Concept Learning

Concept Learning

Concept Learning — Begreppsinlärning

Learning of a boolean function from examples

Categories

- "Nice weather"
- "Dog"
- "Motor vehicle"
- "Criminal offence"

Subsets of a superset X

Terminology

c The concept to learn

$$c(x) \rightarrow 0/1, \quad x \in X$$

h Hypothesis, Result of the learning ("guessed c")

$$h(x) \rightarrow 0/1, \quad x \in X$$

H Hypotheses space (Hypotesrum), All conceivable hypotheses (before data arrives)

$$h \in H$$

D Set of available training data

$$D \subseteq X$$

Terminology

Two kinds of training examples

Positive example:

$$x : c(x) = 1, \quad x \in D$$

Negative example:

$$x : c(x) = 0, \quad x \in D$$

Example of a *concept*

"Nice Weather"

Let each "weather instance" x_i be composed of four attributes:

$x_1 =$ <Sunny, Warm, Windy, Dry>
$x_2 =$ <Cloudy, Warm, Calm, Dry>
$x_3 = \cdots$

Generally: $Sky \times Temperature \times Wind \times Humidity$

Assume that the attributes can only take on certain discrete values:

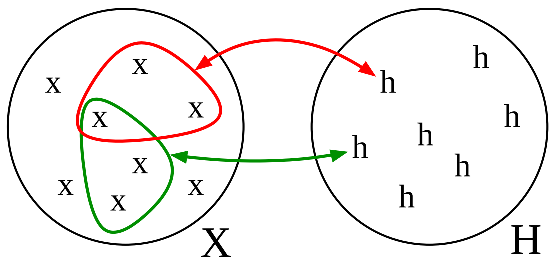
Sky	$\in$ { Sunny, Cloudy, Rainy }
Temp	$\in$ { Warm, Cold }
Wind	$\in$ { Windy, Calm }
Humid	$\in$ { Humid, Dry }

Number of possible weathers: $|X| = 3 \cdot 2 \cdot 2 \cdot 2 = 24$

Typical training samples

$x_1 =$ <Sunny, Warm, Windy, Dry> $\rightarrow$ Nice
$x_2 =$ <Sunny, Warm, Windy, Humid> $\rightarrow$ Nice
$x_3 =$ <Rainy, Cold, Windy, Humid> $\rightarrow$ Bad
$x_4 =$ <Sunny, Warm, Calm, Humid> $\rightarrow$ Nice

What does the hypotheses space H look like?



Each hypothesis h corresponds to one subset of X

How many hypotheses can we choose from?
How many subsets does X have?

$$|H| = 2^{|X|}$$
$$|H| = 2^{24} = 16777216$$

It is necessary to make restrictions!

Example of a Restriction

Assume that the concept is always a conjunction of attribute values

Examples of concepts c of this kind

Sunny & Warm
Cold & Calm & Dry

How many hypotheses do we have now?

Sky	Temperature	Wind	Humidity
Sunny Cloudy Rainy $\star$	Warm Cold $\star$	Windy Calm $\star$	Dry Humid $\star$

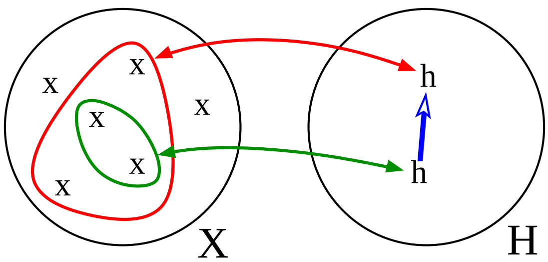
$4 \cdot 3 \cdot 3 \cdot 3 = 108$

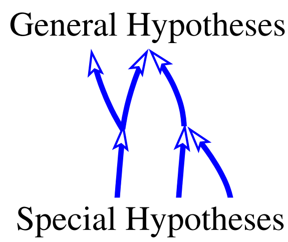
Learning $\equiv$ search for a hypothesis which matches all examples

Use the structure of H to search faster

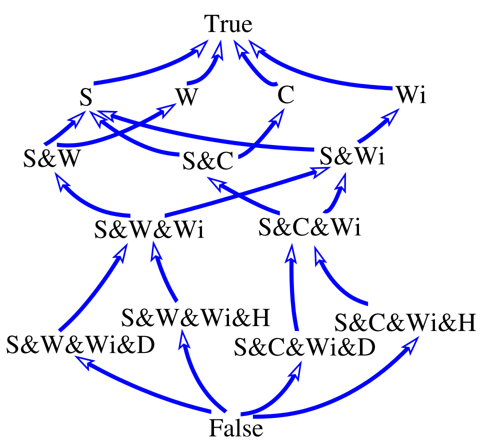
Some hypotheses are more general than others

Partial order between pairs of hypotheses (partiell ordning)





Most General in our example: "All weathers are nice"
Most Special in our example: "No weather is nice" (!)



Find-S algorithm

Start from the Most Special hypothesis and generalize when necessary.

$\hat{h} \leftarrow$ most special hypothesis in H
for $e \leftarrow$ next example:
  if positive example:
    generalize $\hat{h}$ to cover e too

Returns the most special hypothesis which is consistent (konsistent) with all examples.

