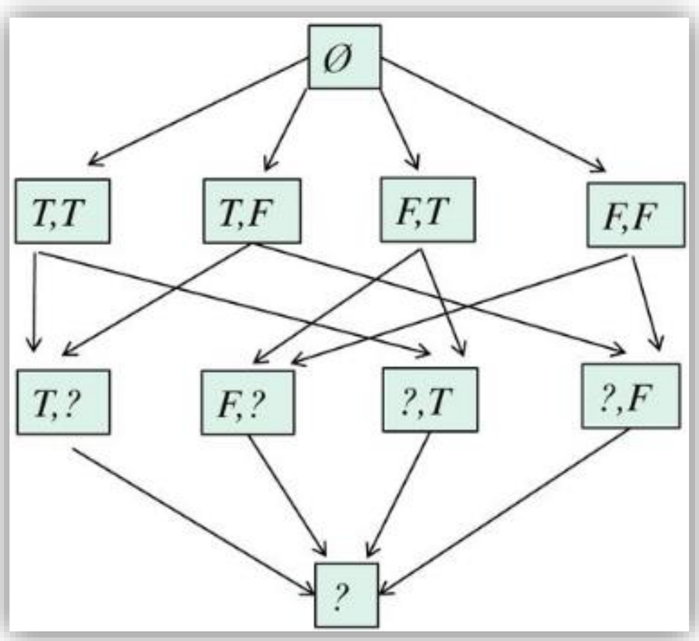
Aufgabenstellungen:
unten auf den Folien, vor jeweiligen Lösungen

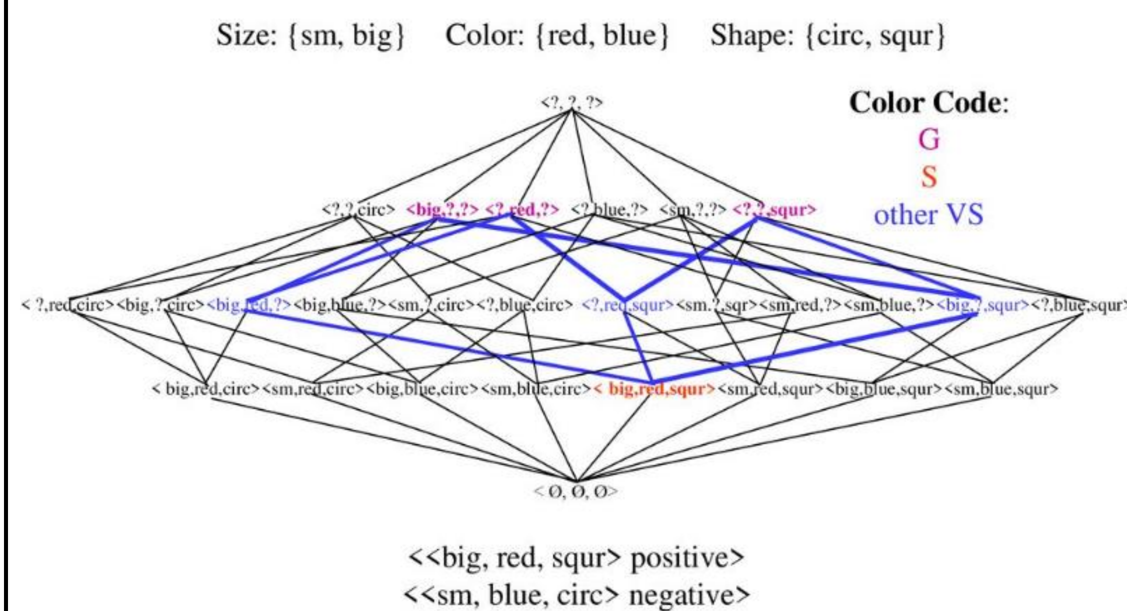
***Anwendungsbeispiele** zur heutigen Sitzung:*

konjunktive Hypothesen

***Hypothesengitter** mit zwei 2-wertigen Attributen*



Versionenraum [S; G] – Attributen-#: 3, |Attribut| = 2



QuickFact: eine Instanz aka Beispiel (z.B. $\langle big, red, squr\rangle$) **erfüllt** eine Hypothese (z.B. $\langle ?, red, ?\rangle$) heißt dass die beiden durch einen **Kantenzug** verbunden sind, bei dem sich keine Knoten auf der gleichen Ebene befinden.

