h_da

HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

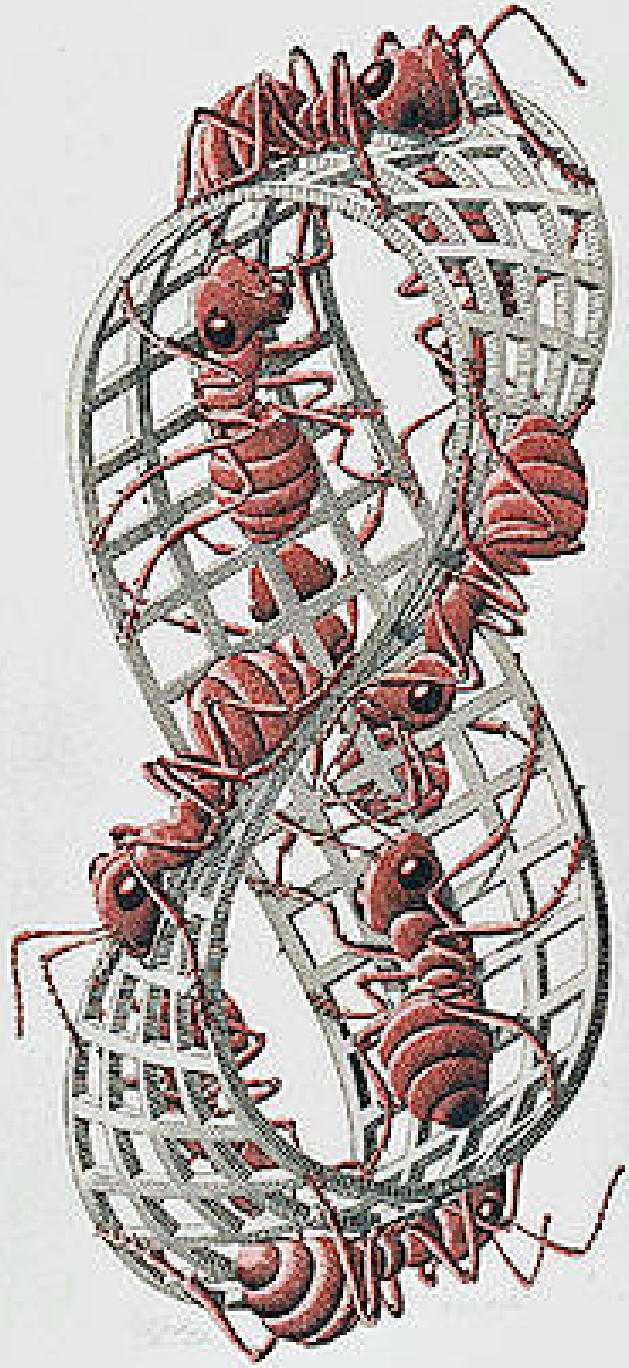
Wissensbasierte Diagnostik

Kap.3: Die Entscheidungsbaumtechnik Teil 1

Dr. Norbert Waleschkowski

c/o Semantis Information Builders GmbH
www.semantis-ib.de

Vorlesung Master-Studiengang
Wintersemester 2009/10



M.C. Escher: Möbiusband II (1963)

Kap. 3.1: Entscheidungsbäume

Ein Klassifikationsproblem (1)

- Gegeben seine 10 Datensätze, in denen Länge, Höhe, Form und Typ von Fahrzeugen zusammengestellt sind.

Nr.	Länge > 5 m	Höhe < 1,5m	Form	Fahrzeugtyp
1	nein	ja	A	O-Typ
2	unbekannt	nein	A	O-Typ
3	unbekannt	ja	B	M-Typ
4	nein	ja	B	O-Typ
5	ja	ja	B	M-Typ
6	ja	ja	A	M-Typ
7	unbekannt	ja	A	M-Typ
8	ja	nein	A	M-Typ
9	unbekannt	nein	B	O-Typ
10	nein	nein	A	O-Typ

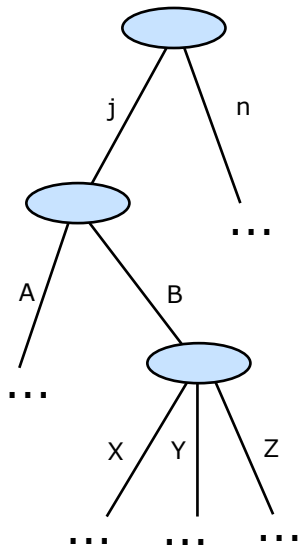
Anmerkung: Die Daten in diesem Beispiel sind völig willkürlich gewählt und entsprechen nicht der Realität.

- Um welchen Fahrzeugtyp handelt es sich beim Datensatz 11 ?

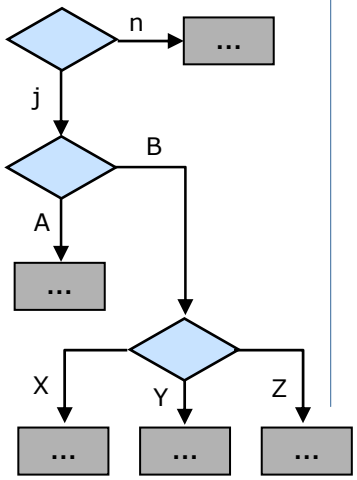
11	ja	nein	B	???

Ein Klassifikationsproblem (2)

- Bei dieser Aufgabe handelt es sich um ein Klassifikations-problem.
- Ein Klassifikationsproblem hat große Ähnlichkeit mit einem Diagnoseproblem.
  - Die Menge der bekannten Daten kann man als Symptom- oder Fehlerbild bezeichnen.
  - Die Daten gewinnt man aufgrund von Fragen bzw. von Tests.
  - Die Schlußfolgerung ist die Diagnose.
- Wie kommt man nun zu einer Diagnose?
- Dieses Problem läßt sich mittels eines Entscheidungsbaumes lösen.



bzw.



Ein Klassifikationsproblem (3)

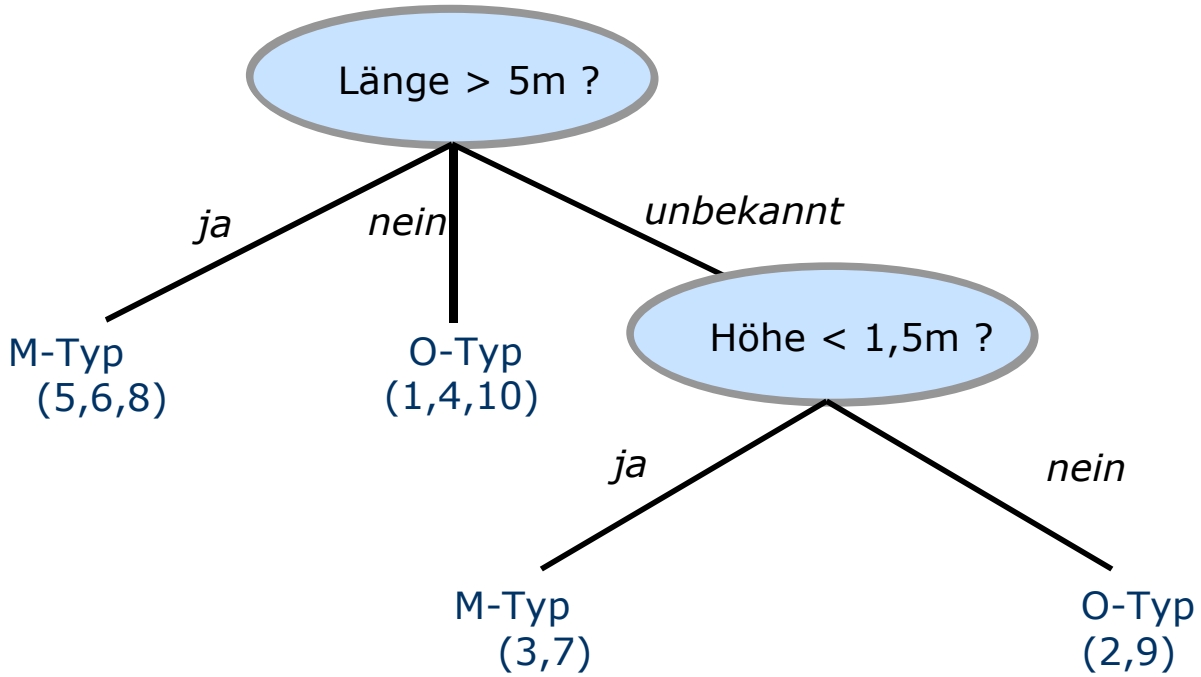
- Ein Entscheidungsbaum ist ein geordneter, gerichteter Baum im Sinne der Graphentheorie.
- Jeder Knoten des Baumes mit Ausnahme der Blätter enthält seinerseits ein Entscheidungsproblem, das Teil des Gesamtproblems ist. Diese Knoten heißen Entscheidungsknoten. Die Blätter nennt man auch terminale Knoten.
- Die Kanten zwischen einem Entscheidungsknoten und seinen Söhnen repräsentieren mögliche Entscheidungsalternativen für das zugehörige Entscheidungsproblem.
- Die Blätter enthalten die Ergebnisse der Entscheidungsfolgen. Diese Ergebnisse können gewisse Situationen - z.B. eine Diagnose - oder auszuführende Aktionen sein.
- Die Auswertung eines Entscheidungsproblems beginnt bei der Wurzel. Solange noch kein Blatt erreicht ist, löst man das Entscheidungsproblem des gerade betrachteten Knotens durch Wahl einer der möglichen Alternativen und verzweigt zum entsprechenden Sohn.

