2. Die Generalisierungs-Halbordnung

Gewinnung von Wissen aus Daten:
beinhaltet Schritt der *Verallgemeinerung* (*Generalisierung*)

- im logischen Sinne (deterministisch)

oder

- im statistischen Sinne (auf Basis von Häufigkeits- bzw. Wahrscheinlichkeitsaussagen)

Im Data Mining werden z.T. auch beide Herangehensweisen vermischt

in diesem Kapitel: logische Generalisierung

"x verallgemeinert y" ist eine Relation:
Theorie der Relationen, der partiellen Ordnungen anwendbar

2.1. Mathematische Grundlagen aus der Ordnungstheorie

- *Ordnungsrelation*: nach Bourbaki einer der zentralen Strukturtypen der Mathematik
- eng verwandt mit algebraischen Strukturen

(zweistellige) *Relation R* auf einer Menge *M*:

Menge von Paaren (a, b) mit $a, b \in M$
d.h. Teilmenge des kartesischen Produkts $M \times M$ (= Menge aller Paare von Elementen aus *M*)

für (a, b) $\in R$ schreibt man $a\ R\ b$

Beispiele für Relationen:
"ist kleiner als", "ist Teilmenge von", "ist Teiler von", "hat nichtleeren Durchschnitt mit", "ist Verallgemeinerung von"

Konverse Relation R^{-1}: alle Paare werden gespiegelt,
also $R^{-1} = \{ (b, a) \mid (a, b) \in R \}$.

Besondere Eigenschaften von Relationen:

R heißt ... falls für alle $a, b, c \in M$ gilt ... :

reflexiv	$a \, R \, a$
irreflexiv	nicht $a \, R \, a$
symmetrisch	$a \, R \, b \Rightarrow b \, R \, a$
antisymmetrisch (oder identitiv)	$a \, R \, b$ und $b \, R \, a \Rightarrow a = b$
asymmetrisch	$a \, R \, b \Rightarrow$ nicht $b \, R \, a$
transitiv	$a \, R \, b$ und $b \, R \, c \Rightarrow a \, R \, c$
total (oder linear)	$a \, R \, b$ oder $b \, R \, a$
konnex	$a \, R \, b$ oder $b \, R \, a$ oder $a = b$

wichtige Kombinationen von Eigenschaften:

R Halbordnung (auch: partielle Ordnung; *partial order*)	R reflexiv, antisymmetrisch und transitiv
R totale Ordnung (auch: lineare Ordnung, Vollordnung; *total order*)	R antisymmetrisch, transitiv und total (dann auch reflexiv)
R Äquivalenzrelation	R reflexiv, symmetrisch und transitiv

(M, R) heißt *partiell geordnete Menge*, wenn R Halbordnung auf M ist.
Dann oft Notation $\leq$ statt R.

Hilfssatz:
(a) Jede partielle Ordnung $\leq$ auf M definiert eine Relation $<$ auf M gemäß $a < b \;\; :\Leftrightarrow \;\; a \leq b$ und $a \neq b$,
die irreflexiv, asymmetrisch und transitiv ist.

(b) Jede irreflexive und transitive Relation $<$ auf M ist auch asymmetrisch und definiert eine partielle Ordnung $\leq$ auf M gemäß $a \leq b \;:\Leftrightarrow\; a < b$ oder $a = b$.

(c) Die durch (a) und (b) gegebene Beziehung zwischen Relationen ist bijektiv, d.h. jede partielle Ordnung $\leq$ in reflexiver Schreibweise korrespondiert umkehrbar eindeutig mit einer partiellen Ordnung $<$ in irreflexiver Schreibweise.

(d) $<$ konnex $\Leftrightarrow$ $\leq$ total.

Dualitätsprinzip für partiell geordnete Mengen:

- Mit R (bzw. $\leq$) ist stets auch R^{-1} (bzw. $\geq$) eine partielle Ordnungsrelation auf M, die duale Ordnung zu R.
- Zu einer Aussage A, die außer rein logischen Bestandteilen nur das Zeichen $\leq$ enthält bekommt man die duale Aussage D(A), wenn man in A das Zeichen $\leq$ durch $\geq$ ersetzt.
- A gilt genau dann in einer partiell geordneten Menge, wenn $D(A)$ in der dualen partiell geordneten Menge gilt.
- Behauptet ein Lehrsatz zwei zueinander duale Aussagen, so genügt es, nur die eine zu beweisen; die andere ergibt sich "dual", d.h. mit dem gleichen Beweis für die duale Ordnung.

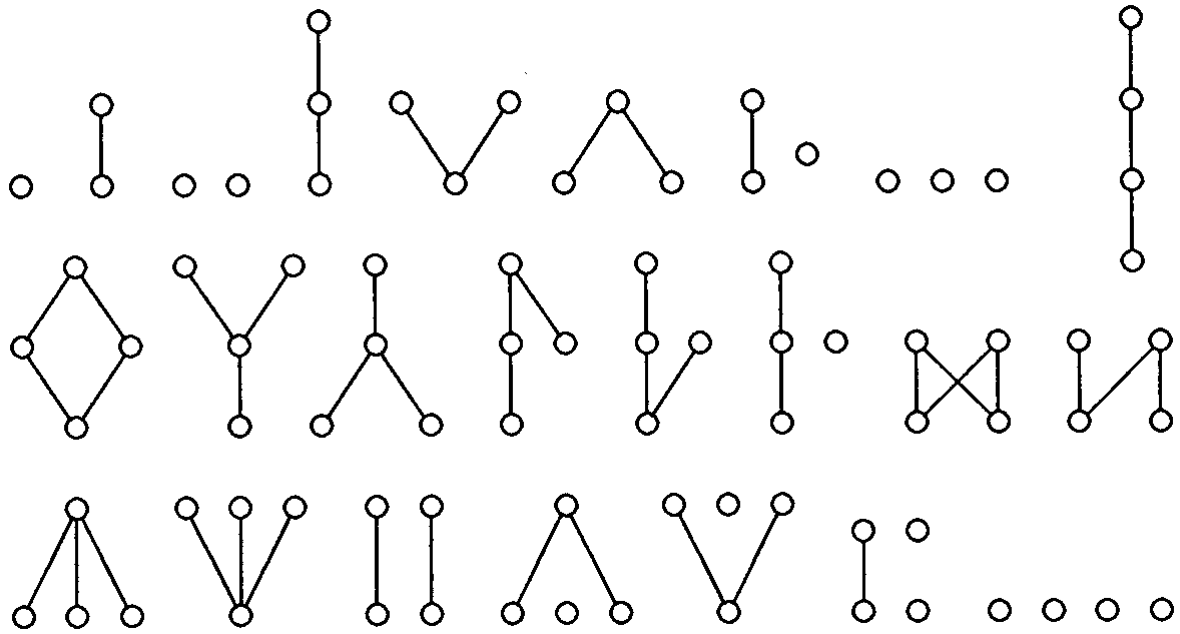
In einer partiell geordneten Menge nennt man a einen *unteren Nachbarn* von b (und b einen oberen Nachbarn von a; auch: b bedeckt a),
wenn $a < b$ ist und kein Element c existiert mit $a < c < b$.