***Die Schlüsselbegriffe** der heutigen Sitzung:*

Trainingsdaten
Hypothesen
Kandidatenelimination
Entscheidungsbaum
Entropie
Informationsgewinn

Inhaltsverzeichnis

Aufgabe T5.1: *induktives Lernen - Hypothesenraum*

Es werde mit dem Hypothesenraum aus Def. 5.1 des Vorlesungsskripts (S. 128) gearbeitet.

Gegeben sind die folgenden Trainingsdaten:

#	Himmel	Lufttemperatur	Luftfeuchtigkeit	Wind	Wasser	Vorhersage	Sport_macht_Spaß
1	sonnig	warm	hoch	stark	kalt	wechselhaft	ja
2	regnerisch	kalt	hoch	stark	warm	wechselhaft	nein
3	sonnig	warm	hoch	stark	warm	gleichbleibend	ja
4	sonnig	warm	normal	stark	warm	gleichbleibend	ja

Das Attribut Sport_macht_Spaß hängt von verschiedenen Umweltfaktoren ab. Insbesondere gibt es für das Attribut Himmel drei, für Lufttemperatur, Luftfeuchtigkeit, Wind, Wasser und Vorhersage jeweils zwei verschiedene mögliche Werte.

Erklären Sie, warum die Anzahl der semantisch verschiedenen möglichen Hypothesen für das Attribut Sport_macht_Spaß 973 beträgt!

Angenommen, es gäbe ein weiteres Attribut Wasserströmung, welches drei verschiedene Werte annehmen kann, wie groß wären dann die Anzahl der möglichen Instanzen und der (syntaktisch und semantisch verschiedenen) möglichen Hypothesen ?

Wie würden diese Werte sich ändern, wenn ganz allgemein ein Attribut A mit k verschiedenen Werten hinzukäme ?

Definition 5.1

Es sei $\Omega = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N$ *ein Objektraum. Dann heißen die Elemente aus*

$$\mathcal{H} = (\mathcal{X}_1 \cup \{?\}) \times \ldots \times (\mathcal{X}_N \cup \{?\})$$

Hypothesen *über* Ω.
Die Menge $\mathcal{H}$ *heißt* **Hypothesenraum** *über* Ω.

Ein Beispielobjekt $x \in \Omega$ *„genügt“ der Hypothese* $h \in \mathcal{H}$ *(*x **erfüllt** h *bzw.* $h \models x$*) genau dann, wenn gilt:*

$$\forall i = 1, \ldots, N : \quad (h_i = ?) \vee (h_i = x_i)$$

Oft wird noch (∅, ∅, ∅, ..., ∅) als niemals erfüllbare Hypothese zum Hypothesenraum hinzugenommen.

Aufgabe T5.1

> **QuickFact**: alternativ zur obigen Definition, das Symbol der leeren Menge $\emptyset$ wird zu jedem Attribut $\aleph_i$ hinzugefügt (also zusätzlich zum Fragezeichen ?) – dieses lässt sich wie “es ist kein Wert beim Attribut $\aleph_i$ zugelassen” interpretieren (**Hausaufgabe**: für welche Anwendungszwecke bräuchte man das?), – in so einer Fassung wären alle Hypothesen mit einer oder mehreren $\emptyset$-Komponenten semantisch äquivalent (zur leeren Hypothese), doch offenbar nicht syntaktisch

Lösung:

Schritt #1: Übersetzung der Attribute in die Symbolsprache

Himmel	$\aleph_1$	***Wind***	$\aleph_4$
Lufttemperatur	$\aleph_2$	***Wasser***	$\aleph_5$
Luftfeuchtigkeit	$\aleph_3$	***Wettervorhersage***	$\aleph_{6=N}$

Schritt #2: Anzahl der Werte pro Attribut

$\lvert\aleph_1\rvert$	$\lvert\aleph_2\rvert$	$\lvert\aleph_3\rvert$	$\lvert\aleph_4\rvert$	$\lvert\aleph_5\rvert$	$\lvert\aleph_6\rvert$
3	2	2	2	2	2

Schritt #3: Anzahl der ***sem***antisch verschiedenen Hypothesen

$$|\mathcal{H}|_{sem} = \left|\prod_{i=1}^{6} \aleph_i \cup \{?\}\right|_{sem} + |\{(\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset)\}|_{sem} = \prod_{i=1}^{6}|\aleph_i \cup \{?\}|_{sem} + 1 = \prod_{i=1}^{6}(|\aleph_i|_{sem}+1)+1 = \prod_{i=1}^{6}(|\aleph_i|+1)+1$$

$$= (3+1)\cdot(2+1)\cdot(2+1)\cdot(2+1)\cdot(2+1)\cdot(2+1)+1 = (3+1)^1\cdot(2+1)^5+1 = 4^1\cdot 3^5+1 = 4\cdot 243+1 = 972+1$$

