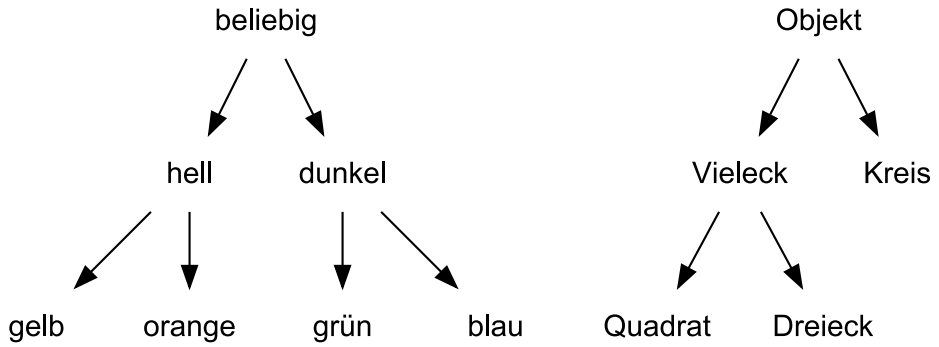
Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2015/2016
3. Übungsblatt

Aufgabe 1 Version Space, Generalisierung und Spezialisierung

Gegeben sei folgende Hierarchie von Begriffen:



Beobachtet werden Objekte, die durch Begriffspaare charakterisiert werden, die man an der untersten Ebene dieser Taxonomien finden kann (also z.B. “blaues Dreieck”). Konzepte können auch höherliegende Begriffe verwenden (also z.B. “dunkles Vieleck”). Überlegen Sie sich eine Generalisierungsvorschrift, die diese Taxonomien verwendet.

a) Wie sieht die minimale Generalisierung der Objekte “blauer Kreis” und “grünes Dreieck” aus?

b) Wie sehen minimale Spezialisierungen des Konzepts “helles Objekt” aus, sodaß das Beispiel “oranger Kreis” nicht mehr abgedeckt wird?

c) Gegeben seien folgende S und G-Sets:

G: { dunkles Vieleck, beliebiges Quadrat }
S: { blaues Quadrat }

Skizzieren Sie den Version Space, der durch diese Mengen definiert wird.

d) Wie würden Sie mit Hilfe des oben gegebenen Version Spaces die folgenden Beispiele klassifizieren (mit Begründung):

Objekt	**Klasse**
blaues Quadrat	
blauer Kreis	
blaues Dreieck	

e) Gegeben seien wiederum die S- und G-sets aus c). Wie verändern sich die Sets nach Eintreffen des Beispiels

gelbes Dreieck +

(Betrachten Sie den Candidate Elimination Algorithmus.)

Wie interpretieren Sie dieses Ergebnis?

Aufgabe 2 Version Space und Candidate Elimination Algorithmus

Überlegen Sie sich eine geeignete Sprache, um den Candidate Elimination Algorithmus um die Behandlung von numerischen Daten zu erweitern.

a) Wie sieht eine passende Generalisierungs/Spezialisierungsvorschrift aus?

b) Berechnen Sie den Version Space für folgende Beispiele:

Nr.	**A1**	**A2**	**Klasse**
1	0.5	1.5	−
2	1.1	1.2	+
3	1.8	1.0	+
4	1.5	2.1	−
5	2.1	1.2	−

c) Skizzieren Sie das S-Set, das G-Set, und den Version Space im $\mathbb{R}^2$.

d) Vertauschen Sie die Rolle der positiven und negativen Beispiele. Was passiert dann?

Aufgabe 3 Candidate Elimination Algorithmus

Gegeben sei ein Datensatz mit drei Attributen:

Haarfarbe: *blond*, *braun*, *schwarz*
Größe: *klein*, *groß*
Augenfarbe: *grün*, *blau*

Der Hypothesenraum besteht aus Disjunktionen (Oder-Verknüpfungen) von maximal einem Wert pro Attribut, einer speziellsten Theorie *false*, die keine Beispiele abdeckt, und einer allgemeinsten Theorie *true*, die alle Beispiele abdeckt.

Zum Beispiel deckt die Hypothese *blond* $\vee$ *blau* alle Personen ab, die entweder blond oder blauäugig sind (in der Datenmenge aus Aufgabe 3b) sind das z.B. die Beispiele 1, 3, 4).

Beachte: Hypothesen wie *blond* $\vee$ *braun*, die mehrere Werte desselben Attributs verwenden, sind nicht im Hypothesenraum.

a) Geben Sie in dieser Hypothesensprache alle minimalen Generalisierungen und Spezialisierungen der Hypothese *blond* $\vee$ *blau* an.

b) Folgende Beispiele treffen in dieser Reihenfolge ein:

1	*braun*	*groß*	*blau*	+
2	*braun*	*klein*	*grün*	−
3	*schwarz*	*klein*	*blau*	−
4	*blond*	*klein*	*grün*	+

Das erste Beispiel kodiert also eine Person, die braune Haare und blaue Augen hat und groß ist.

Führen Sie auf diesen Beispielen den Candidate-Elimination Algorithmus zur Berechnung des Version Spaces durch und geben Sie nach jedem Schritt das S-Set und das G-Set an.