Liniendiagramm (Hasse-Diagramm) einer endlichen partiell geordneten Menge:

- Elemente dargestellt als kleine Kreise in der Ebene
- ist a unterer Nachbar von b, so wird der b entspr. Kreis oberhalb des a entspr. Kreises aufgetragen (seitliche Verschiebung zugelassen) und es werden beide Kreise durch eine Linie verbunden

- Dann gilt: $a < b \Leftrightarrow$ der Kreis, der *b* darstellt, ist von dem *a* darstellenden Kreis durch einen aufsteigenden Linienzug erreichbar.

Liniendiagramme aller möglichen partiell geordneten Mengen mit bis zu 4 Elementen (aus Ganter & Wille 1996):

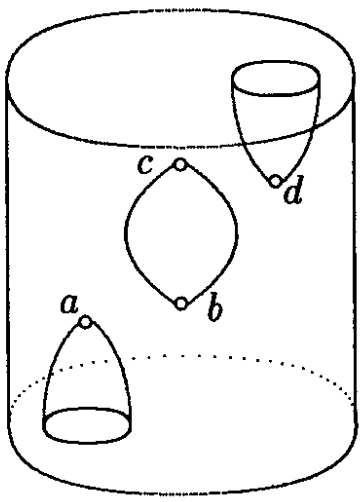


Definitionen:

- 2 Elemente *a*, *b* einer partiell geordneten Menge heißen *vergleichbar*, falls $a \leq b$ oder $b \leq a$, sonst *unvergleichbar*.
- Eine Teilmenge, in der je 2 Elemente stets vergleichbar sind, heißt eine *Kette*.
- Eine Teilmenge, in der je 2 Elemente stets unvergleichbar sind, heißt eine *Antikette*.
- Die *Weite* einer endlichen geordneten Menge ist die maximale Mächtigkeit einer Antikette.
- Die *Länge* einer endlichen geordneten Menge ist die maximale Mächtigkeit einer Kette minus Eins.

Sind a, b, c, d Elemente einer partiell geordneten Menge, so definiert man:

- das *Intervall*
  $[b, c] := \{ x \mid b \leq x \leq c \}$
- das *Hauptideal* (oder uneigentliches, nach unten offenes Intervall)
  $(a] := \{ x \mid x \leq a \}$
- den *Hauptfilter* (oder uneigentliches, nach oben offenes Intervall)
  $[d) := \{ x \mid x \geq d \}$.



Verallgemeinertes Intervall: statt einzelner Elemente b, c werden für die Begrenzung auch Mengen zugelassen
$[B, C] := \{ x \mid \exists\, b \in B, c \in C: b \leq x \leq c \}$

Sei $(M, \leq)$ partiell geordnete Menge, $T \subseteq M$.
$s \in M$ *untere Schranke* (*lower bound*) von $T :\Leftrightarrow \forall t \in T: s \leq t$.

Dual def. man: *obere Schranke* (*upper bound*).

$a \in T$ *minimales Element* von $T :\Leftrightarrow \forall t \in T: t \leq a \Rightarrow t = a$.

Dual: *maximales Element*.

Wenn ein Element $d \in M$ existiert, das maximales Element der Menge aller unteren Schranken von T ist, so heißt d das

Infimum von T, geschrieben: $d = \inf T$ oder $d = \bigwedge T$. ("größte untere Schranke von T")

Dual: *Supremum* sup T, $\bigvee T$. ("kleinste obere Schranke von T")

Eine partiell geordnete Menge heißt *Verband* (*lattice*), wenn für jede 2-elementige Teilmenge ein Infimum und ein Supremum existieren.
(*vollständiger* Verband, wenn für jede nichtleere Teilmenge ein Infimum und ein Supremum existieren.)

Bezeichnung: inf $\{a, b\}$ = inf$(a, b) = a \wedge b$
(gemeint ist hier nicht das logische "und"!)
sup $\{a, b\}$ = sup$(a, b) = a \vee b$

2.2. Die Generalisierungs-Halbordnung und der Versionenraum

Wir betrachten wieder einen Datenraum aus n-Tupeln; die Komponenten sind die *Attribute*:

Datenraum $\Omega = X_1 \times X_2 \times ... \times X_n$

Die Elemente heißen *Instanzen*.

Aufgabe des "Begriffslernens" (*concept learning*):

In Ω habe eine Komponente, z.B. X_n , nur 2 Werte ("yes" und "no").
Gegeben sei eine (kleine) Teilmenge D von Ω: die Menge der *Trainingsbeispiele*.
Gesucht ist eine boolesche Funktion (die "Zielfunktion")

$$X_1 \times X_2 \times ... \times X_{n-1} \rightarrow \{\text{yes, no}\},$$

so dass der Funktionswert die Belegung von X_n auf Ω vorhersagt.