Ein Klassifikationsproblem (4)

- Ein Entscheidungsbaum für das oben dargestellte Problem besteht aus Knoten (Ellipse bzw. Raute), welche die Fragen bzw. Tests darstellen, und aus Kanten, welche die Antworten bzw. Testergebnisse repräsentieren.
- Ein Blatt bzw. eine Endstelle bzw. ein terminales Element des Baumes wird dargestellt durch eine Zeichenkette bzw. eine Box und ist das Ergebnis eines Entscheidungsprozesses bzw. die Diagnose.

Ein Klassifikationsproblem (5)

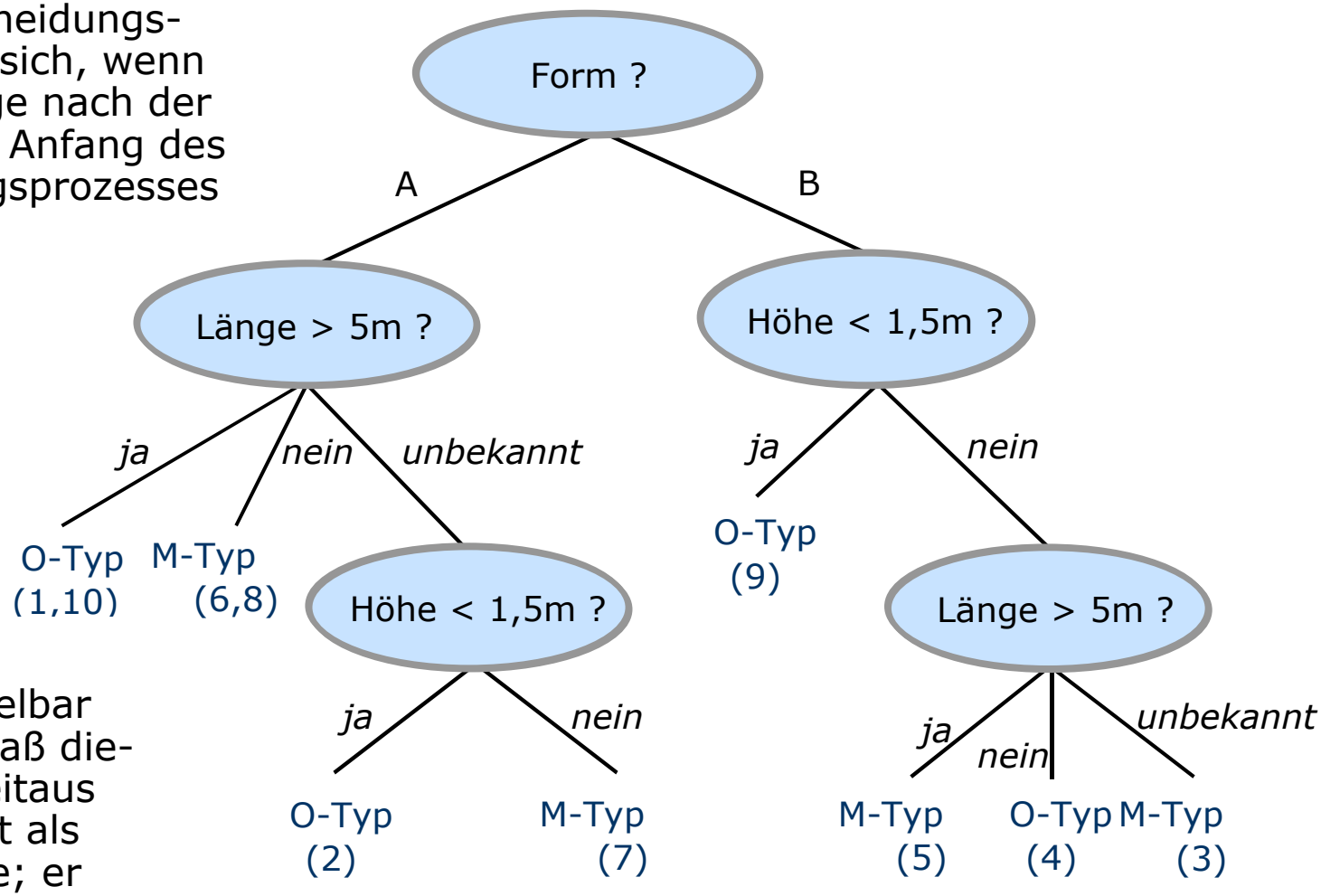
- Die Abbildung zeigt, wie man auf dem kürzesten Wege zur richtigen Entscheidung kommt. Dieser Baum ist der optimale Entscheidungsbaum. (Die Zahlen in den Klammern geben die Satznummern der Sätze an, auf welchen der Entscheidungs-prozess beruht.)



Optimaler Entscheidungsbaum

Ein Klassifikationsproblem (6)

- Dieser Entscheidungs-baum ergibt sich, wenn man die Frage nach der Form an den Anfang des Entscheidungsprozesses stellt.



Nicht-optimaler Entscheidungsbaum

- Es ist unmittelbar erkennbar, daß die-ser Baum weitaus komplexer ist als der vorherige; er führt allerdings zu denselben Ergeb-nissen.

Ein Klassifikationsproblem (7)

- An diesem kleinen Beispiel erkennt man bereits einige grundsätzliche Probleme bei der Verwendung von Entscheidungsbäumen.
- Das eigentliche Kernproblem besteht darin, daß man in Abläufen und nicht in Ursachen denkt. Mit der Struktur eines Entscheidungsbaumes wird sofort der Ablauf des Entscheidungsprozesses festgelegt.
- Änderungen im Entscheidungsbaum sind i.d.R. "globale" Änderungen, d.h. sie haben Auswirkungen auf den gesamtem Entscheidungsbaum.
- Eine Struktur ist dagegen dann änderungsfreundlich, wenn Änderungen lokalen Charakter haben, d.h. wenn sich eine Änderung nicht auf die gesamte Struktur auswirkt bzw. wenn man die Änderung vornehmen kann, ohne die gesamte Struktur zu kennen.
- Nehmen wir etwa an, es soll stärker differenziert werden, z.B. nach M-Typ-Limousine und -Coupe. Hierzu muß man weitere Kriterien hinzuziehen. Es stellt sich die Frage, wie und wo diese neuen Kriterien im Baum untergebracht werden können.

