




Concept Learning

SGA07_DATASCI

18th February 2020



Module Overview

- Rule-based classification: Concept Learning
- General-to-Specific Ordering
- Find-S Algorithm
- List-then-Eliminate Algorithm



Book Keeping

- Regression Task/Practice Lab due today
- Reminder : Group task submission due 28th February
- Final Project : You should have gotten data & start exploratory data mining
- Collection of module slides (with some animation) will be shared



Outcome

After this Module, you will;

- Get an overview of classification data mining process
- Understand the basis of concept learning as a data mining technique
- Be introduced to 2 search-based algorithms (Find-S & List-then-Eliminate)



Classification Data Mining

- Supervised Learning
 - Rule-based (DT)
 - Probabilistic (NB)
 - Linear Classification (SVM)
 - Distance based learning (K-means)
- Unsupervised Learning
 - Clustering (K-cluster)
 - Association rule mining

“

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

”

Classification Process

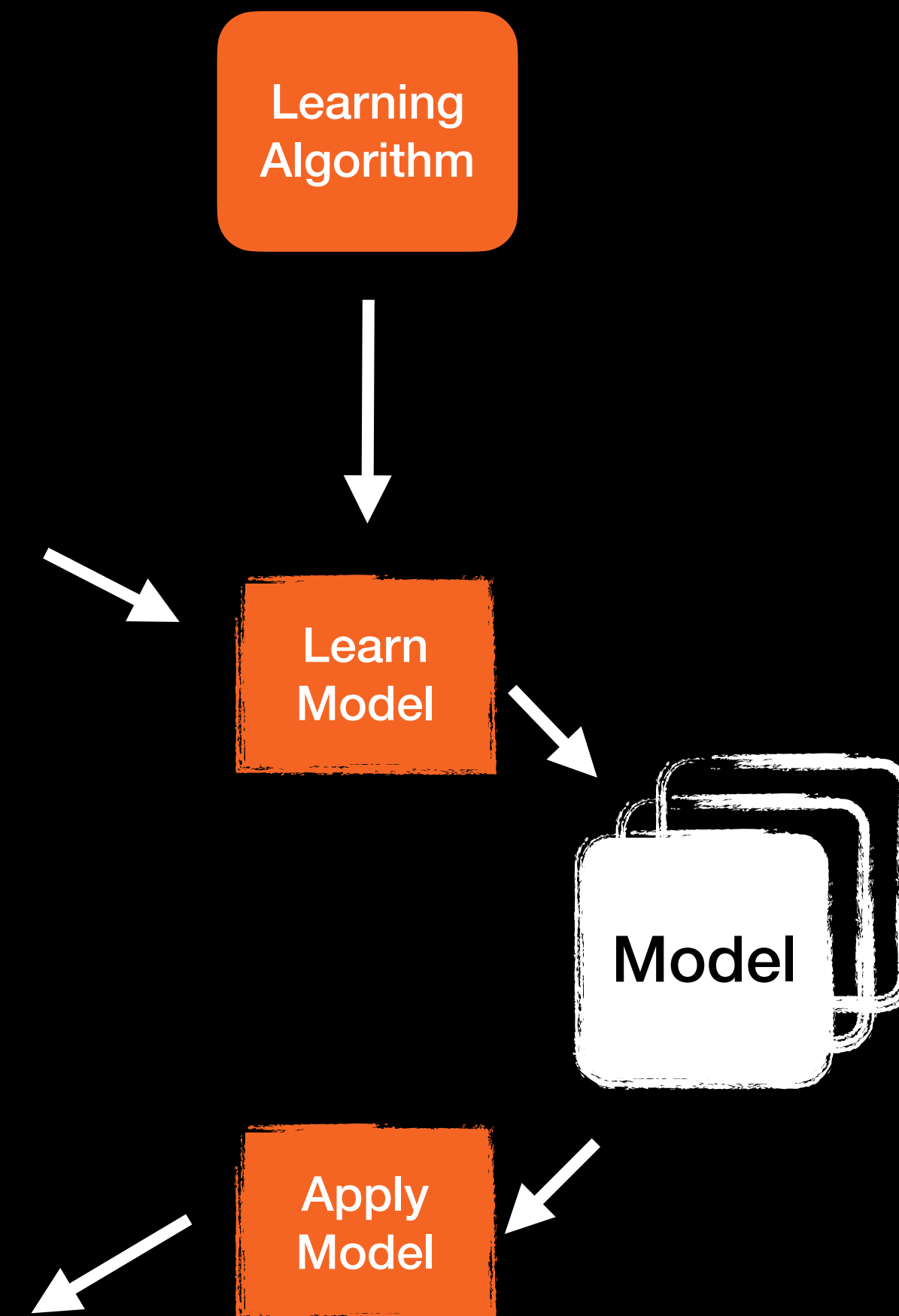
- Given a set of instances – training set:
 - Each instance contains a set of attributes, one of the attributes is the class.
- Find model for class attribute as function of values of other attributes.
- Aim to assign unseen instances to a class as accurately as possible:
 - Dataset divided into training and test sets, former used to build model and latter to validate it