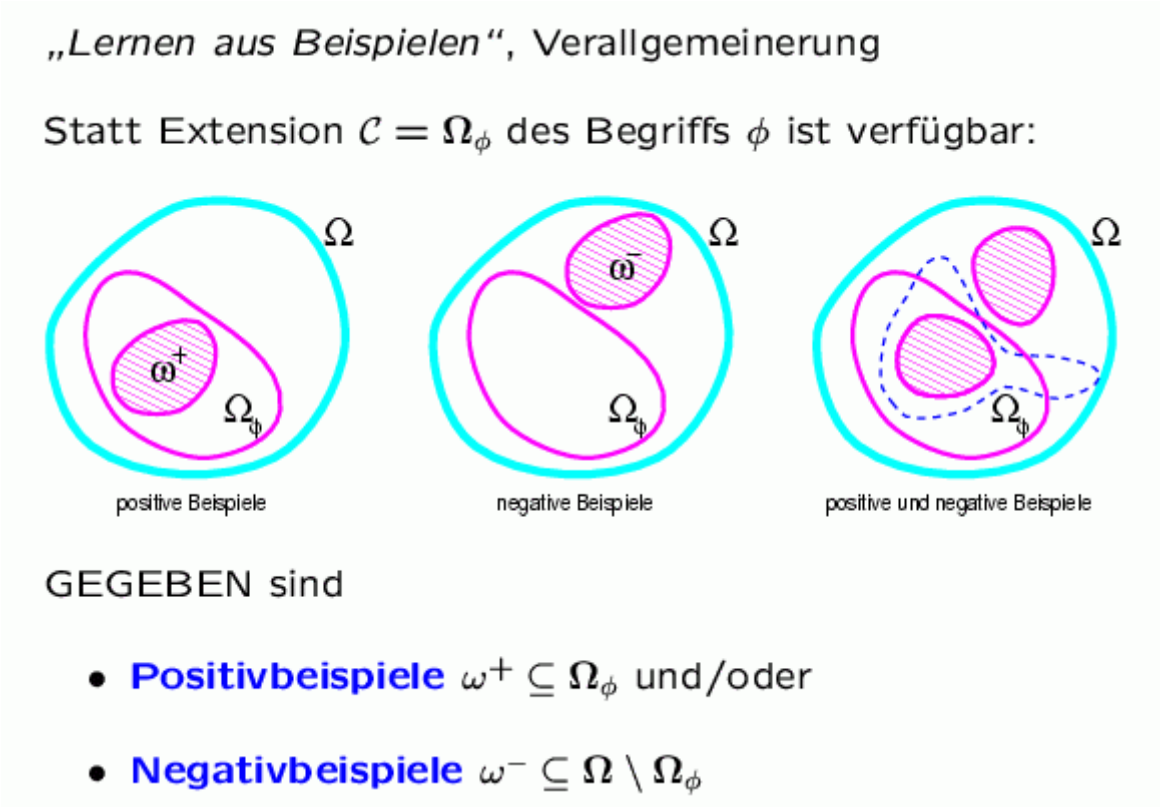
Beispiel:
Trainingsmenge D

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

(aus Mitchell 1997)

Die Teilmenge Ω_ϕ von Ω, für die die *n*-te Komponente "Yes" ist, kann als Extension (Umfang) eines *Begriffs* interpretiert werden. (zu "Begriffen" später mehr!)
Hier: Begriff "gute Tage zum Sport-Treiben".

Die Trainingsmenge *D* unterteilt sich in Positiv- und Negativbeispiele.



D = $\omega^+ \cup \omega^-$.

Grundannahme des induktiven Lernens (inductive learning hypothesis):

> Eine Funktion, die die Zielfunktion auf einer genügend großen Menge von Trainingsbeispielen gut approximiert, wird sie auch auf anderen, unbeobachteten Beispieldaten (und somit auf ganz Ω) gut approximieren.

Probleme beim induktiven Lernen:

Übergeneralisierung

- EFFEKT — es werden *Oberbegriffe* von ϕ gelernt
- ABHILFE — Negativbeispiele bereitstellen

Überspezialisierung

- EFFEKT — es werden *Unterbegriffe* von ϕ gelernt
- ABHILFE — *repräsentative* Positivstichprobe

Fehlgranulation

- EFFEKT — *Überanpassung* oder *Unteranpassung*
- ABHILFE — adäquates Sortiment von Intensionen

Natürlichkeit, Fortsetzbarkeit

- EFFEKT — gelernte Verallgemeinerung versagt bei Wiederabruf (Prädiktion, Extrapolation)
- ABHILFE — *Occam's razor*: einfachste Erklärung

(aus Schukat-Talamazzini 2002).

Als "Bausteine" der zu approximierenden Zielfunktion werden *Hypothesen* verwendet.
Das können grundsätzlich beliebige Teilmengen von Ω sein (bzw. gleichwertig: boolesche Funktionen auf Ω).

Oft: Einschränkung auf *konjunktiv verknüpfte* Attributbelegungen als Hypothesen.

Objekte z.B.

$$(sunny, warm, normal, strong, warm, same)$$

Hypothesen $\widehat{=}$ *partielle Attributbelegungen* z.B.

$$(sunny, ?, ?, strong, ?, ?)$$

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 $\boldsymbol{x} \in \Omega$ *„genügt" der Hypothese* $h \in \mathcal{H}$ *(*$\boldsymbol{x}$ **erfüllt** h *bzw.* $h \models \boldsymbol{x}$*) 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.

Hypothesenhalbordnung

Jede Hypothese h ist durch die Menge

$$\Omega(h) \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \Omega \mid h \models \boldsymbol{x}\}$$

aller passenden Objekte charakterisiert.

$\mathcal{H}$ erbt von $\mathfrak{P}\Omega$ die *Mengenrelation*

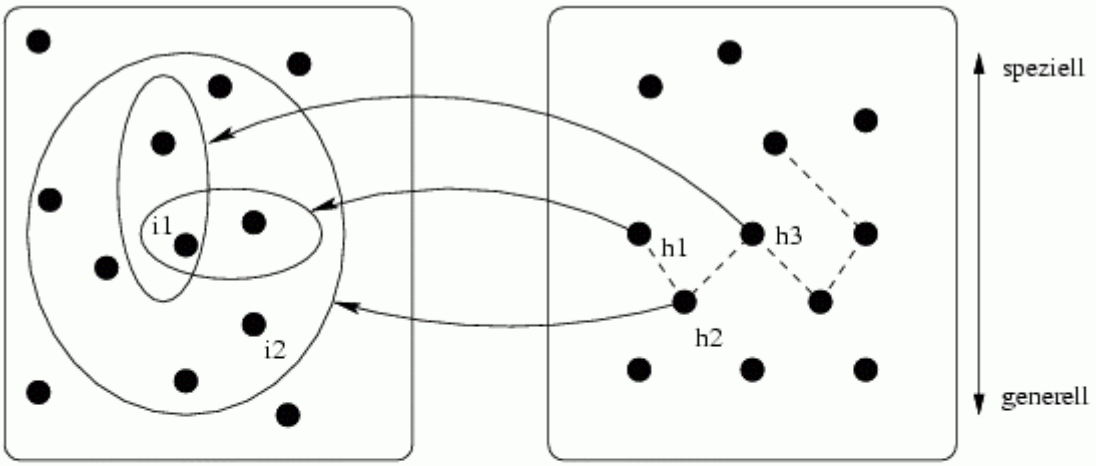
$$h \supseteq h' \quad \text{gdw.} \quad \forall \boldsymbol{x} \in \Omega: \; (h' \models \boldsymbol{x} \Rightarrow h \models \boldsymbol{x})$$