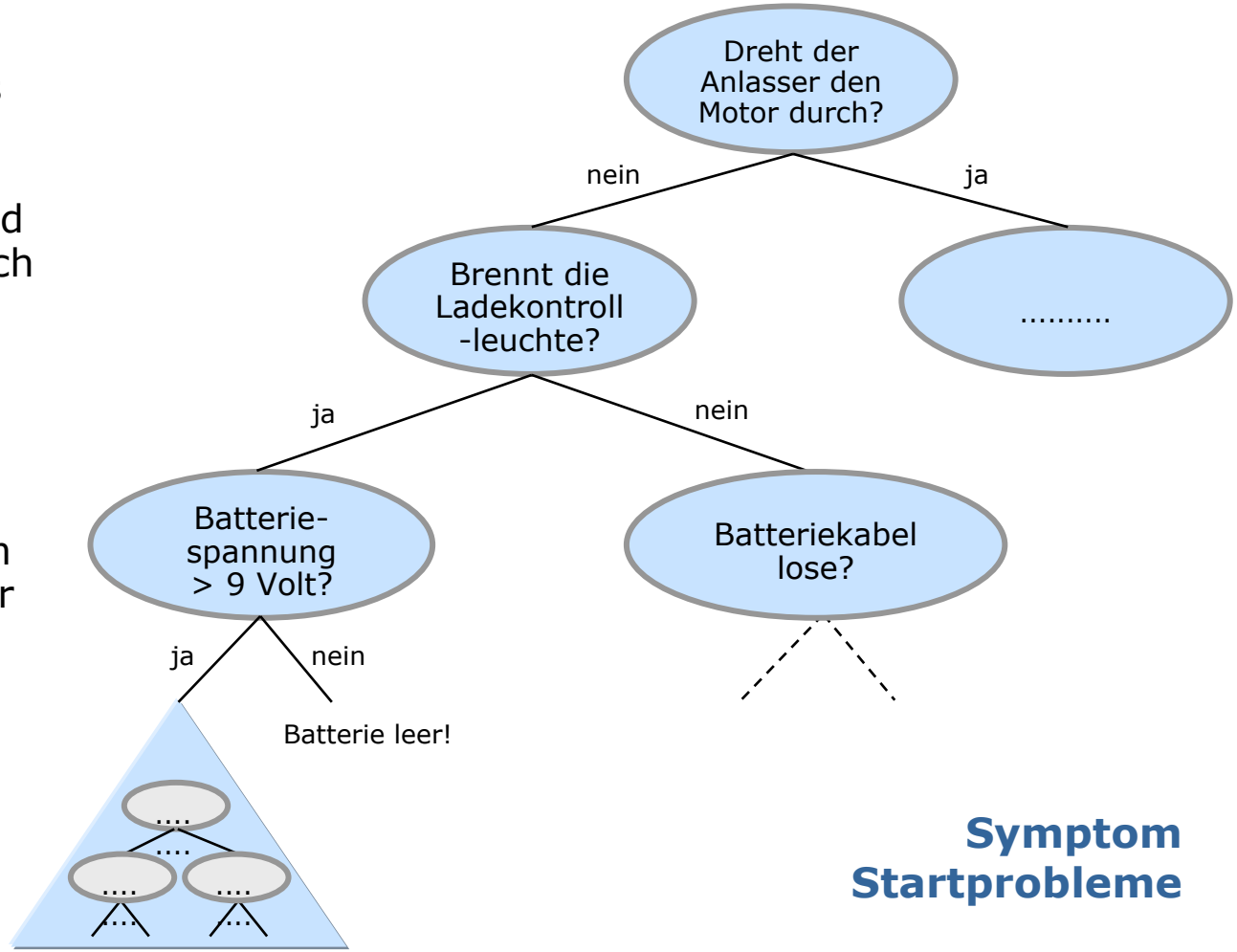
Ein Klassifikationsproblem (8)

- Um diese Informationen richtig einzuordnen, muss der gesamte Entscheidungsbaum inspiziert und meist umorganisiert werden. Es ist i.a. nicht, zumindest nicht ohne weiteres möglich, die Änderung lokal zu begrenzen.
- Obwohl das diskutierte Beispiel sehr klein ist, muss man dennoch schon etwas genauer hinschauen, um zu wohlstrukturierten Entscheidungsbäumen zu gelangen.
- Wie kommt man nun zu angemessenen Entscheidungsbäumen bei sehr großen Datenmengen?
- Es kann z.B. sehr leicht passieren, dass man mit einem gegebenen Entscheidungsbaum eine bestimmte "Diagnose" möglicherweise sehr effizient finden kann, in anderen Fällen dagegen nur eine geringe Effizienz erzielt.

Ein Beispielproblem zur Diagnose (1)

- Betrachten wir nun die Verwendung von Entscheidungsbäumen an einem kleinen Beispiel aus der Fahrzeugdiagnose.

- Das Beispiel ist stark vereinfacht und etwas konstruiert. Es dient lediglich zur besseren Veranschaulichung und erhebt keinen Anspruch auf sachliche Angemessenheit.
- Es soll untersucht werden, warum ein Motor nicht anspringt. Ein mögliches Problem ist eine schwache oder defekte Batterie.
- Die Abbildung zeigt einen kleinen Ausschnitt aus einem entsprechenden Entscheidungsbaum.



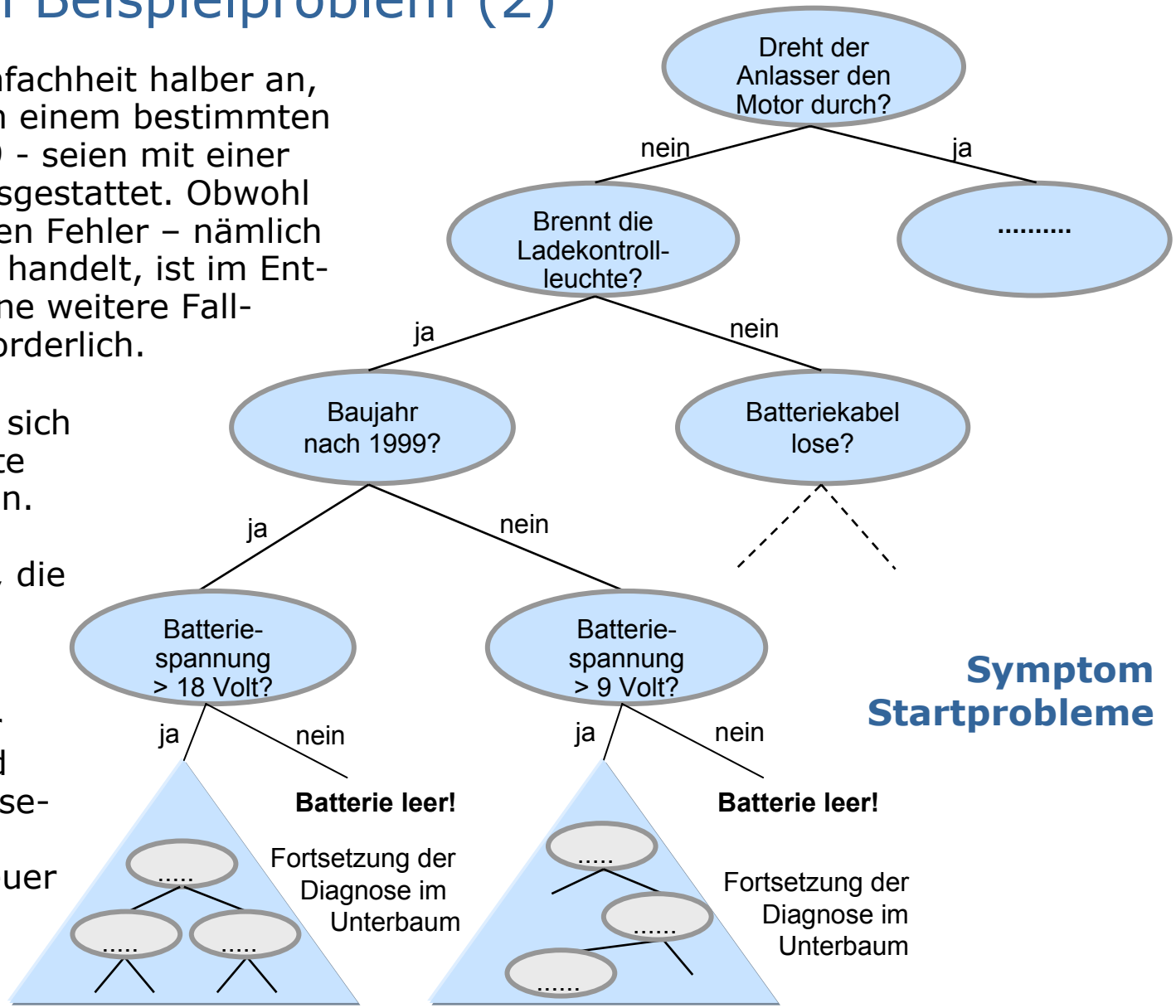
Symptom Startprobleme

Ein Beispielproblem (2)

- Nehmen wir der Einfachheit halber an, alle Fahrzeuge nach einem bestimmten Baujahr - z.B. 1999 - seien mit einer 24-Volt-Batterie ausgestattet. Obwohl es sich um denselben Fehler – nämlich Batterieprobleme – handelt, ist im Entscheidungsbaum eine weitere Fallunterscheidung erforderlich.

- Dieses Problem läßt sich noch durch geeignete Parametrierung lösen.

- Nun nehmen wir an, die Fehlerwahrscheinlichkeiten unterscheiden sich im Sommer und Winter erheblich. Dann sind ganz andere Diagnoseabläufe erforderlich. Es muß ein völlig neuer Entscheidungsbaum erstellt werden.



Das zentrale Entscheidungsbaum-Problem (1)

- Die bisherigen Beispiele sowie einfache Überlegungen zeigen typische Probleme im Zusammenhang mit Entscheidungsbäumen auf:

1. Es ist i.a. à priori nicht bekannt, welche Frage bzw. welcher Test am Anfang des Entscheidungsprozesses stehen sollte.
2. Stellt man fest, dass ein Entscheidungsbaum nicht effizient aufgebaut ist, muss der neue Entscheidungsbaum von Grund auf neu erstellt werden.
3. Neue Fragen bzw. Tests lassen sich nachträglich kaum einbauen, ohne den gesamten Baum zu inspizieren.
4. Struktur und Ablauf sind untrennbar miteinander verwoben, d.h., mit der Struktur des Baumes ist gleich der gesamte Ablauf festgelegt.
5. Betrachtet man einen Entscheidungsbaum, den man nicht kennt bzw. nicht selbst erstellt hat, so hat man große Schwierigkeiten, den Fokus des Systems zu erkennen. (Man weiß z.B. nicht, welche Fehlerhypothese gerade untersucht wird. Der Techniker in der Werkstatt fragt sich, worauf die Untersuchung eigentlich hinauslaufen soll.)

