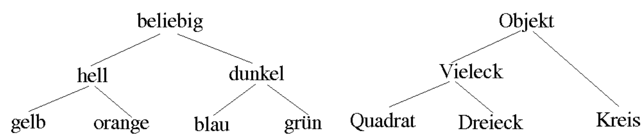
Maschinelles Lernen: Symbolische Ansätze

Musterlösung für das 3. Übungsblatt

Aufgabe 1

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.

1. Wie sieht die minimale Generalisierung der Objekte "blauer Kreis" und "grünes Dreieck" aus?

   Lösung: Gemeint war hier eine Generalisierung der beiden Objekten zusammen. Geht man von "grün" und "blau" den kürzesten Weg nach oben, landet man bei "dunkel". Geht man von "Kreis" und "Objekt" den minimalen Schritt nach oben, landet man bei "Objekt".

   $\Rightarrow$ "dunkles Objekt"

2. Wie sehen minimale Spezialisierungen des Konzepts "helles Objekt" aus, sodaß das Beispiel "oranger Kreis" nicht mehr abgedeckt wird?

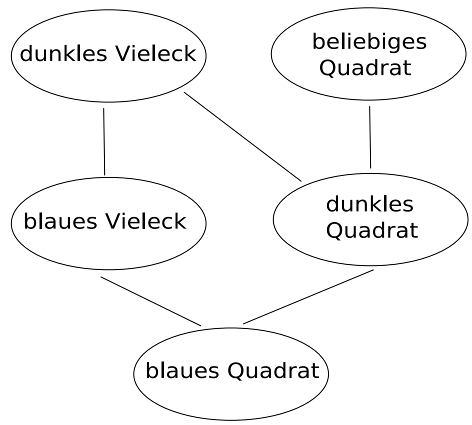
   Lösung: Entweder bei "hell" eine Ebene weiter runter (ohne bei "orange" zu landen) oder bei "Objekt" (ohne bei "Kreis" zu landen):

   "gelbes Objekt" und "helles Vieleck". "Gelbes Vieleck" ist keine zulässige Lösung, da diese Spezialisierung nicht minimal ist.

3. 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.



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

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

Lösung: Das erste Beispiel ist im S-set enthalten und wird daher als positiv klassifiziert. Das zweite Beispiel ist nicht im G-set enthalten und wird daher als negativ klassifiziert. Beispiel 3 ist ein “dunkles Vieleck”, aber kein “beliebiges Quadrat”. Im S-set ist es nicht enthalten. Daher würde es nicht klassifizierbar sein und ein “?” erhalten. (Vergleiche Folie 29)

5. Gegeben seien wiederum die S- und G-sets aus Punkt 3. Wie verändern sich die Sets nach Eintreffen des Beispiels

gelbes Dreieck $+$

Lösung: Betrachtet wird der Candidate Elimination Algorithm für das Auftreten eines positiven Beispiels. Hier wird jede Hypothese im G-set entfernt, die das Beispiel nicht abdeckt. Da “gelbes Dreieck” $\notin G$ ist, werden beide Hypothesen entfernt. Dann gilt $G = \{\}$. Da zum S-set nur die Hypothesen hinzugefügt werden, für die es eine Hypothese im G-set gibt, die genereller ist, werden keine Hypothesen hinzugefügt (da das G-set leer ist). Außerdem wurde das aktuelle Beispiel nicht abgedeckt. Daher gilt $S = \{\}$.

Wie interpretieren Sie dieses Ergebnis?

Mit der Hinzunahme des Beispiels “gelbes Dreieck” + ist das Konzept mit dem vorliegenden Version Space nicht mehr lernbar. Wenn man sich die Beispiele betrachten würde, die zur Entstehung des Version Spaces geführt haben, und das aktuelle Beispiel hinzunehmen würde, so wäre diese Beispielmenge mit dem Candidate Elimination Algorithmus nicht mehr lernbar.

Aufgabe 2

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

1. Wie sieht eine passende Generalisierungs/Spezialisierungsvorschrift aus?

   Lösung: Wir verwenden als Hypothesensprache Intervalle. Bei Generalisierungen sind diese abgeschlossen und bei Spezialisierungen entsprechend offen. Demnach sind die Hypothesen des G-Sets offen beziehungsweise die des S-Sets abgeschlossen.

   Spezialisierungen erfolgen durch Hinzunahme beziehungsweise Einschränkung von Intervallen. Generalisierungen entstehen durch Wegfallen beziehungsweise Erweiterung von Intervallen.

2. 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	$-$

   Lösung: Wir verwenden den wie oben beschrieben modifizierten Candidate Elimination Algorithmus.