(h ist „allgemeiner" oder „genereller" als h')

($\wp\Omega$: Potenzmenge von Ω, d.h. Menge aller Teilmengen)

auch $h \leq h'$ für "*h* allgemeiner als *h*' " (allgemeinere Hypothesen stehen weiter "unten")

Mit Einschluss von (∅, ..., ∅) ist *H* im Fall, dass alle Wertemengen der Attribute endlich sind, sogar ein Verband.



i_1 : $(sunny, warm, high, strong, cool, same)$
i_2 : $(sunny, warm, high, light, warm, same)$

h_1 : $(sunny, ?, ?, strong, ?, ?)$
h_2 : $(sunny, ?, ?, ?, ?, ?)$
h_3 : $(sunny, ?, ?, ?, cool, ?)$

$h_2 \supseteq h_1, h_3$

Def.:

Eine Hypothese $h \in \mathcal{H}$ heißt **konsistent** *mit den Lerndaten (ω^+, ω^-) genau dann wenn gilt:*

$$x \in \omega^+ \quad \Rightarrow \quad h \models x$$
$$x \in \omega^- \quad \Rightarrow \quad h \not\models x$$

Ein erster Hypothesenfindungsalgorithmus für das induktive Lernen:

FIND-S

findet eine "speziellste Hypothese", die konsistent mit den Positivbeispielen ist, durch sukzessive Generalisierung.

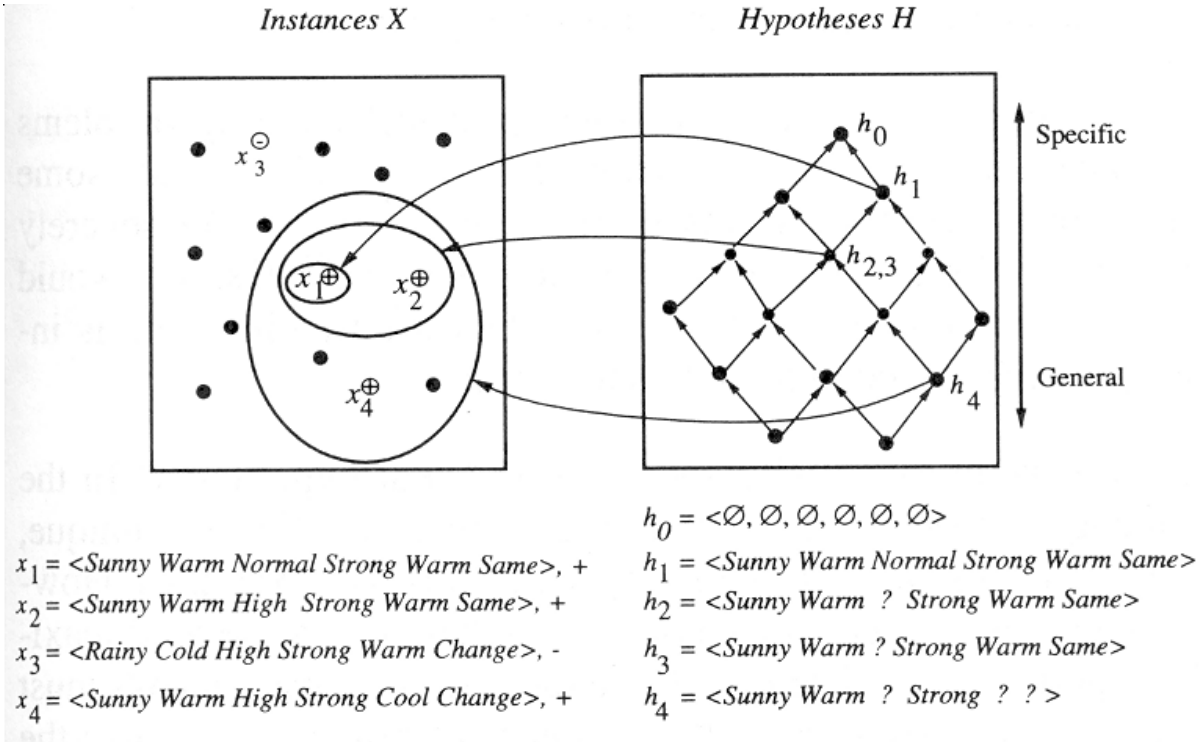
Start mit der speziellsten Hypothese überhaupt:
$h_\varnothing$ = (∅, ∅, ∅, ..., ∅).

FIND-S:
(1) Initialisierung: Setze $h := h_{\varnothing}$

(2) Generalisierung (Schleife):
für jedes positive Trainingsbeispiel $x \in \omega^+$:
h := speziellste Verallgemeinerung von h, die mit x konsistent ist, d.h. $h := h \wedge x$ (Infimum im Hypothesenraum bzgl. Generalisierungshalbordn.)
(bei negativen Trainingsbeispielen passiert nichts!)

(3) Terminierung nach Abarbeitung aller Positivbeispiele:
Das Ergebnis ist h.

Ablauf in unserem Beispiel:



- Der Algorithmus terminiert immer mit der speziellsten Hypothese in H, die mit den positiven Trainingsdaten konsistent ist.
- Wenn die Zielfunktion in H enthalten ist (also rein konjunktiv aus Attributbelegungen zusammengesetzt werden kann) und wenn die Trainingsdaten korrekt sind, ist das Ergebnis auch mit den Negativbeispielen konsistent.