> **FunFact**: die Anzahl von allen möglichen Zielfunktionen (die wir durch Hypothesen approximieren wollen) mit 6 Variablen wäre viel größer – wir betrachten aber einen relativ kleinen Teil davon, und zwar die konjunktiven Hypothesen

Schritt #4: Anzahl der Instanzen und der *sem*antisch (hier: *aka* *syn*taktisch) verschiedenen Hypothesen bei einem weiteren k-wertigen Attribut $\aleph_7 := A$

$$|\Omega_{+A}| = \left|\prod_{i=1}^{6+1}\aleph_i\right| = \prod_{i=1}^{7}|\aleph_i| = |\aleph_7|\cdot\prod_{i=1}^{6}|\aleph_i| = |A|\cdot\prod_{i=1}^{6}|\aleph_i| = k\cdot\prod_{i=1}^{6}|\aleph_i| = k\cdot 3\cdot 2\cdot 2\cdot 2\cdot 2\cdot 2 = k^1\cdot 3^1\cdot 2^5 = k\cdot 96$$

$$|\mathcal{H}_{+A}|_{s\_} = \left|\prod_{i=1}^{6+1}\aleph_i \cup \{?\}\right|_{s\_} + |\{(\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset)\}|_{s\_} = \prod_{i=1}^{7}|\aleph_i \cup \{?\}|_{s\_} + 1 = |\aleph_7 \cup \{?\}|_{s\_}\cdot\prod_{i=1}^{6}|\aleph_i \cup \{?\}|_{s\_} + 1 = |\aleph_7 \cup \{?\}|_{s\_}\cdot\left(\prod_{i=1}^{6}|\aleph_i \cup \{?\}|_{s\_} + 1 - 1\right) + 1$$

$$= |\aleph_7 \cup \{?\}|_{s\_}\cdot(|\mathcal{H}|_{s\_} - 1) + 1 = (|A|_{s\_}+1)\cdot(|\mathcal{H}|_{s\_}-1)+1 = (|A|+1)\cdot(973-1)+1 = (k+1)\cdot 972+1$$

Aufgabe T5.2: *Kandidateneliminationsverfahren*

Gegeben sind die folgenden Trainingsdaten:

Origin	Manufacturer	Color	Decade	Type	Example Type
Japan	Honda	Blue	1980	Economy	Positive
Japan	Toyota	Green	1970	Sports	Negative
Japan	Toyota	Blue	1990	Economy	Positive
USA	Chrysler	Red	1980	Economy	Negative
Japan	Honda	White	1980	Economy	Positive

Führen Sie den Algorithmus der Candidate-Elimination mit den Trainingsdaten in der angegebenen Reihenfolge durch und geben Sie die dabei erzeugte Sequenz der Begrenzungsmengen S und G an.

Aufgabe T5.2

Schritt #1: theoretische Vorüberlegungen

Beim Kandidateneliminationsalgorithmus muss man die Mengen $\boldsymbol{S}$ und $\boldsymbol{G}$ Schritt für Schritt so verändern, damit jede Hypothese aus dem Intervall (in anderen Worten, aus dem **Versionenraum**) $[\boldsymbol{S}; \boldsymbol{G}]$ konsistent mit allen schon betrachteten Beispielen bleibt.

Dabei muss man nur **kleinstmögliche Änderungen** vornehmen – sonst wird man entweder unnötigen / überflüssigen Kram sammeln, oder einige konsistente Hypothesen weglassen / verlieren.

Weitere wichtige Überlegungen:

$\boldsymbol{S}$ - die Menge der speziellsten (aka spezifischsten aka kleinsten) Hypothesen für unseren zu lernenden Begriff (hier: „Example Type“); diese Menge dient als ***die unterste Schranke*** für alle Hypothesen, die mit den schon betrachteten Beispielen konsistent sein müssen. Also ganz am Anfang (beim nullten Schritt) enthält $\boldsymbol{S}$ nur die kleinste, nie erfüllbare Hypothese: $\boldsymbol{S_0} = \{(\emptyset, \emptyset, \emptyset, \emptyset, \emptyset)\}$.