Das zentrale Entscheidungsbaum-Problem (2)

6. Es gibt keine Möglichkeit, einen Entscheidungsbaum auf Vollständigkeit zu untersuchen.

7. Ein großer Entscheidungsbaum ist nur schwer wartbar. (Bsp.: Beim Symptom "*Startschwierigkeiten*" möchte man im Sommer zunächst den Fehler *"Defekter Anlasser",* im Winter den Fehler *"Defekte Batterie"* zuerst untersuchen. Dieser Sachverhalt lässt sich nicht abbilden, ohne den Unterbaum zu duplizieren und umzuordnen. Die Unterbäume sind nur spezifisch verwendbar, da man das Problem impliziter Annahmen hat, nämlich im jeweiligen Unterbaum die Annahme, es herrsche Sommer bzw. Winter.)

8. Ein Entscheidungsbaum entspricht aus der Sicht der Informatik einer Struktur, die aus lauter *if-then-else* und *goto*-Konstrukten besteht.

9. Ein Entscheidungsbaum ist nicht ohne weiteres objektivierbar. Zwei Experten kommen zu unterschiedlichen Baumstrukturen.

- Das zentrale Entscheidungsbaum-Problem

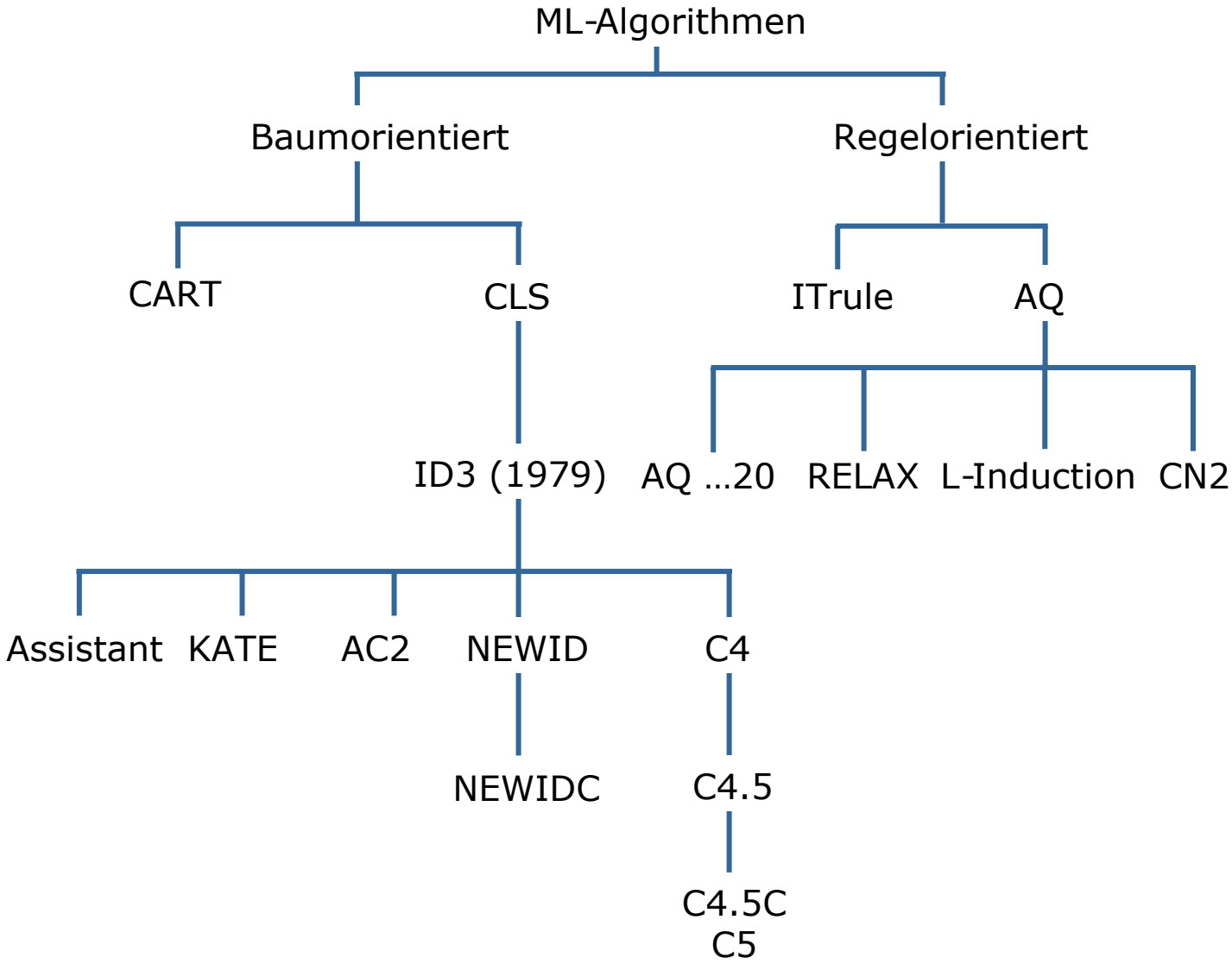
  Das Kernproblem liegt darin, dass mit einem Entscheidungsbaum gleichzeitig auch der Ablauf der Entscheidungsfindung festgelegt wird.

Kap. 3.2: Entscheidungsbaum-Lernalgorithmen

Erstellung effizienter Entscheidungsbäume durch ML-Algorithmen (1)

- Ohne maschinelle Unterstützung ist das Aufstellen effizienter Entscheidungsbäume sehr schwierig.
- Es gibt eine Reihe von Algorithmen, die hier Unterstützung liefern können.
- Die folgende Abbildung zeigt die wichtigsten dieser zum Gebiet des Maschinellen Lernens (Machine Learning = ML) gehörenden Algorithmen.
- Der Klassiker dieser Algorithmen ist der *ID3*-Algorithmus von Quinlan [Quinlan 1979].
- Ausgehend von *ID3* wurde eine Reihe von Weiterentwicklungen und Verbesserungen (z.B. *NEWID* und *C4.5* sowie Varianten wie *NEWIDC* und *C4.5C*) vorgenommen, die im folgenden kurz behandelt werden.

Erstellung effizienter Entscheidungsbäume durch ML-Algorithmen (2)



Entscheidungsbaum-Lernen

- In jedem Knoten eines Entscheidungsbaumes steckt ein Verzweigungs-kriterium für ein Attribut.
- Beim Entscheidungsbaum zum Klassifikationsproblem auf S.2 sind dies die Attribute „Form", „Länge" und „Höhe". Das Klassifikationsziel ist der Fahrzeugtyp; seine Ausprägungen heißen Klassen.
- Um einen effizienten Entscheidungsbaum aufzubauen, sollten die „wichtigeren" Attribute weiter oben im Baum angeordnet werden als die „unwichtigeren". Doch was heißt hier „wichtiger", und wie läßt sich feststellen, ob ein Attribut „wichtiger" ist als ein anders.
- Wir benötigen hierzu einige Grundlagen aus der Informationstheorie wie Informationsgehalt, Entropie, Transinformation etc.. Daher erfolgt zunächst ein kurzer Exkurs in die Informationstheorie.

Kap. 3.2.EX: Exkurs in die Informationstheorie

Begriff der Information (1)

- Der zentrale Begriff in der Informationstheorie ist der Informationsbegriff. Aber was ist eigentlich Information?
- Begriffe wie Signale, Nachrichten, Daten und Information werden häufig synonym gebraucht, obwohl sie voneinander abzugrenzen sind.

- **Signal**
  Ein Signal ist ein physikalisch wahrnehmbarer Tatbestand.
  Bspe.: ein Farbsignal (rote Ampel), ein Zeichen (Buchstabe), ein Ton.

- **Nachricht**
  Eine Nachricht ist eine Folge von Signalen. Hier liegt der Schwerpunkt auf dem Übertragungsvorgang. Dieser Begriff setzt also einen Sender und einen Empfänger der Nachricht voraus.

- **Datum**
  Ein Datum meint eher die zeitraumbezogene Speicherung von Zeichen.

- **Information**
  Daten werden i.a. als Information bezeichnet, wenn sie einen Nutzen für ein Entscheidungsproblem besitzen, d.h. wenn sie die Entscheidungen des Empfängers beeinflussen können.

Ein Laplace-Experiment ist ein Zufallsereignis, bei dem alle möglichen Ausgänge die gleiche Eintrittswahrscheinlichkeit besitzen.