Concrete example: "Nice Weather" assuming that this concept is a conjunction of attributes.

Initial Hypothesis: Current Hypothesis: $< \emptyset, \emptyset, \emptyset, \emptyset >$ (Maximally pessimistic) <Sunny, Warm, Windy, Dry> <Sunny, Warm, Windy, $\star$ > <Sunny, Warm, $\star, \star$ >

Training examples:

x_1 = <Sunny, Warm, Windy, Dry> $\rightarrow$ Nice
x_2 = <Sunny, Warm, Windy, Humid> $\rightarrow$ Nice
x_3 = <Rainy, Cold, Windy, Humid> $\rightarrow$ Bad
x_4 = <Sunny, Warm, Calm, Humid> $\rightarrow$ Nice

Final hypothesis: "Nice Weather" $\equiv$ Sunny $\wedge$ Warm

Problems with Find-S

- Impossible to know if only one unique hypothesis remains.
- Why should we prefer the most specific hypothesis?
- We will not detect inconsistent data since all negative examples are ignored.
- What happens if there are more equally specific hypotheses?

Version Space (VS)

The set of all hypotheses consistent with the examples seen so far.

- $VS \subseteq H$
- $|VS| = 1$ One unique solution
- $VS = \emptyset$ Inconsistent examples

The List-then-Eliminate algorithm

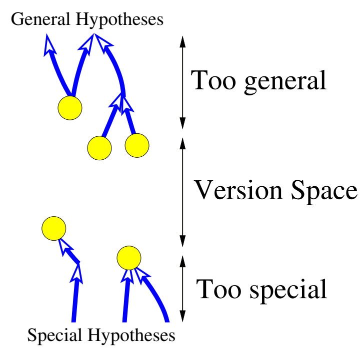
Direct representation of the Version Space (VS)

$VS \leftarrow H$
for $e \leftarrow$ next example:
remove all hypotheses from VS which are not consistent with e

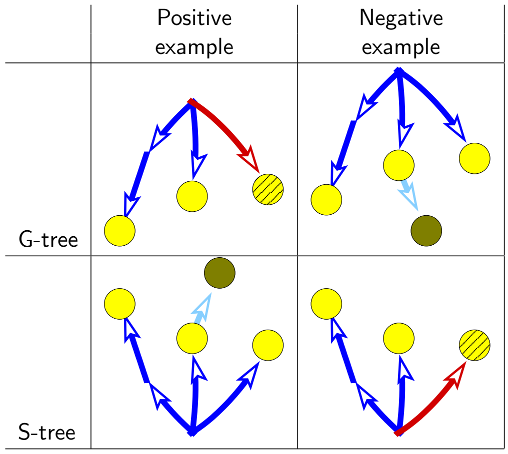
Problem: H is normally too large!

Candidate Elimination

- Efficient representation of the Version Space
- Utilizes the partial ordering between hypotheses.



$G \leftarrow$ most general hypotheses in H
$S \leftarrow$ most special hypotheses in H
for $e \leftarrow$ next example:
  if positive example:
    $G \leftarrow G - \{\text{hypotheses not including } e\}$
    $S \leftarrow \text{generalize } S \text{ to include } e$
    Remove "general duplicates" from S
  else:
    $S \leftarrow S - \{\text{hypotheses including } e\}$
    $G \leftarrow \text{specialize } G \text{ not to include } e$
    Remove "special duplicates" from G
  Clean G from hypotheses not more general than something in S
  Clean S from hypotheses not more special than something in G



Concrete example: "Nice Weather" assuming that this concept is a conjunction of attributes.

$$G = \{< \star, \star, \star, \star >\}$$
$$G = \{<\text{Sunny}, \star, \star, \star >, <\text{Cloudy}, \star, \star, \star >, < \star, \text{Warm}, \star, \star >, < \star, \star, \text{Calm}, \star >, < \star, \star, \star, \text{Dry}>\}$$
$$G = \{<\text{Sunny}, \star, \star, \star >, < \star, \text{Warm}, \star, \star >\}$$

$x_1 =$	<Sunny, Warm, Windy, Dry>	$\rightarrow$ Nice
$x_2 =$	<Sunny, Warm, Windy, Humid>	$\rightarrow$ Nice
$x_3 =$	<Rainy, Cold, Windy, Humid>	$\rightarrow$ Bad
$x_4 =$	<Sunny, Warm, Calm, Humid>	$\rightarrow$ Nice

$$S = \{<\text{Sunny, Warm}, \star, \star >\}$$
$$S = \{<\text{Sunny, Warm, Windy}, \star >\}$$
$$S = \{<\text{Sunny, Warm, Windy, Dry}>\}$$

Bias

Our learning algorithm is not *unbiased* (objektiv) since it can not choose among all possible hypotheses.

Induction Bias — The choice of learning algorithm influences the result

Unbiased Learner A learning algorithm where all hypotheses are treated equally

Restriction Bias Restriction of which hypotheses are allowed

Preference Bias Tendency to prefer certain hypotheses before others

Is an *Unbiased Learner* better?

All subsets of X are equally likely.

Knowledge about $x_1, x_2, \ldots, x_n$ will reveal nothing about x_{n+1}

Without bias it becomes impossible to generalize to unseen examples $x \notin D$.