$\boldsymbol{G}$ - die Menge der generellsten (aka größten) Hypothesen für unseren zu lernenden Begriff (hier - „Example Type“); diese Menge dient als ***die oberste Schranke*** für alle Hypothesen, die mit den schon betrachteten Beispielen konsistent sein müssen. Also ganz am Anfang (beim nullten Schritt) besteht $\boldsymbol{G}$ aus der größten, stets erfüllbaren Hypothese: $\boldsymbol{G_0} = \{(?\,,?\,,?\,,?\,,?\,)\}$.

Um Hypothesen miteinander vergleichen zu können, muss man eine Relation (eine **Halbordnung**) haben, die den Vergleich zwischen 2 Hypothesen ermöglicht (natürlich nicht alle Hypothesen sind vergleichbar - deshalb reden wir über die **Halb**ordnung und nicht über die totale Ordnung).

Diese Relation basiert nämlich auf der Inklusion folgender Art:

wir sagen dass eine Hypothese $\boldsymbol{h_1}$ genereller als eine andere Hypothese $\boldsymbol{h_2}$ ist (bzw. h_2 spezieller als h_1 ist) und drücken das $\boldsymbol{h_1} \supseteq \boldsymbol{h_2}$ aus („$\supset$“ bedeutet „$\supseteq$ und $\neq$“) genau dann, wenn jede Instanz (d.h. jedes Beispiel), die $\boldsymbol{h_2}$ erfüllt, auch $\boldsymbol{h_1}$ erfüllt.

Und eine Instanz $\boldsymbol{x}$ erfüllt eine Hypothese $\boldsymbol{h}$ genau dann, wenn jede Komponente von $\boldsymbol{h}$ entweder gleich dem Fragezeichen ? ist, oder gleich der entsprechenden Komponente von $\boldsymbol{x}$ ist.

Jede Instanz (d.h. jedes Beispiel) *per se* ist eine Hypothese, die ausschließlich von sich selbst erfüllt wird.

Aufgabe T5.2

Theoretische Vorüberlegungen

Einige Beispiele:

h_1	h_2	**Vergleich *(Relation)***
$(\boldsymbol{Japan}, \boldsymbol{Honda}, \boldsymbol{Blue}, \mathbf{1980}, \boldsymbol{Economy})$	$(Japan, ?, Blue, ?, Economy)$	$\boldsymbol{h_1 \subset h_2}$
$(\boldsymbol{Japan}, \boldsymbol{Honda}, \boldsymbol{Blue}, \mathbf{1980}, \boldsymbol{Economy})$	$(Japan, Toyota, Green, 1970, Sports)$	nicht vergleichbar
$(\boldsymbol{Japan}, ?, \boldsymbol{Blue}, ?, \boldsymbol{Economy})$	$(?, Honda, ?, ?, ?)$	nicht vergleichbar
$(\boldsymbol{Japan}, ?, \boldsymbol{Blue}, ?, \boldsymbol{Economy})$	$(?, ?, Blue, ?, Economy)$	$\boldsymbol{h_1 \subset h_2}$

Interessanter Hinweis: man kann praktisch **ein Gitter aus allen Hypothesen** aufbauen (***empfohlen!*** – verschafft einen anschaulichen Überblick):

unterste Ebene **[Basis]** wird nur die kleinste Hypothese (∅, ∅, ∅, ∅, ∅) enthalten,
nächste Ebene **[Grundgeschoss]** – alle möglichen **Instanzen** (Beispiele),
weitere Ebene [1. OG] – alle möglichen 5-Tupel mit **einem** einzigen Fragezeichen,
noch weitere Ebene [2. OG] – alle möglichen 5-Tupel mit **zwei** Fragezeichen,
…
oberste Ebene **[Dachgeschoss]**– nur die größte Hypothese $(?, ?, ?, ?, ?)$

***Anregung** [NICHT klausurrelevant]*: falls man bisschen tiefer in Mathe wühlen will, kann man die Anzahlen von Elementen auf jeder Ebene beobachten – die weisen ein echt schönes Muster auf. Auch die Art der Vernetzung ($m\!:\!n$ Beziehungen) ist bemerkenswert.

Aufgabe T5.2

Theoretische Vorüberlegungen

Positive Beispiele:

Wenn wir ein Positivbeispiel (+) $\boldsymbol{x}$ an der Reihe haben, müssen wir sicher gehen, dass wir unser aktuelles Intervall $[\boldsymbol{S}; \boldsymbol{G}]$ so verändern, dass wir keine von diesem Beispiel erfüllte Hypothese (aus $[\boldsymbol{S}; \boldsymbol{G}]$) verlieren und dabei keine von diesem Beispiel nicht erfüllte Hypothese (aus $[\boldsymbol{S}; \boldsymbol{G}]$) behalten (**Hausaufgabe**: was würde es im Hypothesengitter heißen?).

