmus

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

… zum Generieren von Entscheidungsbäumen

function	*DT(E; A; default)*	
Eingabe:	E	Menge von Beispielen
	A	Menge von Attributen
	default	Default-Klassikation
Ausgabe:	Entscheidungsbaum	

Vollständiger Lernalgorithmus

… zum Generieren von Entscheidungsbäumen

- Verwenden Sie das **Konzept der Entropie** (siehe Aufgabenblatt) um den besten Entscheidungsbaum zu bestimmen!

- So können Sie den **Informationsgehalt** eines Attributes, der durch den jeweiligen **Informationsgewinn** bestimmt wird, berechnen.

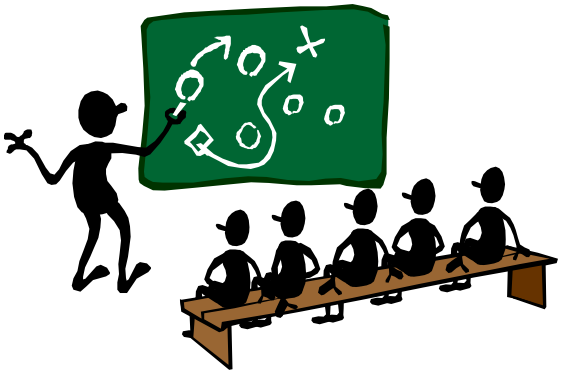
- Und **recherchieren** Sie weiter in dem Thema (siehe weitere Literaturempfehlungen sowie die Inhalte der Vorlesung Datenanalyse)!

Gruppenarbeit

Wissenstransfer, erste Überlegungen zum Projekt

1. Diskussion in Gruppen

2. Präsentation im Plenum



3. Diskussion im Plenum


GO!

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Software Engineering I

3. Semester

(20.08.2012 – 09.11.2012)

Prof. Dr. Dagmar Monett Díaz

Dagmar.Monett-Diaz@hwr-berlin.de