Training Data

Tid	Att_1	Att_2	Att_3	Class
1	Yes	Large	125k	No
2	No	Medium	100k	No
3	No	Small	70k	No
4	Yes	Medium	120k	No
5	No	Large	95k	Yes
6	Yes	Medium	60k	No
7	No	Large	220k	No

Test Data

Tid	Att_1	Att_2	Att_3	Class
8	No	Small	55k	?
9	Yes	Medium	80k	?
10	Yes	Large	110k	?





Concept Learning

- Rule-based classification technique
- Boolean-valued function
- Learning based on given structure
- General-to-specific search
- Best fits the training data

“

A phenomenon of searching through predefined space of potential hypothesis that best fits the training example.

”

Simona Halep (4)

Let's assume you work as part of Simona Halep's coaching team. Simona is the fourth highest ranked female tennis player. You are to predict if Simona would enjoy her match based on given attributes for a particular day.

Simona Halep Data

Day_ID	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoyMatch
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
5	Rainy	Warm	Normal	Weak	Warm	Same	?

Concept Learning Notation

- Given:
 - Instances X : Possible days, each described by the attributes
 - Sky (with possible values Sunny, Cloudy, and Rainy),
 - AirTemp (with values Warm and Cold),
 - Humidity (with values Normal and High),
 - Wind (with values Strong and Weak),
 - Water (with values Warm and Cool), and
 - Forecast (with values Same and Change).
 - Hypotheses H : Each hypothesis is described by a conjunction of constraints on the attributes Sky, AirTemp, Humidity, Wind, Water, and Forecast. The constraints may be "?" (any value is acceptable), "0" (no value is acceptable), or a specific value.
 - Target concept c : EnjoySport : $X \rightarrow \{0,1\}$
 - Training examples D : Positive and negative examples of the target function.
- Determine:
 - A hypothesis h in H such that $h(x) = c(x)$ for all x in X .

General-to-Specific Hypothesis

$h(x) = 1$ If instance x satisfies all constraints

$h(x) = 0$ Otherwise

Most general hypothesis $\langle ?, ?, ?, ?, \dots, ? \rangle$

Most specific hypothesis $\langle \emptyset, \emptyset, \emptyset, \emptyset, \dots, \emptyset \rangle$