Man braucht keine Änderungen zu $\boldsymbol{S}$ und $\boldsymbol{G}$ falls alles in Ordnung bleibt – also Änderungen ***(elementweise** und dabei mit **kleinstmöglichen Schritten)*** sind nur in dem Fall vorzunehmen, wenn es Elemente aus $\boldsymbol{S}$ oder aus $\boldsymbol{G}$ gibt, die vom aktuellen Positivbeispiel **nicht erfüllt** werden:

1. für jedes **solche Element** aus $\boldsymbol{S}$:
   PS1: wir **generalisieren** aka vergrößern dieses Element **zu allen kleinstmöglichen Hypothesen**, die von diesem Beispiel erfüllt werden (**Hausaufgabe**: kann man mehrere von solchen kleinstmöglichen, vom aktuellen positiven Beispiel erfüllten Hypothesen haben oder wird es nur eine einzige geben?).

   PS2: dabei stellen wir sicher, dass es keine miteinander vergleichbaren Elementpaare durch die **PS1**-Generalisierungen entstanden sind – wenn aber schon, dass müssen wir die Menge $\boldsymbol{S}$ vom unnötigen Kram reinigen, das heißt, **aus jedem vergleichbaren Paar $\boldsymbol{h_1} \subseteq \boldsymbol{h_2}$** einfach die generellere (also größere; in diesem Fall ist es $\boldsymbol{h_2}$) **entfernen**.

   PS3: nach diesen Generalisierungen, wir müssen sicher gehen, dass die Menge $\boldsymbol{S}$ ihrer Definition entspricht (sonst wird unser Versionenraum $[\boldsymbol{S}; \boldsymbol{G}]$ verletzt – **Hausaufgabe**: überlegt, warum?), das heißt, **für jedes Element $\boldsymbol{h_S}$ aus $\boldsymbol{S}$ findet man so ein Element $\boldsymbol{h_G}$ aus $\boldsymbol{G}$, dass $\boldsymbol{h_S} \subseteq \boldsymbol{h_G}$ – falls nicht, dann sind solche Elemente aus $\boldsymbol{S}$ zu entfernen.**

2. für jedes **solche Element** aus $\boldsymbol{G}$:
   PG: klar, wenn man dieses Element genereller macht (durch die Umwandlung von konfliktierenden Komponenten zum Fragezeichen ?), wird die generalisierte Hypothese vom aktuellen Beispiel zwar erfüllt, aber diese Generalisierung wird das Intervall $[\boldsymbol{S}; \boldsymbol{G}]$ vergrößern und dadurch die Inkonsistenz in Bezug auf die vergangenen (bereits betrachteten) Beispielen einführen (**Hausaufgabe**: versucht zu verstehen, warum?).
   Das heißt, es bleibt uns nichts anderes übrig, als dieses Element einfach **aus $\boldsymbol{G}$ zu entfernen**.

Aufgabe T5.2

Theoretische Vorüberlegungen

Negative Beispiele:

Bei Negativbeispielen (-) $\boldsymbol{x}$ verläuft alles praktisch identisch zu Positivbeispielen, wegen Dualität. Dieses heißt, dass man statt der **PG**, **PS1**, **PS2**, **PS3** Vorgänge die **NS**, **NG1**, **NG2**, **NG3** Vorgänge hat (**N** steht für Negativbeispiele, **G** – für die Menge $\boldsymbol{G}$ und **S** – für die Menge $\boldsymbol{S}$):

bei **NG1** werden erfüllte Elemente spezialisiert aka spezifiziert aka verkleinert *(bis diese vom aktuellen Negativbeispiel nicht erfüllt werden - solche Spezialisierungen sind nicht einzigartig und von daher* ***muss man alle möglichen Spezialisierungen bilden****)*,

bei **NG2** werden die spezielleren **entfernt**,

bei **NG3** werden alle Elemente $\boldsymbol{h_G}$ mit der Eigenschaft $\nexists \boldsymbol{h_S} \in \boldsymbol{S}: \boldsymbol{h_S} \subseteq \boldsymbol{h_G}$ aus $\boldsymbol{G}$ **entfernt**.

Bei **NS** werden alle Elemente, die vom aktuellen negativen Beispiel erfüllt werden, aus $\boldsymbol{S}$ **entfernt**.