Wir betrachten das einfachste Laplace-Experiment (ein Ereignis mit zwei Ausgängen) , z.B. das Werfen einer fairen Münze. Die beiden Ausgänge – Kopf oder Zahl – treten mit gleicher Wahrscheinlichkeit ½ auf. Diesem Ereignis wird nun von Shannon der Informationswert 1 bit zugeordnet.

Ein Zufallsereignis mit 4 (bzw. 8) möglichen Ausgängen mit je ¼ (bzw. 1/8) Wahrscheinlichkeit sollte dann einen Informationswert von 2 (bzw.3) bit besitzen, denn ein solches Ereignis kann aufgefasst werden wie zwei (bzw. drei) unabhängige Münzwürfe.

Ein Laplace-Ereignis mit n möglichen Ausgängen sollte dann einen Informationswert von $-\log_2(1/n)$ $[=\log_2(n)]$ besitzen.

Dieser Informationsbegriff lässt sich auch als die durchschnittliche Anzahl von Fragen interpretieren, die man stellen muss, um den Ausgang eines Zufallsexperiments herauszufinden. Beispiel: Wir suchen einen Namen in einem Telefonbuch mit 1024 Seiten. Nach durchschnittlich 10 Fragen $[\log_2(1/1024)=10]$ – jeweils Halbieren und Weitersuchen im linken bzw. rechten Teilbuch – haben wir die richtige Seite identifiziert.

Begriff der Information (2)

- Information und Nutzen der Information liegen im allgemeinen Sprachgebrauch nahe beieinander. Dieser Informationsbegriff wird daher auch als pragmatischer Informationsbegriff bezeichnet.
- Der im folgenden zugrundegelegte klassische Begriff der Information wirkt daher auf den ersten Blick abstrakt und unmotiviert.
- Der Ansatz von Shannon besteht darin, den Informationsgehalt I eines Ereignisses e an Hand seiner Eintrittswahrscheinlichkeit p zu messen.
- Grundgedanke: Ein Ereignis hat nur dann Informationscharakter, wenn sein Eintreten ungewiss ist. Dabei bleibt die Bedeutung des Ereignisses völlig unberücksichtigt.
- Ein sicheres Ereignis e hat danach keinen Informationswert:

$$p(e) = 1 \Rightarrow I = I(e) = 0$$

- Allgemein gilt für ein Ereignis e mit Eintrittswahrscheinlichkeit p(e)

$$I(e) = \log(1/p(e)) = -\log p(e)$$

Bem.: Im folgenden gelte stets **$\log = \log_2$.**

Basisumrechnung:
$\log_b r = \log_a r / \log_a b$

Ein motivierendes Beispiel (1)

Claude Elwood Shannon (30.4.1916 - 24.2.2001) gilt als einer der Begründer der Informationstheorie. In 1948 veröffentlichte er seine bahnbrechende Arbeit „A Mathematical Theory of Communication". Er war lange für die AT&T Bell Labs tätig und war von 1958-78 Professor am MIT (Cambridge, Mass.)

- Wir betrachten eine Zufallsvariable Z mit den möglichen Ausgängen A, B, C und D mit gleichen Eintrittswahrscheinlichkeiten ¼.*

p(Z=A) = 1/4	p(Z=B) = 1/4	p(Z=C) = 1/4	p(Z=D) = 1/4

- Dabei kann z.B. folgende Zeichenkette erscheinen:

  CBBACBDAADBABCC …

- Die Zeichen sollen über eine binäre serielle Verbindung übertragen werden. Dazu können wir z.B. folgende Kodierung verwenden:

  A=00, B=01, C=10, D=11

- Dann ergibt sich für die o.a. Zeichenkette

  1001010010011100001101000110 10….

* nach einem Beispiel von Andrew W. Moore von der Carnegie Mellon University (www.cs.cmu.edu/~awm).

Ein motivierendes Beispiel (2)

- Nun möge sich die Wahrscheinlichkeitsverteilung wie folgt ändern:

p(Z=A) = 1/2	p(Z=B) = 1/4	p(Z=C) = 1/8	p(Z=D) = 1/8

- Frage: Ist es möglich, die Kodierung so zu verändern, dass durchschnittlich nur noch 1,75 bits pro Zeichen benötigt werden?
- Die Antwort lautet: „Ja." Betrachte z.B. die folgende Kodierung:

A	0
B	10
C	110
D	111

- Dies ist nur eine von mehreren Möglichkeiten.

Ein motivierendes Beispiel (3)

- Wir nehmen nun an, Z hätte 3 gleichverteilte Ausgänge.

p(Z=A) = 1/3	p(Z=B) = 1/3	p(Z=C) = 1/3

- Die folgende Kodierung verwendet 2 bits je Zeichen.

A	00
B	01
C	10

- Diese Kodierung lässt sich ebenfalls verkleinern, etwa durch folgende Kodierung:

A	0
B	10
C	11

Damit ergäben sich durchschnittlich 5/3 bits pro Zeichen.

- Theoretisch kann sogar eine Kodierung gefunden werden, die nur 1,58496… bits je Zeichen erfordert.

Ein motivierendes Beispiel (4)

- Wir betrachten nun den allgemeinen Fall von n Ausgängen E_i mit Eintrittswahrscheinlichkeit p_i.

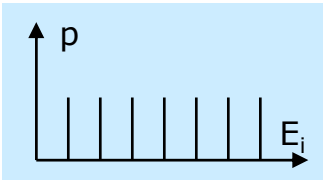
$p(Z=E_1) = p_1$	$p(Z=E_2) = p_2$	…	$p(Z=E_n) = p_n$

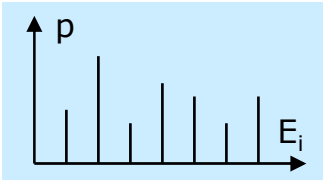
- Frage: Wie groß ist die kleinstmögliche Anzahl von bits – im Durchschnitt –, um einen gemäß der Verteilung erzeugten Zeichenstrom über eine binäre serielle Leitung zu übertragen?

- Diese Größe läßt sich wie folgt berechnen:

$$H(X) = - p_1 \cdot \log_2 p_1 - p_2 \cdot \log_2 p_2 - \ldots - p_n \cdot \log_2 p_n = - \sum_i p_i \cdot \log_2 p_i$$

- Sie heißt Entropie und ist ein Maß für den mittleren Informationsgehalt pro Zeichen einer Quelle. Man könnte bei vorliegender Information auch von einem Maß für „beseitigte Unsicherheit" sprechen. Oder: Der Informationsgehalt eines Zeichens ist umgekehrt proportional zum Logarithmus der Wahrscheinlichkeit, mit der man es erraten kann.





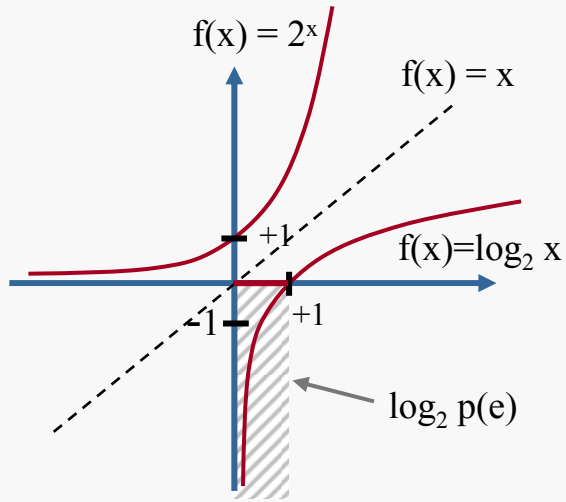
- Die Entropie ist offensichtlich hoch, wenn die Zeichen gleichmäßig verteilt sind. Die Entropie ist offensichtlich niedriger, wenn die Verteilung Berge und Täler aufweist. Dann können Zeichen offensichtlich leichter erraten werden.

- Wir wollen im folgenden zeigen, daß die Entropie maximal ist, wenn die Ereignisse gleichverteilt sind.