Aber:

- nicht immer muss die speziellste konsistente Hypothese die richtige sein
- Inkonsistenzen (Widersprüche) in den Trainingsdaten werden von FIND-S nicht erkannt
- in manchen Fällen (reellwertige Attribute) gibt es keine eindeutige, speziellste konsistente Hypothese in H: FIND-S wäre durch Backtracking zu erweitern.

bessere Algorithmen?

erster Ansatz:
List-then-eliminate

ALGORITHMUS — Kandidatenelimination

1 *INITIALISIERUNG*
Setze $H \leftarrow \mathcal{H}$

2 *GENERALISIERUNG / SPEZIALISIERUNG*
Eliminiere für alle $x \in \omega^+ \cup \omega^-$

- *Fall* $x \in \omega^+$: *alle* $h \in H$ *mit* $h \not\models x$
- *Fall* $x \in \omega^-$: *alle* $h \in H$ *mit* $h \models x$

3 *TERMINIERUNG*
Das Ergebnis ist h, *falls* $H = \{h\}$ *ist.*

Nachteil: *H* kann riesengroß sein, vollständige Auflistung und Durchsuchen von *H* nicht praktikabel

⇒ Ordnungsstruktur von *H* besser ausnutzen!

Die Menge der mit den Lernbeispielen konsistenten Hypothesen

$$\{h \in \mathcal{H} \mid h \text{ konsistent mit } (\omega^+, \omega^-)\}$$

heißt **Versionenraum** *von* (ω^+, ω^-) *bezüglich* $\mathcal{H}$ *und wird mit* $\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-)$ *(oder* $\mathfrak{V}$*) bezeichnet.*

Der VR besitzt **minimale** Elemente

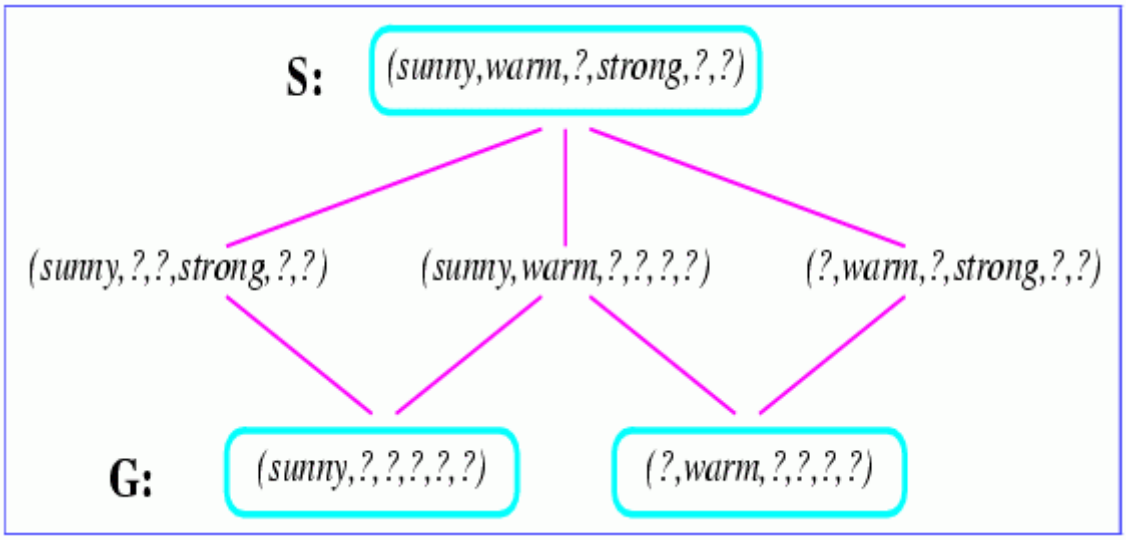
$$\mathfrak{V}_S \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V} : \; h' \subseteq h \Rightarrow h' = h\}$$

und **maximale** Elemente:

$$\mathfrak{V}_G \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V} : \; h \subseteq h' \Rightarrow h' = h\}$$

("minimale" Elemente bzgl. Inklusion $\subseteq$ sind *maximale* Elemente bzgl. $\leq$, und dual)

Beispiel: ein Versionenraum mit 6 Hypothesen



- statt des ganzen Versionenraumes nur S und G heranziehen?

BEISPIEL:
Der *vollständige* Hypothesenraum

$$\mathcal{H} \;=\; \mathfrak{P}\Omega \;=\; \{\mathcal{Y} \mid \mathcal{Y} \subseteq \Omega\}$$

bedingt Versionenräume der Gestalt:

$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) \;=\; \{h \in \mathcal{H} \mid \omega^+ \subseteq h \subseteq \Omega \setminus \omega^-\}$$

Die minimalen/maximalen Elemente

$$\mathfrak{V}_S \;=\; \{\omega^+\} \qquad \text{und} \qquad \mathfrak{V}_G \;=\; \{\Omega \setminus \omega^-\}$$

sind eindeutig und es gilt die Intervalldarstellung

$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) \;=\; [\mathfrak{V}_S, \mathfrak{V}_G]_{\mathcal{H}}$$

- Wenn wir verallgemeinerte Intervalle zulassen, gilt diese Darstellung sogar für beliebige Hypothesenräume!

Versionenraum-Darstellungssatz:

Für den Versionenraum $\mathfrak{V}$ der Beispieldaten ω^+ und ω^- bezüglich $\mathcal{H}$ gilt die Intervalldarstellung

$$\mathfrak{V} = [\mathfrak{V}_S, \mathfrak{V}_G]_{\mathcal{H}} \quad .$$

Dabei sind $\mathfrak{V}_S$ und $\mathfrak{V}_G$ die Mengen der $\subseteq$-minimalen ($\subseteq$-maximalen) Elemente des Versionenraums $\mathfrak{V}$.

(kurz: S und G.)

S: "specific boundary", Menge der *am wenigsten allgemeinen* (d.h. speziellsten) Elemente von *H*, die mit der Trainingsmenge *D* konsistent sind

G: "general boundary", Menge der allgemeinsten Elemente von *H*, die mit *D* konsistent sind

Damit: kompaktere Darstellung von Hypothesenmengen im Verlauf der Kandidatenelimination möglich.