Zum Schluss fassen wir unsere Mengen in einem Mengenintervall $[\boldsymbol{S}; \boldsymbol{G}]$ zusammen, das aus einem Element (im besten Fall), aus mehreren Elementen oder aus gar keinem Element (d.h. die leere Menge) bestehen kann.

Dieses Intervall lässt sich folgendermaßen intuitiv interpretieren: all die Hypothesen, die zu diesem Intervall gehören, sind mit den Lerndaten (hier: mit 5 Beispieldaten) konsistent.

Wenn wir beispielsweise nur ein einziges Element im oben genannten Intervall bekommen, dann sagt man, dass der Begriff (induktiv) gelernt wurde.

QuickFacts:
die kleinste Generalisierung aka Vergrößerung einer Hypothese erfolgt durch die Umwandlung einer deren Komponenten (diese muss dabei ein Attributenwert und nicht ein Fragezeichen sein) in das Fragezeichen ?

die umgekehrte Reihenfolge der Umwandlung (d.h. eine Fragezeichen-Komponente zu einem Wert entsprechendes Attributes) wäre die kleinste Spezifizierung aka Verkleinerung

Aufgabe T5.2

Schritt #2: Lösung

I Beispiel) $x = (x_1, x_2, x_3, x_4, x_5) = (Japan, Honda, Blue, 1980, Economy)$ - Positivbeispiel

Aktuelle Menge S: $\{(\emptyset, \emptyset, \emptyset, \emptyset, \emptyset)\}$
Aktuelle Menge G: $\{(?,?,?,?,?)\}$

Teilschritt PG: $G = \{(?,?,?,?,?)\}$ **– wird zur aktuellen G**
Teilschritt PS1: $S = \{(Japan, Honda, Blue, 1980, Economy)\}$
Teilschritt PS2: $S = \{(Japan, Honda, Blue, 1980, Economy)\}$
Teilschritt PS3: $S = \{(Japan, Honda, Blue, 1980, Economy)\}$ **– wird zur aktuellen S**

II Beispiel) $x = (x_1, x_2, x_3, x_4, x_5) = (Japan, Toyota, Green, 1970, Sports)$ - Negativbeispiel

Aktuelle Menge S: $\{(Japan, Honda, Blue, 1980, Economy)\}$
Aktuelle Menge G: $\{(?,?,?,?,?)\}$

Teilschritt NS: $S = \{(Japan, Honda, Blue, 1980, Economy)\}$ **– wird zur aktuellen S**
Teilschritt NG1: $G = \{(?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy)\}$
Teilschritt NG2: $G = \{(?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy)\}$
Teilschritt NG3: $G = \{(?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy)\}$ **– wird zur aktuellen G**

III Beispiel) $x = (x_1, x_2, x_3, x_4, x_5) = (Japan, Toyota, Blue, 1990, Economy)$ - Positivbeispiel

Aktuelle Menge S: $\{(Japan, Honda, Blue, 1980, Economy)\}$
Aktuelle Menge G: $\{(?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy)\}$

Teilschritt PG: $G = \{(?, ?, Blue, ?, ?), (?, ?, ?, ?, Economy)\}$ **– wird zur aktuellen G**
Teilschritt PS1: $S = \{(Japan, ?, Blue, ?, Economy)\}$
Teilschritt PS2: $S = \{(Japan, ?, Blue, ?, Economy)\}$
Teilschritt PS3: $S = \{(Japan, ?, Blue, ?, Economy)\}$ **– wird zur aktuellen S**

Aufgabe T5.2

Lösung:

IV Beispiel) $x = (x_1, x_2, x_3, x_4, x_5) = (USA, Chrysler, Red, 1980, Economy)$ - Negativbeispiel

Aktuelle Menge S: $\{(Japan, ?, Blue, ?, Economy)\}$
Aktuelle Menge G: $\{(?, ?, Blue, ?, ?), (?, ?, ?, ?, Economy)\}$

Teilschritt NS: $S = \{(Japan, ?, Blue, ?, Economy)\}$ **– wird zur aktuellen S**

Teilschritt NG1: $G = \left\{\begin{matrix} (?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy), (?, Honda, ?, ?, Economy), (?, Toyota, ?, ?, Economy), \\ (?, ?, Blue, ?, Economy), (?, ?, Green, ?, Economy), (?, ?, ?, 1970, Economy), (?, ?, ?, 1990, Economy) \end{matrix}\right\}$