Die Entropie (1)

- Sei S eine Informationsquelle von Symbolen bzw. eine diskrete reelle Zufallsvariable, $S = \{s_1, s_2, ..., s_q\}$ mit den Eintrittswahrscheinlichkeiten $\{p(s_1), p(s_2), ..., p(s_q)\}$.
- Ein Zeichengenerator könnte gemäß dieser Wahrscheinlichkeitsverteilung Zeichen ausgeben.
- Ist die Eintrittswahrscheinlichkeit eines Zeichens unabhängig von den bisher gesendeten Zeichen, so heißt S diskrete gedächtnislose Quelle.
- Der Informationsgehalt eines ausgegebenen Zeichens ist damit

$$I(s_i) = \log (1/p(s_i)) = -\log p(s_i)$$



Die log-Funktion ist die Umkehrfunktion der e-Funktion

- Der gewichtete Informationsgehalt der Informationsquelle S beträgt dann

$$\langle I \rangle = \sum_{i=1}^{q} p(s_i) I(s_i) = -\sum_{i=1}^{q} p(s_i) \log p(s_i)$$

- Er ist definiert als die gewichtete Summe des Informationsgehalts aller Einzelereignisse.
- <*I*> heißt auch Entropie H(S) der Quelle S.

$$\langle I \rangle = H(S) = -\sum_{i=1}^{q} p(s_i) \log p(s_i)$$

Die Entropie (2)

Die Entropie kann als ein Maß für die Menge an Zufallsinformation bezeichnet werden, die in einem oder mehreren Zufallsereignissen oder in einer Informationsfolge steckt. Die Einheit der Entropie heißt 1 Shannon (Sh) und ist definiert als die Informationsmenge, die in einer Zufallsentscheidung eines Wurfes einer idealen Münze enthalten ist. Die beiden möglichen Ergebnisse - Kopf oder Zahl – besitzen dieselbe Eintrittswahrscheinlicheit p = 0,5.

- Dieser Entropiebegriff entspricht formal dem physikalischen Begriff der Entropie. In der Physik ist die Entropie assoziiert mit dem Begriff eines Maßes für die Unordnung eines Systems.
- In der Informationstheorie ist das System mit der größten „Unordnung" eines, bei welchem alle q Zeichen mit gleichen Eintrittswahrscheinlichkeiten 1/q auftreten. Ein solches System besitzt die höchste (maximale) Entropie. Es gilt dann nämlich

$$H(S) = -\sum_{i=1}^{q} \frac{1}{q} \log(\frac{1}{q}) = -\log(\frac{1}{q}) = \log q$$

- Wir wollen nun zeigen, daß Quellen mit gleichverteilten Eintrittswahrscheinlichkeiten ihrer Elemente die höchste Entropie besitzen.
- Wir betrachten zwei Quellen S_1 und S_2 mit q Zeichen und den Eintrittswahrscheinlichkeiten p_{1i} und p_{2i}. Dann ist

$$H_1 - H_2 = -\sum_i (p_{1i} \log p_{1i} - p_{2i} \log p_{2i}) \quad (*)$$

- Wir wollen zeigen, daß stets $H_1 \leq H_2$ gilt, wenn die Elemente von S_2 gleichverteilte Eintrittswahrscheinlichkeiten besitzen.

Die Entropie (3)

- Durch Addition der Null und einige einfache Termumformungen erhält man

$$H_1 - H_2 = -\sum_i (p_{1i} \log p_{1i} - p_{2i} \log p_{2i})$$

Addition der Null

$$= -\sum_i (p_{1i} \log p_{1i} + p_{1i} \log p_{2i} - p_{1i} \log p_{2i} - p_{2i} \log p_{2i})$$

$$= -\sum_i \left(p_{1i} \log(\frac{p_{1i}}{p_{2i}}) + (p_{1i} - p_{2i}) \log p_{2i} \right)$$

$$= -\sum_i p_{1i} \log(\frac{p_{1i}}{p_{2i}}) - \sum_i (p_{1i} - p_{2i}) \log p_{2i}$$

- Wir nehmen nun an, S_2 sei die Quelle mit identischen Eintrittswahrscheinlichkeiten. Dann gilt $H_2 = \log q$.
- Ferner gilt für den 2. Summanden der rechten Seite wg. $p_{2i} = 1/q$

$$-\sum_i (p_{1i} - p_{2i}) \log p_{2i} = -\log(\tfrac{1}{q}) \sum_i (p_{1i} - p_{2i})$$

$$= -\log(\tfrac{1}{q}) \left(\sum_i p_{1i} - \sum_i p_{2i} \right) = 0,$$

$$\text{da} \sum_i p_{1i} = 1 \quad \text{und} \quad \sum_i p_{2i} = 1.$$

Die Entropie (4)

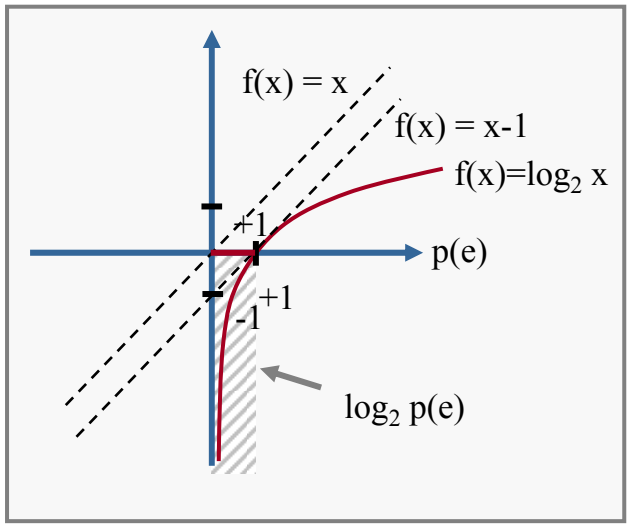
- Also reduziert sich der Ausdruck auf

$$H_1 - H_2 = -\sum_i p_{1i} \log(\frac{p_{1i}}{p_{2i}}) \text{ bzw.}$$

$$H_1 - \log q = -\sum_i p_{1i} \log(\frac{p_{1i}}{p_{2i}})$$

$$= \sum_i p_{1i} \log(\frac{p_{2i}}{p_{1i}})$$

- Aus log x ≤ x-1 (s. Abb. links) folgt für die rechte Seite



$$\sum_i p_{1i} \log(\frac{p_{2i}}{p_{1i}}) \le \sum_i p_{1i} (\frac{p_{2i}}{p_{1i}} - 1)$$

$$= \sum_i p_{2i} - \sum_i p_{1i} = 0,$$

$$\text{da } \sum_i p_{2i} = 1 \text{ und } \sum_i p_{1i} = 1.$$

Daraus folgt dann
$H_1 - \log q \le 0$ bzw. $H_1 \le \log q$,
also $H_1 \le H_2$.

- Die Gleichheit gilt, falls S_1 ebenfalls gleichverteilte Eintrittswahrscheinlichkeiten aufweist.

Die Maßeinheit Shannon (Sh)

- Ein Shannon (Sh) ist die nach Claude Shannon benannte dimensionslose Einheit für den Informationsgehalt einer Nachricht. Sie ist definiert als die theoretisch minimale Anzahl an bit, mit der eine Information abgebildet werden kann.
- Es geht hierbei aber nicht darum, wieviele ganze reale Bits (bei einer gegebenen Kodierung) notwendig sind, um die Nachricht zu übertragen oder zu speichern, weshalb der Informationsgehalt i.a. eine reelle Zahl ist.
- (Die Einheit „Shannon" wurde geschaffen, um die von Shannon verwendete Einheit „bit" zu ersetzen, nicht zu verwechseln mit dem klassischen Begriff „Bit". Dieser letztgenannte Begriff sollte hierbei per Konvention groß – also „Bit"– geschrieben werden, um eine Unterscheidung zu „bit" zu ermöglichen, was jedoch in der Praxis häufig nicht beachtet wurde.)
- Bsp.: Wir nehmen an, es soll in einer Nachricht ein Buchstabe übertragen werden. Dabei sei das Auftreten jeder der 26 Buchstaben des Alphabets gleich wahrscheinlich. Der Informationsgehalt dieser Nachricht beträgt also

$$-\log_2 (1/26) = 4{,}7004\ldots \text{ Sh.}$$

- Die Datenmenge, die benötigt wird, um diesen Informationsgehalt darzustellen, beträgt dagegen mindestens 5 Bit.

Bsp.: Nochmal der Münzwurf

Beispiel Münzwurf:

- Bei einer idealen bzw. fairen Münze sind die Münzbilder „Kopf" und „Zahl" gleichwahrscheinlich. Wir betrachten die Zufallsvariable X (Münzbild) über der diskreten Menge Z ={ Kopf, Zahl} mit den Wahrscheinlichkeiten p(Kopf) = p(Zahl) = ½, also P(X∈Z) = 1. Dann berechnet sich die Entropie zu

  $H(X) = -(½ \cdot \log ½ + ½ \cdot \log ½) = -(\log ½) = 1.$

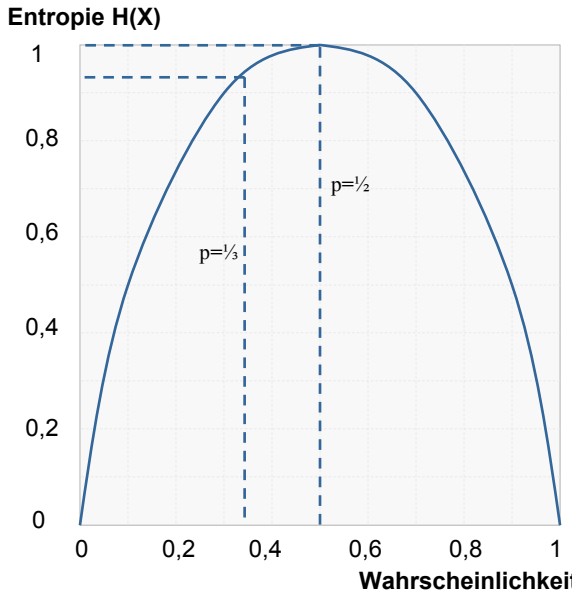
- Wir nehmen nun an, die Münze sei unfair bzw. „gezinkt". Die Wahrscheinlichkeit für das Münzbild „Kopf" sei 1/3, für „Zahl" dann dementsprechend 2/3. Dann ist die Entropie

  $H(X) = -(⅓ \cdot \log ⅓ + ⅔ \cdot \log ⅔) \approx 0{,}9183$

- Geht die Wahrscheinlichkeit eines Münzbildes gegen 1, so geht die Entropie gegen 0.