CANDIDATE-ELIMINATION Lernalgorithmus:

führt stets 2 Mengen von Hypothesen G und S mit, die einfach spezifizierbar sind

- G wird initialisiert mit {(?, ?, ?, ..., ?)} (beschreibt Ω)
- S wird initialisiert mit {($\varnothing$, $\varnothing$, $\varnothing$, ..., $\varnothing$)} (beschreibt $\varnothing$)
- [G, S] ist am Schluss verallg. Intervalldarstellung der gesuchten Hypothesenmenge

(Intervall hier notiert bzgl. $\leq$)

ALGORITHMUS — VR-Kandidatenelimination

1 *INITIALISIERUNG*

$$G \leftarrow \{\Omega\} \quad \textit{und} \quad S \leftarrow \{\emptyset\}$$

+2 *POSITIVE BEISPIELE*

Für alle $x \in \omega^+$*:*

Entferne alle $h \in G$ *mit* $h \not\models x$
Für alle $h \in S$*:*

Generalisiere h *zu* h' *mit* $h' \models x$
Behalte $h' \in S$*, falls* h' *spezieller als* G

Entferne alle nichtminimalen $h \in S$

−2 *NEGATIVE BEISPIELE*

Für alle $x \in \omega^-$*:*

Entferne alle $h \in S$ *mit* $h \models x$
Für alle $h \in G$*:*

Spezialisiere h *zu* h' *mit* $h' \not\models x$
Behalte $h' \in G$*, falls* h' *allgemeiner als* S

Entferne alle nichtmaximalen $h \in G$

3 *TERMINIERUNG*

Das Ergebnis ist h*, falls* $G = \{h\} = S$ *ist.*

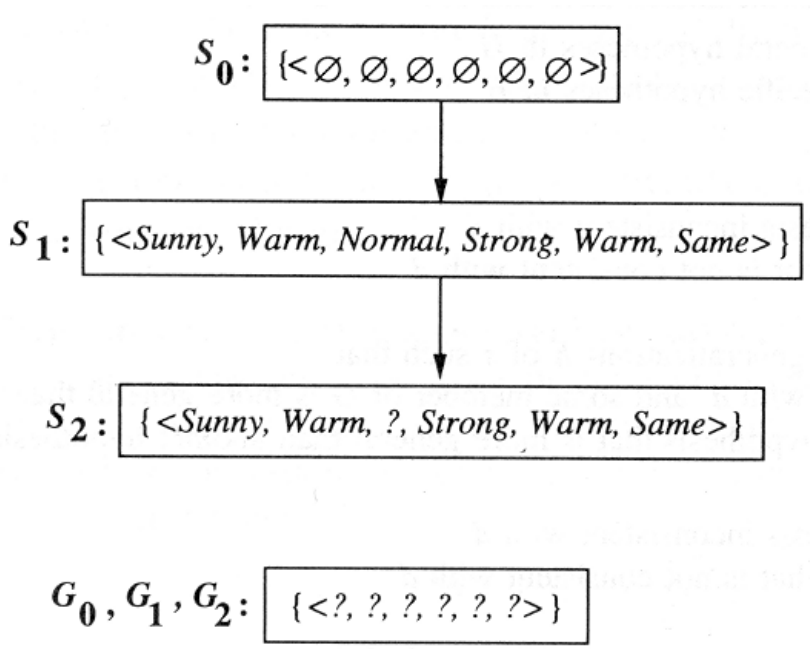
("minimal", "maximal" hier im Sinne der Inklusion)

Am Schluss enthält der berechnete Versionenraum [G, S] in jedem Fall *alle* Hypothesen aus *H*, die mit *D* konsistent sind, und *nur diese*. (Verbesserung gegenüber FIND-S!)

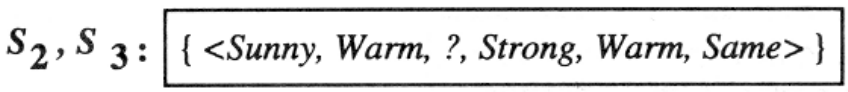
Beachte: der Algorithmus behandelt Positiv- und Negativ-beispiele in dualer Weise ("gleichberechtigt").

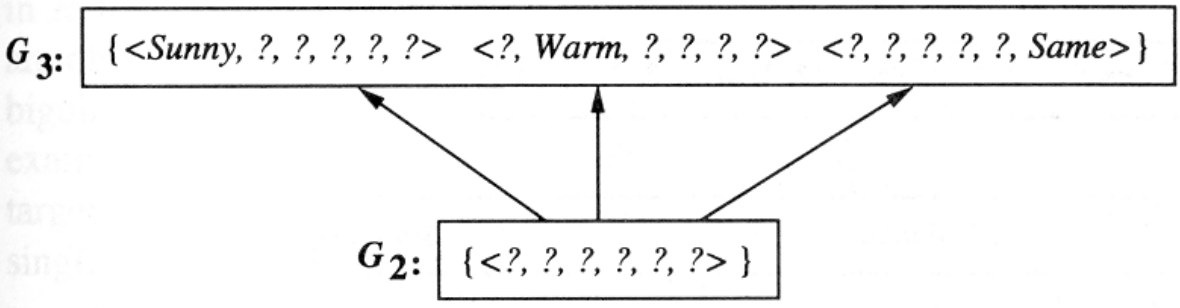
Anwendung auf unser "Sport"-Beispiel:

die ersten beiden Trainingsbeispiele werden verarbeitet.
1. <Sunny, Warm, Normal, Strong, Warm, Same, **YES**>
2. <Sunny, Warm, High, Strong, Warm, Same, **YES**>