   (a) Wir beginnen mit:
   - $G_0 = \{<?, ?>\}$
   - $S_0 = \{<\emptyset, \emptyset>\}$

   (b) Als nächsten erhalten wir folgendes Beispiel (0.5, 1.5, $-$):
   - $S_1 = S_0$, da keine Hypothese in S_0 das Beispiel abdeckt
   - $\{<?, ?>\}$ deckt das Beispiel ab, deshalb muß diese Hypothese entfernt und minimal spezialisiert werden. Alle möglichen minimalen Spezialisierung werden hinzugefügt, da G_0 die generellste Hypothese beinhaltet.
     $\Rightarrow G_1 = \{<(-\infty, 0.5), ?>, <(0.5, \infty), ?>, <?, (-\infty, 1.5)>, <?, (1.5, \infty)>\}$

   (c) Nächstes Beispiel (1.1, 1.2, $+$):
   - Wir entfernen alle Hypothesen aus G_1, die das Beispiel nicht abdecken: $<(-\infty, 0.5), ?>, <?, (1.5, \infty)>\}$
     $\Rightarrow G_2 = \{<(0.5, \infty), ?>, <?, (-\infty, 1.5)>\}$
   - $<\emptyset, \emptyset>$ aus S_1 deckt das Beispiel nicht ab und muß generalisiert werden:
     $\Rightarrow S_2 = \{<[1.1, 1.1], [1.2, 1.2]>\}$ (alle Hypothesen in G_2 sind genereller)

(d) Nächstes Beispiel (1.8, 1.0, +):

- Alle Hypothesen aus G_2 decken das Beispiel ab:
  $\Rightarrow G_3 = G_2$
- $< [1.1, 1.1], [1.2, 1.2] >$ aus S_2 deckt das Beispiel nicht ab und muß generalisiert werden:
  $\Rightarrow S_3 = \{< [1.1, 1.8], [1.0, 1.2] >\}$

(e) Nächstes Beispiel (1.5, 2.1, −):

- Keine Hypothese in S_3 deckt das Beispiel ab:
  $\Rightarrow S_4 = S_3$
- Wir spezialisieren alle Hypothesen in G_3, die das Beispiel abdecken: $< (0.5, \infty), ? >$
  Mögliche Spezialisierungen werden hinzugefügt, falls eine Hypothese in G_3 spezifischer ist:
  - $< (0.5, 1.5), ? >$, keine Hypothese ist spezifischer
  - $< (1.5, \infty), ? >$, keine Hypothese ist spezifischer
  - $< (0.5, \infty), (-\infty, 2.1) >$, $< [1.1, 1.8], [1.0, 1.2] >$ ist spezifischer
  - $< (0.5, \infty), (2.1, \infty) >$, keine Hypothese ist spezifischer

  Alle anderen Hypothesen in G_3 bleiben unverändert.
  $\Rightarrow G_4 = \{< (0.5, \infty), (-\infty, 2.1) >, <?, (-\infty, 1.5) >\}$

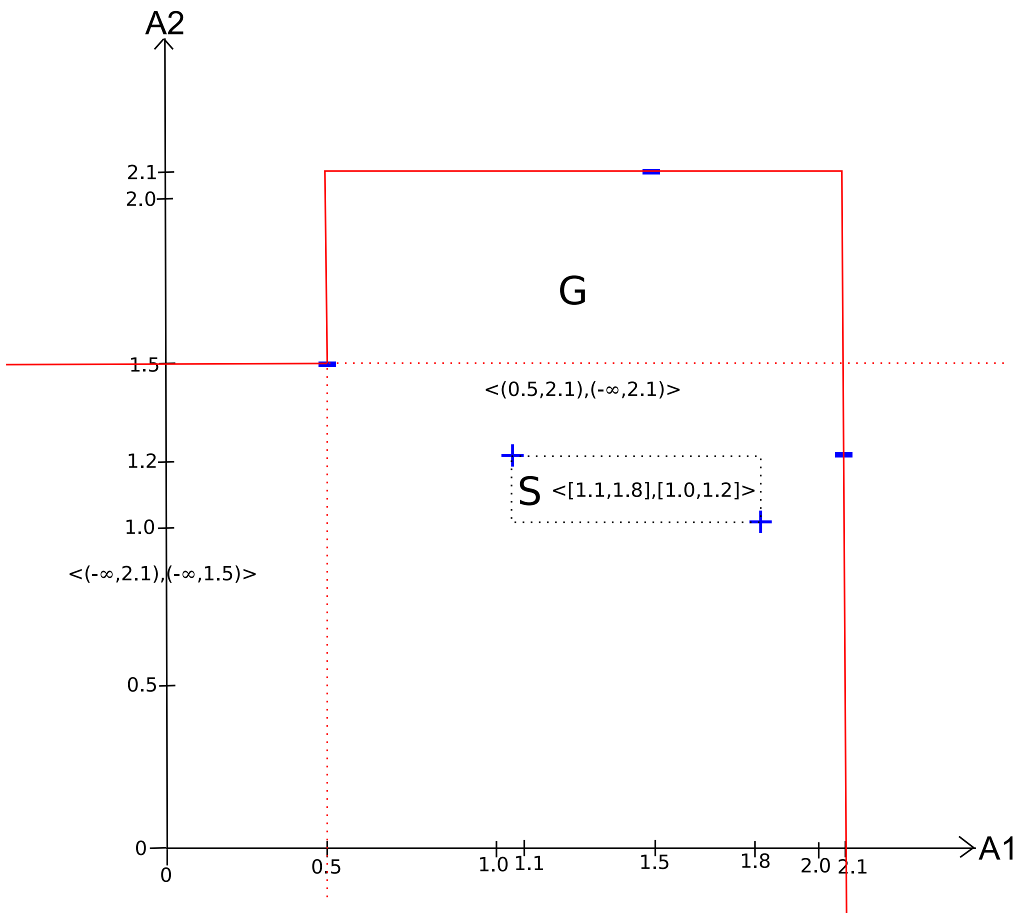
(f) Nächstes Beispiel (2.1, 1.2, −):