  Sei p(Kopf) = 1, dann ist H(X) = 0.

- Die rechte Grafik veranschaulicht, daß die Entropie maximal ist, wenn eine Gleichverteilung vorliegt,



Bsp.: Alphabet

Platz	Zeichen	Relative Häufigkeit
1.	E	17,40 %
2.	N	9,78 %
3.	I	7,55 %
4.	S	7,27 %
5.	R	7,00 %
6.	A	6,51 %
7.	T	6,15 %
8.	D	5,08 %
9.	H	4,76 %
10.	U	4,35 %
11.	L	3,44 %
12.	C	3,06 %
13.	G	3,01 %
14.	M	2,53 %
15.	O	2,51 %
16.	B	1,89 %
17.	W	1,89 %
18.	F	1,66 %
19.	K	1,21 %
20.	Z	1,13 %
21.	P	0,79 %
22.	V	0,67 %
23.	ß	0,31 %
24.	J	0,27 %
25.	Y	0,04 %
26.	X	0,03 %
27.	Q	0,02 %

Häufigkeitsverteilung der Buchstaben im deutschen Alphabet mit „ß".

Bei der Berechnung der Entropie wurde lediglich die statistische Verteilung der Zeichen verwendet; häufige Buchstabenkombinationen wie „sch", „st" oder „ch" und andere sprachliche Phänomene wurden nicht berücksichtigt.

Redundanz in Alphabeten:

- Bei einer Gleichverteilung der Zeichen im lateinischen Alphabet (26 Zeichen) beträgt die Entropie $H(Z) = H_{max}(Z) = \log_2(26) \approx 4{,}7004$ Sh.
- In natürlichen Sprachen sind die Zeichen nicht gleichverteilt. Im Deutschen taucht z.B. das „E" am häufigsten auf, mehr als 9 Mal häufiger als „B" und „W". Die Tabelle informiert über die relative Häufigkeit der Zeichen. Legt man diese statistische Verteilung der Zeichen zugrunde – das „ß" rechnen wir hier dem „s" zu – ergibt sich eine Entropie von
  $$H_{deu}(Z) \approx 4{,}0629 \text{ Sh/Zeichen.}$$
- Damit ergibt sich (wie in allen natürlichen Sprachen) eine gewisse Redundanz (Weitschweifigkeit) im „deutschen" Alphabet. Die Redundanz ist definiert als $R := H_{max}(Z) - H_{deu}(Z)$. Es ergibt sich
  $$R = H_{max}(Z) - H_{deu}(Z) \approx 0{,}6375 \text{ Sh.}$$
- Für alle Zeichen ergibt sich damit eine Gesamtredundanz GR von
  $$GR = 26*R \approx 16{,}575 \text{ Sh.}$$
- Dividiert man diesen Wert durch den durchschnittlichen Informationsgehalt eines Zeichens im „deutschen" Alphabet, so ergibt sich
  $$GR/H_{deu}(X) \approx 4{,}08$$
- Das Alphabet enthält also eine Redundanz von 4 Zeichen, d.h. es könnte also rechnerisch um 4 Zeichen reduziert werden, ohne daß Information verloren geht.

Bedingte Entropie – Motivierendes Beispiel (1)

- Frage: Läßt sich die Outputgröße Y vorhersagen, wenn man die Inputgröße X kennt?

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- Betrachte die linke Tabelle.* Eine Zeile repräsentiert einen Studenten. Die linke Spalte X enthält sein Hauptfach. Die rechte Spalte Y gibt an, ob der Student Computerspiele mag.
- Was läßt sich aus der Tabelle ableiten?
  - p(magComp = Ja) = 0,5
  - p(Fach=Math & magComp =Nein) = 0,25
  - p(Fach=Math) = 0,5
  - p(magComp = Ja | Fach=Kunst) = 0
- Ferner gilt:
  - H(X) = 1,5
  - H(Y) = 1

* nach einem Beispiel von Andrew W. Moore von der Carnegie Mellon University (www.cs.cmu.edu/~awm).

Bedingte Entropie – Motivierendes Beispiel (2)

- Definition der spezifischen bedingten Entropie:

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- H(Y | X=x) ist die Entropie von Y für alle Sätze der Tabelle, in denen X den Wert x annimmt.
- Beispiele:
  - H(Y | X = Math)
    = -p(Ja)·log p(Ja) – p(Nein)·log p(Nein)
    = - ½·log ½ - ½·log ½
    = 1
  - H(Y | X = Kunst)
    = -p(Ja)·log p(Ja) – p(Nein)·log p(Nein)
    = 0·log 0 – 1·log 1
    = 0 wg. (*)
  - H(Y | X = Inform)
    = -p(Ja)·log p(Ja) – p(Nein)·log p(Nein)
    = -1·log 1 – 0·log 0
    = 0 wg. (*)

$$(*)\quad \lim_{x\to 0}(x\cdot\log x)=\lim_{x\to 0}\left(\frac{\log x}{\frac{1}{x}}\right)=\lim_{x\to 0}\left(\frac{(\log x)'}{(\frac{1}{x})'}\right)=\lim_{x\to 0}\left(\frac{\frac{1}{x\cdot\ln 2}}{-\frac{1}{x^2}}\right)=-\lim_{x\to 0}\left(\frac{x}{\ln 2}\right)=0$$

nach der Regel von de l'Hospital

Bedingte Entropie – Motivierendes Beispiel (3)

- Definition der spezifischen bedingten Entropie:

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- H(Y | X) ist die durchschnittliche spezifische Entropie von Y

  und ist eine Antwort auf folgende Problemstellung:
  Wie hoch ist die bedingte Entropie von Y für einen zufällig gezogenen Satz, wenn dabei der jeweilige X-Wert bekannt ist?

  = der erwarteten Anzahl von Bits, um Y zu übertragen, wenn beide Seiten den Wert von X kennen

  $= \sum_i p(X=x_i) \cdot H(Y|\ X= x_i)$

Bedingte Entropie – Motivierendes Beispiel (4)

- Definition der spezifischen bedingten Entropie:

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- H(Y|X) heißt die bedingte Entropie von Y und ist definiert als

$$\sum_i p(X=x_i) \cdot H(Y|\ X=\ x_i)$$

- Beispiel:

x_i	$p(x_i)$	$H(Y|X=x_i)$
Mathematik	0,5	1
Kunst	0,25	0
Informatik	0,25	0

- H(Y|X) = 0,5*1 + 0,25*0 + 0,25*0 = 0,5

Information Gain – Der Informationsgewinn

- Definition des Information Gain (Informationsgewinn):

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- IG(Y|X) = H(Y) – H(Y|X) ist der mit der Kenntnis von X verbundene Informations-gewinn

- Beispiel:

  H(Y) = 1

  H(Y|X) = 0,5

  IG(Y|X) = 1 – 0,5 = 0,5

Der relative Informationsgewinn

- Definition des „Relative Information Gain“ (rel. Informationsgewinn):

X = Hauptfach

Y = mag Computerspiele

X	Y
Mathematik	Ja
Kunst	Nein
Mathematik	Nein
Mathematik	Ja
Kunst	Nein
Informatik	Ja
Mathematik	Nein
Informatik	Ja

- $$RIG(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}$$

RIG gibt Antwort auf folgende Problemstellung:
Es ist Y zu übertragen. Welchen Anteil an Bits kann man durchschnittlich einsparen, wenn beide Seiten X kennen?

- Beispiel:

  H(Y|X) = 0,5

  H(Y) = 1

  RIG(Y|X) = (1 – 0,5) / 1 = 0,5

Der Nutzen des IG

- Angenommen, man möchte prognostizieren, ob jemand älter als 80 Jahre wird. Zur Beantwortung dieser Frage könnte man z.B. historische Daten benutzen.
- Daraus könnte man etwa auf folgende Zusammenhänge schließen:

  IG (LangesLeben | Haarfarbe) = 0,01 (geringer Zshg.)

