Übungsblatt 5/Aufgabe 1 [ihr könnt direkt zur Lösung überspringen]

Beim Kandidateneliminationsalgorithmus muss man die Mengen S und G Schritt für Schritt so verändern, damit jede Hypothese aus dem Intervall (in anderen Worten, aus dem **Versionenraum**) $[S; G]$ 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:

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

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

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 einer Inklusion folgender Art:

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

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

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

Einige Beispiele:

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

Interessanter Hinweis: man kann praktisch **ein Gitter aus allen Hypothesen** aufbauen:

> *unterste Ebene* **[Basis]** wird nur die kleinste Hypothese $(\emptyset, \emptyset, \emptyset, \emptyset, \emptyset)$ 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 echt ein schönes Muster auf. Auch die Art der Vernetzung (**m:n** Beziehungen) ist bemerkenswert.

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

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

1. für jedes solche Element aus S:
   PS1: wir **generalisieren** dieses Element **zu einer kleinstmöglichen Hypothese**, die von diesem Beispiel erfüllt wird.
   PS2: dabei stellen wir sicher, dass es keine miteinander vergleichbaren Elementpaare durch die PS1-Generalisierungen entstanden sind – wenn aber schon, dass müssen wir die S vom unnötigen Kram reinigen, das heißt, **aus jedem vergleichbaren Paar $\boldsymbol{h_1 \subseteq h_2}$** 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 S ihrer Definition entspricht (sonst wird unser Versionenraum $[S; G]$ verletzt – überlegt, warum?), das heißt, dass es **für jedes Element $\boldsymbol{h_S}$ aus S so ein Element $\boldsymbol{h_G}$ aus G gibt, dass $\boldsymbol{h_S \subseteq h_G}$ – wenn nicht, dann solche Elemente aus S entfernen.**

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

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

bei **NG1** werden erfüllte Elemente spezialisiert *(bis die vom Negativbeispiel nicht erfüllbar werden - solche Spezialisierungen sind nicht einzigartig und von daher **muss man alle möglichen Spezialisierungen bilden**)*,
bei **NG2** werden die spezielleren entfernt und
bei **NG3** werden alle Elemente h_G mit der Eigenschaft $\nexists h_S \in S: h_S \subseteq h_G$ aus G entfernt.

Bei **NS** werden alle Elemente, die vom negativen Beispiel erfüllt werden, aus S entfernt.

Zum Schluss fassen wir unsere Mengen in einem Mengenintervall $[S; G]$ zusammen, das aus einem Element (der beste 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 angegebenen Lerndaten (in unserem Fall – mit 5 Beispielsdaten) konsistent.

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

Und jetzt geht es mit der tatsächlichen **Lösung** los:

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***

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: $G = \begin{Bmatrix} (?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy), (?, Honda, ?, ?, Economy), (?, Toyota, ?, ?, Economy), \\ (?, ?, Blue, ?, Economy), (?, ?, Green, ?, Economy), (?, ?, ?, 1970, Economy), (?, ?, ?, 1990, Economy) \end{Bmatrix}$

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

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***