Teilschritt NG2: $G = \left\{\begin{matrix} (?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy), (?, Honda, ?, ?, Economy), (?, Toyota, ?, ?, Economy), \\ (?, ?, Green, ?, Economy), (?, ?, ?, 1970, Economy), (?, ?, ?, 1990, Economy) \end{matrix}\right\}$

Teilschritt NG3: $G = \{(?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy)\}$ **– wird zur aktuellen G**

V Beispiel) $x = (x_1, x_2, x_3, x_4, x_5) = (Japan, Honda, White, 1980, Economy)$ - Positivbeispiel

Aktuelle Menge S: $\{(Japan, ?, Blue, ?, Economy)\}$
Aktuelle Menge G: $\{(?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy)\}$

Teilschritt PG: $G = \{(Japan, ?, ?, ?, Economy)\}$ **– wird zur letzten G**
Teilschritt PS1: $S = \{(Japan, ?, ?, ?, Economy)\}$
Teilschritt PS2: $S = \{(Japan, ?, ?, ?, Economy)\}$
Teilschritt PS3: $S = \{(Japan, ?, ?, ?, Economy)\}$ **– wird zur letzten S**

Anregung/Tipp: baut ein Hypothesengitter für diese Trainingsdaten auf und versucht den ganzen Vorgang auf dem erstellten Diagramm zu visualisieren – so werdet ihr am besten verstehen was ihr bei diesem Verfahren wirklich tut bzw. tun solltet, und, inter alia, werdet ihr auch Fehler erkennen/vorbeugen

Aufgabe T5.3: *Entscheidungsbäume - Einführung*

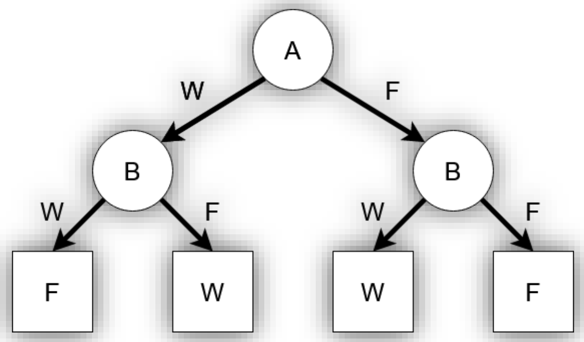
Geben Sie zu folgenden Booleschen Funktionen einen Entscheidungsbaum an:

(a) A xor B ("xor" = exklusives Oder)

(b) $(A \wedge B) \vee (C \wedge D)$

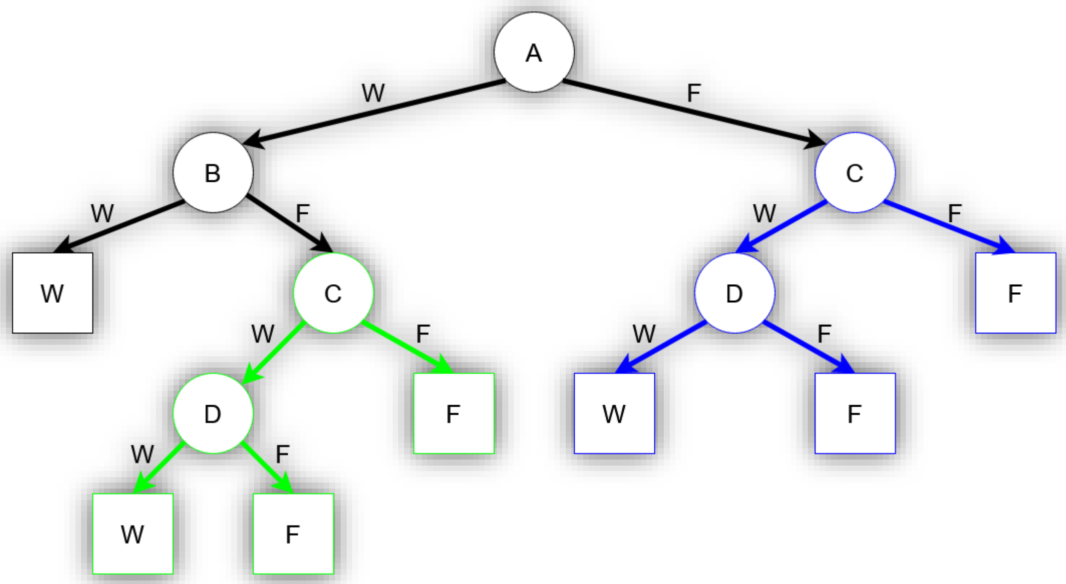
Aufgabe T5.3

Lösung: (a)