- Keine Hypothese in S_4 deckt das Beispiel ab:
  $\Rightarrow S_5 = S_4$
- Wir spezialisieren alle Hypothesen in G_4, die das Beispiel abdecken: $< (0.5, \infty), (-\infty, 2.1) >, <?, (-\infty, 1.5) >$
  Betrachten wir die möglichen Spezialisierungen der ersten Hypothese:
  - $< (0.5, 2.1), (-\infty, 2.1) >$, $< [1.1, 1.8], [1.0, 1.2] >$ ist spezifischer
  - $< (2.1, \infty), (-\infty, 2.1) >$, keine Hypothese ist spezifischer
  - $< (0.5, \infty), (-\infty, 1.2) >$, keine Hypothese ist spezifischer
  - $< (0.5, \infty), (1.2, 2.1) >$, keine Hypothese ist spezifischer

  Die möglichen Spezialisierungen der zweiten Hypothese lauten wie folgt:
  - $< (-\infty, 2.1), (-\infty, 1.5) >$, $< [1.1, 1.8], [1.0, 1.2] >$ ist spezifischer
  - $< (2.1, \infty), (-\infty, 1.5) >$, keine Hypothese ist spezifischer
  - $<?, (-\infty, 1.2) >$, keine Hypothese ist spezifischer
  - $<?, (1.2, 1.5) >$, keine Hypothese ist spezifischer

  $\Rightarrow G_5 = \{< (0.5, 2.1), (-\infty, 2.1) >, < (-\infty, 2.1), (-\infty, 1.5) >\}$

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



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

   (a) Wir beginnen mit:
      - $G_0 = \{<?, ? >\}$
      - $S_0 = \{< \emptyset, \emptyset >\}$

   (b) Als nächsten erhalten wir folgendes Beispiel (0.5, 1.5, +):
      - Alle Hypothesen aus G_0 decken das Beispiel ab:
        $\Rightarrow G_1 = G_0$
      - $< \emptyset, \emptyset >$ aus S_0 deckt das Beispiel nicht ab und muß generalisiert werden:
        $\Rightarrow S_1 = \{< [0.5, 0.5] , [1.5, 1.5] >\}$

   (c) Nächstes Beispiel (1.1, 1.2, $-$):

- Keine Hypothese in S_1 deckt das Beispiel ab:
  $\Rightarrow S_2 = S_1$
- $\{<?, ?>\}$ aus G_1 deckt das Beispiel ab und muß spezialisiert werden:
  - $<(-\infty, 1.1), ?>$, $<[0.5, 0.5], [1.5, 1.5]>$ ist spezifischer
  - $<(1.1, \infty), ?>$, keine Hypothese ist spezifischer
  - $<?, (-\infty, 1.2)>$, keine Hypothese ist spezifischer
  - $<?, (1.2, \infty)>$, $<[0.5, 0.5], [1.5, 1.5]>$ ist spezifischer

  $\Rightarrow G_2 = \{<(-\infty, 1.1), ?>, <?, (1.2, \infty)>\}$

(d) Nächstes Beispiel (1.8, 1.0, $-$):

- Keine Hypothese in S_2 deckt das Beispiel ab:
  $\Rightarrow S_3 = S_2$
- Keine Hypothese in G_2 deckt das Beispiel ab:
  $\Rightarrow G_3 = G_2$

(e) Nächstes Beispiel (1.5, 2.1, $+$):

- Wir entfernen alle Hypothesen aus G_3, die das Beispiel nicht abdecken: $<(-\infty, 1.1), ?>$
  $\Rightarrow G_4 = \{<?, (1.2, \infty)>\}$
- $<[0.5, 0.5], [1.5, 1.5]>$ aus S_3 deckt das Beispiel nicht ab und muß generalisiert werden:
  $\Rightarrow S_3 = \{<[0.5, 1.5], [1.5, 2.1]>\}$

(f) Nächstes Beispiel (2.1, 1.2, $+$):

- Wir entfernen alle Hypothesen aus G_4, die das Beispiel nicht abdecken: $<?, (1.2, \infty)>$
  $\Rightarrow G_5 = \emptyset$
- $<[0.5, 1.5], [1.5, 2.1]>$ aus S_4 deckt das Beispiel nicht ab und muß generalisiert werden. Es können aber keine gültigen Generalisierungen gefunden werden, da $G_5 = \emptyset$ (siehe Skript Folie 26: "some hypothesis...").
  $\Rightarrow S_5 = \emptyset$

Da sowohl das G-set und das S-set leer sind folgt, dass diese Beispielmenge nicht mit dem Algorithmus lernbar ist.