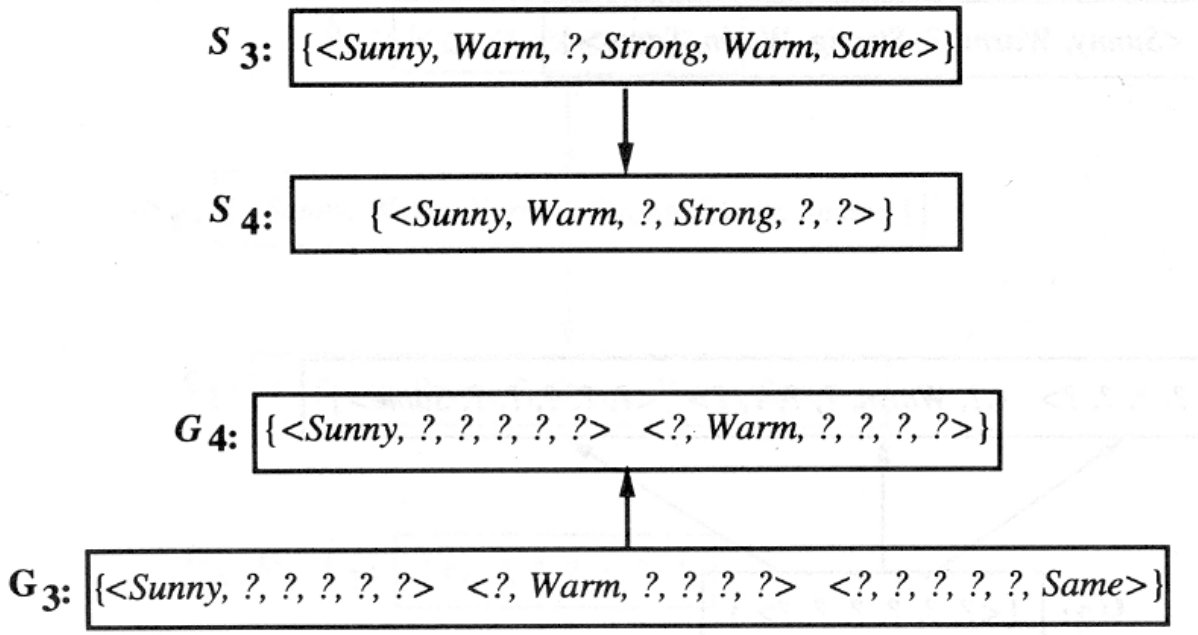


bei den ersten beiden Schritten ändert sich G nicht, aber beim dritten Schritt (erstes Negativbeispiel):
3. <Rainy, Cold, High, Strong, Warm, Change, **NO**>

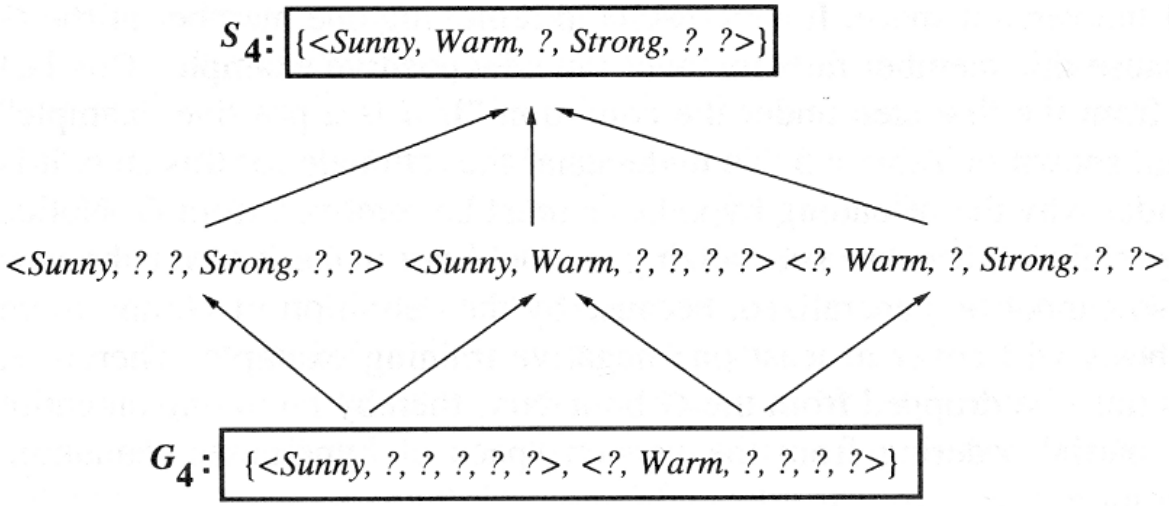




der Schritt vom 3. zum 4. Trainingsbeispiel:
4. <Sunny, Warm, High, Strong, Cool, Change, **YES**>



Endergebnis: Versionsraum für das "Sport"-Beispiel



(aus Mitchell 1997)

CANDIDATE-ELIMINATION konvergiert gegen die Hypothese, die die Zielfunktion korrekt beschreibt, wenn

- die Trainingsbeispiele keine Fehler enthalten,
- die Zielfunktion durch ein Element von *H* beschrieben werden kann (also als Konjunktion von Attributbelegungen).

Online-Anwendung des Verfahrens:
Neue Trainingsbeispiele können angefordert werden, solange die einelementige Zielhypothese noch nicht erreicht ist, d.h. solange noch Mehrdeutigkeit besteht.

Nachteil des Verfahrens:

- bei fehlerhaften Trainingsdaten falsches Ergebnis!

Vorteile des Verfahrens:

- Bei Inkonsistenzen in den Trainingsdaten oder bei Nicht-Darstellbarkeit der Zielfunktion in *H* terminiert der Algorithmus mit dem *leeren* Versionenraum. (Indikatorwirkung!)
- Auch bei unvollständiger Konvergenz (nicht-einelementiges Intervall [G, S] als Ergebnis) können einige neue Instanzen schon sicher klassifiziert werden ("partiell gelernter Begriff").

Beispiele für neue Instanzen zum "Sport"-Beispiel:

Instance	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
A	Sunny	Warm	Normal	Strong	Cool	Change	?
B	Rainy	Cold	Normal	Light	Warm	Same	?
C	Sunny	Warm	Normal	Light	Warm	Same	?
D	Sunny	Cold	Normal	Strong	Warm	Same	?