  IG (LangesLeben | Raucher) = 0,2 (signifikanter Zshg.)

  IG (LangesLeben) | Geschlecht) = 0,25 (signifikanter Zshg.)

  IG (LangesLeben) | EndzifferDerHandynummer) = 0,000001 (kein Zshg.)

- Nach diesen einfachen motivierenden Beispielen nun zur theoretischen Betrachtung der bedingten Entropie.

Die bedingte Entropie (1)

Es wird also lediglich p(x) durch die bedingte Wahrscheinlichkeit p(x|A) ersetzt.

- Wir definieren den Begriff der bedingten Entropie.
- Sei X eine reelle diskrete Zufallsvariabe mit Wertebereich M = $\{x_1, x_2, x_3, \ldots\}$ und sei x_i eine Ausprägung von X. Sei A ein Ereignis mit p(A)>0.

  Dann heißt $H(X|A) = -\sum_i p(x_i \mid A) \cdot \log p(x_i|A)$

  bedingte Entropie von X gegeben A.
- Wir erinnern an die Definition der bedingten Wahrscheinlichkeit: Seien die Ereignisse A, B $\subset \Omega$ (Ergebnisraum) und p(B) > 0. Dann ist die bedingte Wahrscheinlichkeit für A gegeben B bzw. die Wahrscheinlichkeit für A unter (der Bedingung) B definiert als

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

- Beispiel: Betrachte das Werfen eines fairen Würfels. Es handelt sich um ein Laplace-Experiment (, da alle Ergebnisse dieselbe Eintrittswahrscheinlichkeit besitzen).

  Sei A_1 das Ereignis, eine gerade Zahl (Ereignis {2,4,6}) zu würfeln.

  Sei A_2 das zu A_1 komplementäre Ereignis, eine ungerade Zahl (Ereignis {1,3,5}) zu würfeln.

  Dann ist $p(A_1)$= (Anz. günst. Ergebnisse)/(Anz. mögl. Ergebnisse) = $p(A_2)$ = 0,5.

Die bedingte Entropie (2)

Angenommen, wir erhalten die zusätzliche Information, das Ereignis B sei eingetreten, nämlich die geworfene Zahl sei kleiner gleich 3.

Dann beträgt die Wahrscheinlichkeit für das Ereignis A_1 gegeben B

$$p(A_1 \mid B) = \frac{p(A_1 \cap B)}{p(B)} = \frac{p(\{2\})}{p(\{1,2,3\})} = \frac{\frac{1}{6}}{\frac{3}{6}} = 1/3$$

Die Wahrscheinlichkeit für das Ereignis A_2 gegeben B beträgt

$$p(A_2 \mid B) = \frac{p(A_2 \cap B)}{p(B)} = \frac{p(\{1,3\})}{p(\{1,2,3\})} = \frac{\frac{2}{6}}{\frac{3}{6}} = 2/3$$

- Sei nun Y eine reelle diskrete Zufallsvariabe mit Wertebereich L = $\{y_1, y_2, y_3, \ldots\}$ und sei y_j eine Ausprägung von Y. Dann ist die bedingte Entropie von X gegeben Y definiert als gewichtetes Mittel der bedingten Entropien von X gegeben Ereignisse $Y = y_j \in L$, d.h.

$$H(X|Y) = \sum_{j, p(y_j)>0} p(y_j) \cdot H(X|Y = y_j)$$

Verbundentropie bzw. gemeinsame Entropie

- Sei x_i eine Realisierung der Zufallsvariablen X und y_j eine Realisierung der Zufallsvariablen Y. Dann ist die Verbundwahrscheinlichkeit bzw. die gemeinsame Wahrscheinlichkeit $p(x_i, y_j)$ die Wahrscheinlichkeit dafür, daß das Ereignis x_i gleichzeitig mit dem Ereignis y_j auftritt.
- Die Wahrscheinlichkeit, daß das Ereignis x_i auftritt, beträgt dann

$$p(x_i) = \sum_j p(x_i, y_j)$$

- Mit der bedingten Wahrscheinlichkeit läßt sich die Verbundwahrscheinlichkeit bzw. gemeinsame Wahrscheinlichkeit wie folgt ausdrücken:

$$p(x_i, y_j) = p(x_i) \cdot p(y_j \mid x_i) = p(y_j) \cdot p(x_i \mid y_j)$$

- Damit läßt sich nun die Verbundentropie bzw. gemeinsame Entropie definieren. Der mittlere Informationsgehalt der Verbundentropie je Ereignispaar statistisch abhängiger Ereignisse beträgt dann

$$H(X,Y) = -\sum_i \sum_j p(x_i, y_j) \cdot \log\left(p(x_i, y_j)\right)$$

Die Transinformation (1)

- Die Transinformation bzw. gegenseitige Information beschreibt die Stärke des statistischen Zusammenhangs zweier Zufallsgrößen X und Y. Die Ungewißheit einer Zufallsvariablen wird durch Kenntnis einer anderen reduziert.

$$I(X,Y) = H(X) - H(X \mid Y)$$

- Die Transinformation ist also ein Maß für die Information, die eine Zufallsvariable über eine andere enthält. Sie ist die Abnahme der Ungewißheit über eine Zufallsvariable X, wenn man Y kennt.
- Verschwindet die Transinformation, so spricht man von statistischer Unabhängigkeit der beiden Zufallsgrößen. Die Transinformation wird maximal, wenn sich eine Zufallsgröße vollkommen aus der anderen berechnen lässt.
- Nimmt die Transinformation zu, so verringert sich die Unsicherheit über eine Zufallsgröße unter der Voraussetzung, daß die andere bekannt ist. Ist die Transinformation maximal, verschwindet die Unsicherheit folglich.
- Diese Größe heißt auch Information Gain (Informationsgewinn):

$$IG(X,Y) = H(X) - H(X \mid Y)$$

- Der relative Informationsgewinn (Relative Information Gain) ist wie folgt definiert:

$$RIG(X,Y) = \frac{H(X) - H(X \mid Y)}{H(Y)}$$

Die Transinformation (2)

- Die Transinformation läßt sich auch über die Wahrscheinlichkeits-dichtefunktionen für X und Y ausdrücken.

$$I(X,Y) = H(X) - H(X \mid Y)$$

$$= -\sum_{x} p(x) \cdot \log p(x) - \left(-\sum_{x,y} p(x,y) \cdot \log p(x \mid y) \right)$$

$$= -\sum_{x,y} p(x,y) \cdot \log p(x) + \sum_{x,y} p(x,y) \cdot \log p(x \mid y)$$

$$= \sum_{x,y} p(x,y) \cdot \log \frac{p(x \mid y)}{p(x)}$$

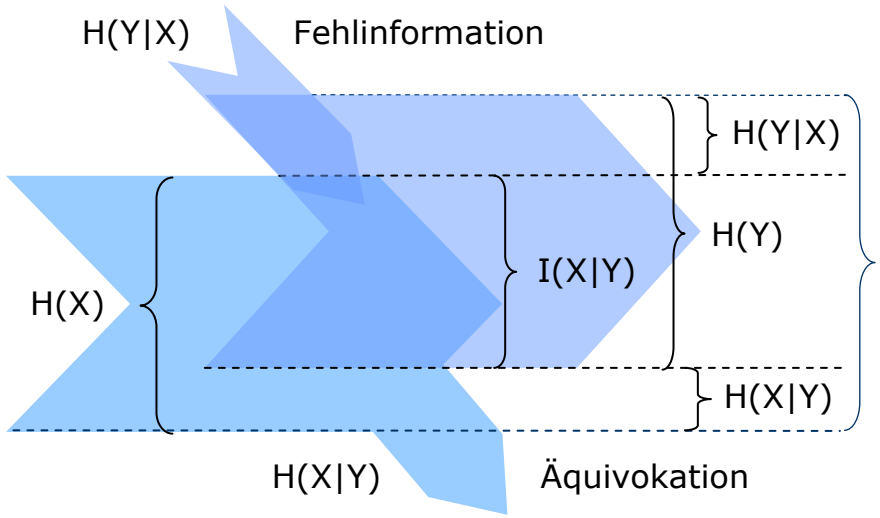
$$= \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

- Aus Symmetriegründen gilt ferner:

$$I(X,Y) = H(Y) - H(Y \mid X)$$

Die Transinformation (3)

- Die folgende Grafik veranschaulicht die Gesamtzusammenhänge.



$$H(X,Y) = H(X) + H(Y) - I(X,Y)$$

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = IG(X,Y)$$

- Bei jedem Kommunikationsprozess treten durch gewisse Übertragungsphänomene verschiedene Arten von Informationen auf:
  - Fehlinformation/Irrelevanz: Hinzufügen irrelevanter, störender Information
  - Äquivokation: Verlust relevanter Information
  - Transinformation: erwünschte relevante Information, die tatsächlich vom Sender zum Empfänger übertragen wird,