QuickFact: die Reihenfolge von zu belegenden Variablen wäre gleichgültig, sobald ihr euch um die Verringerung von Verästelungen nicht kümmert – die Entscheidungsbäume sind daher alles andere als eindeutig

Lösung: (b)



Aufgabe T5.4: *Entscheidungsbäume – Entropie, Informationsgewinn*

Man betrachte folgende Menge von Trainingsbeispielen:

Instanz	a_1	a_2	Klassifikation
1	W	W	+
2	W	W	+
3	W	F	-
4	F	F	+
5	F	W	-
6	F	W	-

Wie groß ist die Entropie der Trainingsbeispiele in Bezug auf die Klassifikation als Zielfunktion? Wie groß ist der Informationsgewinn (*information gain*) von a_2 relativ zu diesen Trainingsbeispielen?

Aufgabe T5.4

Lösung:

Schritt #1: Entropie $\mathbf{H}$ (aka *Überraschungsgrad* aka *Unbestimmtheitsmaß* aka *erwarteter Informationsgehalt [hier: in Bits]*) der Trainingsdaten $\boldsymbol{S} = S^{+} + S^{-}$

$$\boldsymbol{H}(\boldsymbol{S}) = \frac{|S^+|}{|S|}\cdot\log_2\frac{|S|}{|S^+|} + \frac{|S^-|}{|S|}\cdot\log_2\frac{|S|}{|S^-|} = \frac{3}{6}\cdot\log_2\frac{6}{3} + \frac{3}{6}\cdot\log_2\frac{6}{3} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 = 1$$

FunFact: falls ihr keine Fans von Logarithmen seid oder euch die thermodynamische Entropie ablenkt und in den Physik-Dschungel entführt ☺, es stehen auch weitere "Kostenfunktionen" aka "Maße der Unreinheit" zwecks Zerlegung der Datenmenge zur Verfügung – z.B. der simple, aber mächtige Gini-Koeffizient

Schritt #2: Informationsgewinn $\boldsymbol{G}$ vom zweiten Attribut $\boldsymbol{a_2}$

$$\boldsymbol{G}(\boldsymbol{a_2}) = \boldsymbol{H}(\boldsymbol{S}) - \left(\frac{|S_{a_2=W}|}{|S|}\cdot\boldsymbol{H}(\boldsymbol{S_{a_2=W}}) + \frac{|S_{a_2=F}|}{|S|}\cdot\boldsymbol{H}(\boldsymbol{S_{a_2=F}})\right)$$

$$= 1 - \left(\frac{4}{6}\cdot\left(\frac{|S^+_{a_2=W}|}{|S_{a_2=W}|}\cdot\log_2\frac{|S_{a_2=W}|}{|S^+_{a_2=W}|} + \frac{|S^-_{a_2=W}|}{|S_{a_2=W}|}\cdot\log_2\frac{|S_{a_2=W}|}{|S^-_{a_2=W}|}\right) + \frac{2}{6}\cdot\left(\frac{|S^+_{a_2=F}|}{|S_{a_2=F}|}\cdot\log_2\frac{|S_{a_2=F}|}{|S^+_{a_2=F}|} + \frac{|S^-_{a_2=F}|}{|S_{a_2=F}|}\cdot\log_2\frac{|S_{a_2=F}|}{|S^-_{a_2=F}|}\right)\right)$$

$$= 1 - \left(\frac{4}{6}\cdot\left(\frac{2}{4}\cdot\log_2\frac{4}{2} + \frac{2}{4}\cdot\log_2\frac{4}{2}\right) + \frac{2}{6}\cdot\left(\frac{1}{2}\cdot\log_2\frac{2}{1} + \frac{1}{2}\cdot\log_2\frac{2}{1}\right)\right) = 1 - \left(\frac{4}{6}\cdot 1 + \frac{2}{6}\cdot 1\right) = 1 - 1 = 0$$

QuickFact: Informationsgewinn gleich Null sagt uns dass es nix Schlimmeres gäbe als das zweite Attribut vor dem ersten Attribut für die Aufteilung der Datenmenge zu verwenden – dessen Teilmengen (sprich, $S_{a_2=W}$ und $S_{a_2=F}$ Zeilengruppen) bewahren die ursprüngliche maximale Entropie (d.h. den "fifty-fifty"-Fall) in Bezug aufs Zielattribut "*Klassifikation*"