- A wird von *jeder* Hypothese im erhaltenen Versionenraum als "positiv" klassiert – diese Klassifikation ist verlässlich
- hierfür genügt es, festzustellen, dass A mit jedem Element von S konsistent ist
- B wird von jeder Hypothese im erhaltenen Versionenraum als "negativ" klassifiziert
- hierfür genügt es, festzustellen, dass B mit keinem Element von G konsistent ist

- C kann nicht als positiv oder negativ klassifiziert werden (erfordert neue Trainingsdaten) – gleichviele Hypothesen aus dem erhaltenen Versionenraum *V* klassieren C als positiv wie als negativ
- D: 2 der Hypothesen aus *V* sagen "positiv", 4 "negativ" – mögliche Entscheidung könnte Mehrheitsvotum zugrundelegen

Die Fälle C und D eröffnen allgemein die Frage, wie mit unvollständigem Wissen (*V* nicht einelementig) umzugehen ist

grundsätzlicheres Problem: die Zielfunktion ist eventuell in *H* gar nicht beschreibbar

Beispiel: disjunktiv gebildete Konzepte
"Sky = Sunny **oder** Sky = Cloudy"

in unserem Hypothesenraum nicht repräsentierbar!
3 Trainingsbeispiele für dieses Konzept:

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Cool	Change	Yes
2	Cloudy	Warm	Normal	Strong	Cool	Change	Yes
3	Rainy	Warm	Normal	Strong	Cool	Change	No

CANDIDATE-ELIMINATION konvergiert hier gegen den leeren Versionsraum (denn schon S_2 ist <?, Warm, Normal, Strong, Cool, Change> und damit inkonsistent mit dem 3. Trainingsbeispiel)

$\Rightarrow$ mit dem Zulassen nur von Konjunktionen von Attributbelegungen haben wir eine Verzerrung (*Bias*) eingeführt

Beachte: hier "*induktiver Bias*", i. Ggs. zu Schätzungs-Bias in der Statistik

H hat ein "Ausdrucksdefizit" (nicht alle möglichen Zielfunktionen können als Hypothesen in H ausgedrückt werden)

es gibt verschiedene Möglichkeiten des Umgangs mit dem Bias

Ausdrucksdefizit von $\mathcal{H}$

'BIAS' (Verzerrung)

Fallbasierte Klassifikation
Speichere Lernbeispiele
Klassifiziere $\boldsymbol{x}$ falls $\boldsymbol{x} \in \omega^+$ oder $\boldsymbol{x} \in \omega^-$; sonst *Rückweisung*

Einstimmiges Votum
Berechne den Versionenraum $\mathfrak{V}$
Klassifiziere falls $\boldsymbol{x}$ *allen* oder *keinem* $h \in \mathfrak{V}$ genügt;
sonst *Rückweisung*

Generalkonsens
Berechne den Versionenraum $\mathfrak{V}$
Klasssifiziere $\boldsymbol{x}$ positiv falls $h \in \mathfrak{V}_G \Rightarrow h \models \boldsymbol{x}$
Klasssifiziere $\boldsymbol{x}$ negativ im anderen Fall

Mehrheitsvotum
Berechne den Versionenraum $\mathfrak{V}$
Klassifiziere $\boldsymbol{x}$ gemäß Stimmenverhältnis:

$$n_+(\boldsymbol{x}) = \text{card}\{h \in \mathfrak{V} \mid h \models \boldsymbol{x}\}$$
$$n_-(\boldsymbol{x}) = \text{card}\{h \in \mathfrak{V} \mid h \not\models \boldsymbol{x}\}$$

Occam's Razor
Berechne den Versionenraum $\mathfrak{V}$
Wähle ein $h \in \mathfrak{V}$ mit *minimaler Komplexität* (?!?)
Klassifiziere $\boldsymbol{x}$ gemäß $h \models \boldsymbol{x}$

(aus Schukat-Talamazzini 2002)

Vermeidung des Bias: Wähle $\wp(\Omega)$ als Hypothesenraum (Potenzmenge) – d.h. Menge *aller* Teilmengen

Operationalisierung z.B. durch Zulassung beliebiger Konjunktionen, Disjunktionen und Negationen von Attributbelegungen

CANDIDATE-ELIMINATION kann auch hierfür benutzt werden
(nur mit entspr. veränderter Interpretation von min, max, inf, sup)

$\Rightarrow$ dann keine Gefahr mehr, dass das Zielkonzept nicht in *H* ausdrückbar ist

aber:
jetzt ist keine Generalisierung über die Trainingsdaten hinaus mehr möglich!

S ist stets die Disjunktion der Positivbeispiele
G ist stets die negierte Disjunktion der Negativbeispiele
$\Rightarrow$ die einzigen Instanzen, die eindeutig klassiert werden, sind die Trainingsdaten selbst

damit der Algorithmus gegen einen einelementigen Versionenraum konvergiert, müssen *alle* Elemente von Ω als Trainingsdaten präsentiert werden!

auch das "Mehrheitsvotum"-Prinzip hilft hier nicht weiter:
alle Nicht-Trainingsdaten werden *präzise von der Hälfte* der Hypothesen im Versionenraum positiv klassiert, von der anderen Hälfte negativ.

Schlussfolgerung:
"Futility of bias-free learning"

Ein Lernender, der keine a-priori-Annahmen über die Art des zu lernenden Begriffs macht, hat keine rationale Grundlage, um neue Instanzen zu klassieren.

Charakterisierung induktiver Lernansätze durch die *Art des induktiven Bias*, den sie voraussetzen

präzise Definition des induktiven Bias:

L sei ein induktiver Lernalgorithmus
L(x, D) sei die Klassifikation, die L für x liefert, wenn der Trainingsdatensatz D zugrundegelegt wurde ($x \notin D$)

Der induktive Bias von L ist die Menge B aller Annahmen, die benötigt werden, damit für alle $x \notin D$ L(x, D) logisch beweisbar ist aus (B und D und x).

$$(B \wedge D \wedge x) \quad |— \quad L(x, D)$$

(Man fordert noch die Minimalität von B.)

Beispiele:

- Lernalgorithmus "Rote-Lerner" (fallbasierte Klassifikation): Lernen bedeutet, jedes Trainingsbeispiel zu speichern. Neue Instanzen werden klassiert durch "Nachsehen" im Speicher – wenn sie dort schon gespeichert sind, wird das dortige Klassifikationsergebnis zurückgegeben, sonst erfolgt Rückweisung.

kein induktiver Bias

- Lernalgorithmus CANDIDATE-ELIMINATION, wobei Klassierung nur erfolgt, wenn alle Elemente des erhaltenen Versionsraumes dasselbe Ergebnis liefern (sonst Rückweisung).

Induktiver Bias: das Zielkonzept kann im Hypothesenraum H repräsentiert werden.

- FIND-S

induktiver Bias: Zielkonzept kann in H repräsentiert werden und alle Instanzen, über die kein Wissen vorliegt, sind negativ zu klassieren.

Stärke des induktiven Bias:

Rote-Lerner < CANDIDATE-ELIMINATION < FIND-S

(z.T. nach Mitchell 1997, S. 20–51)