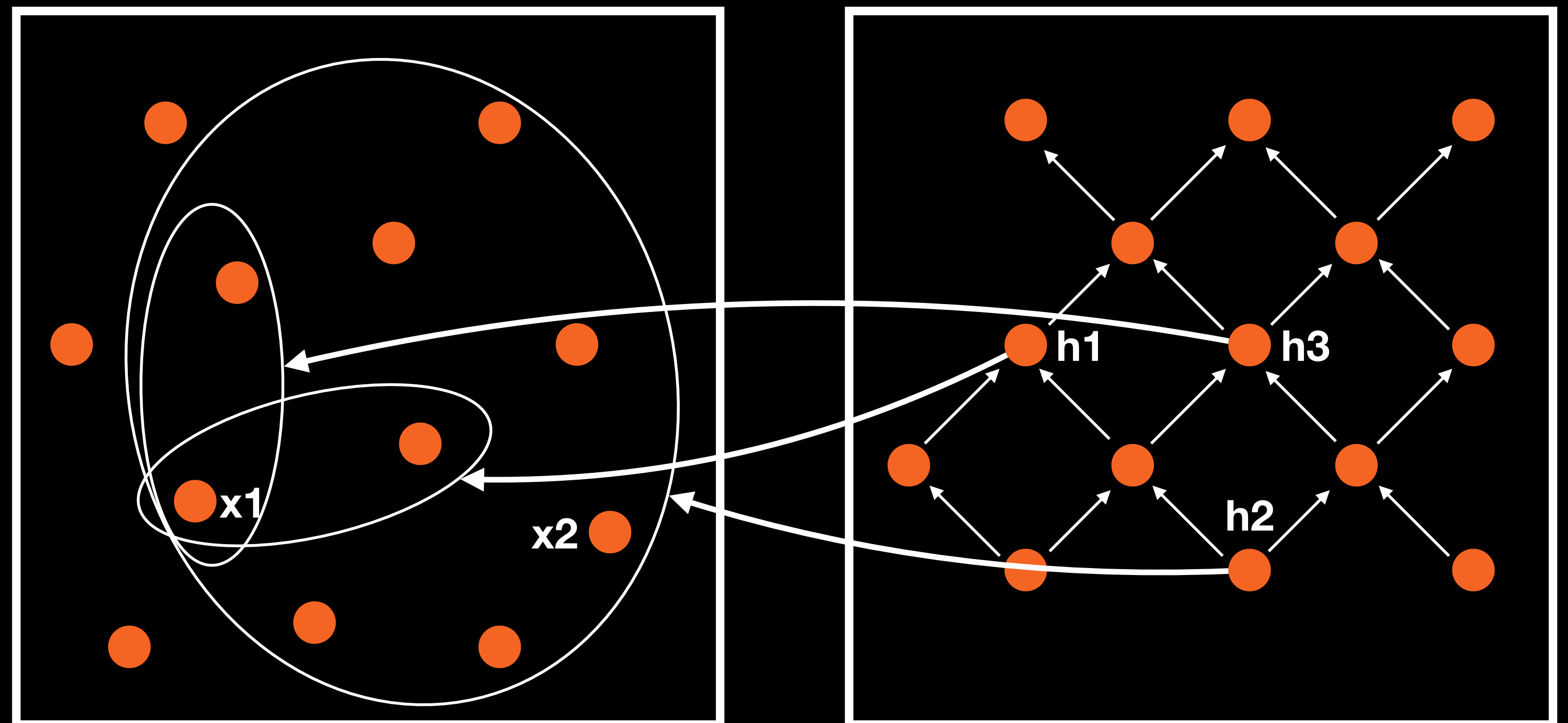
$x_1 = \langle \text{Sunny}, \text{Warm}, \text{High}, \text{Strong}, \text{Cool}, \text{Same} \rangle$

$x_2 = \langle \text{Sunny}, \text{Warm}, \text{High}, \text{Light}, \text{Warm}, \text{Same} \rangle$

$h_1 = \langle \text{Sunny}, ?, ?, \text{Strong}, ?, ? \rangle$

$h_2 = \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle$

$h_3 = \langle \text{Sunny}, ?, ?, ?, \text{Cool}, ? \rangle$





Find-S Algorithm

- Initialise h to most specific hypothesis in H
- For each positive training instance x
 - For each attribute constraint a_j in h
 - if constraint a_j satisfied by x
 - then
 - do nothing
 - else
 - replace a_j in h by next more general constraint satisfied by x
- Output hypothesis h

Find-S: Step 0

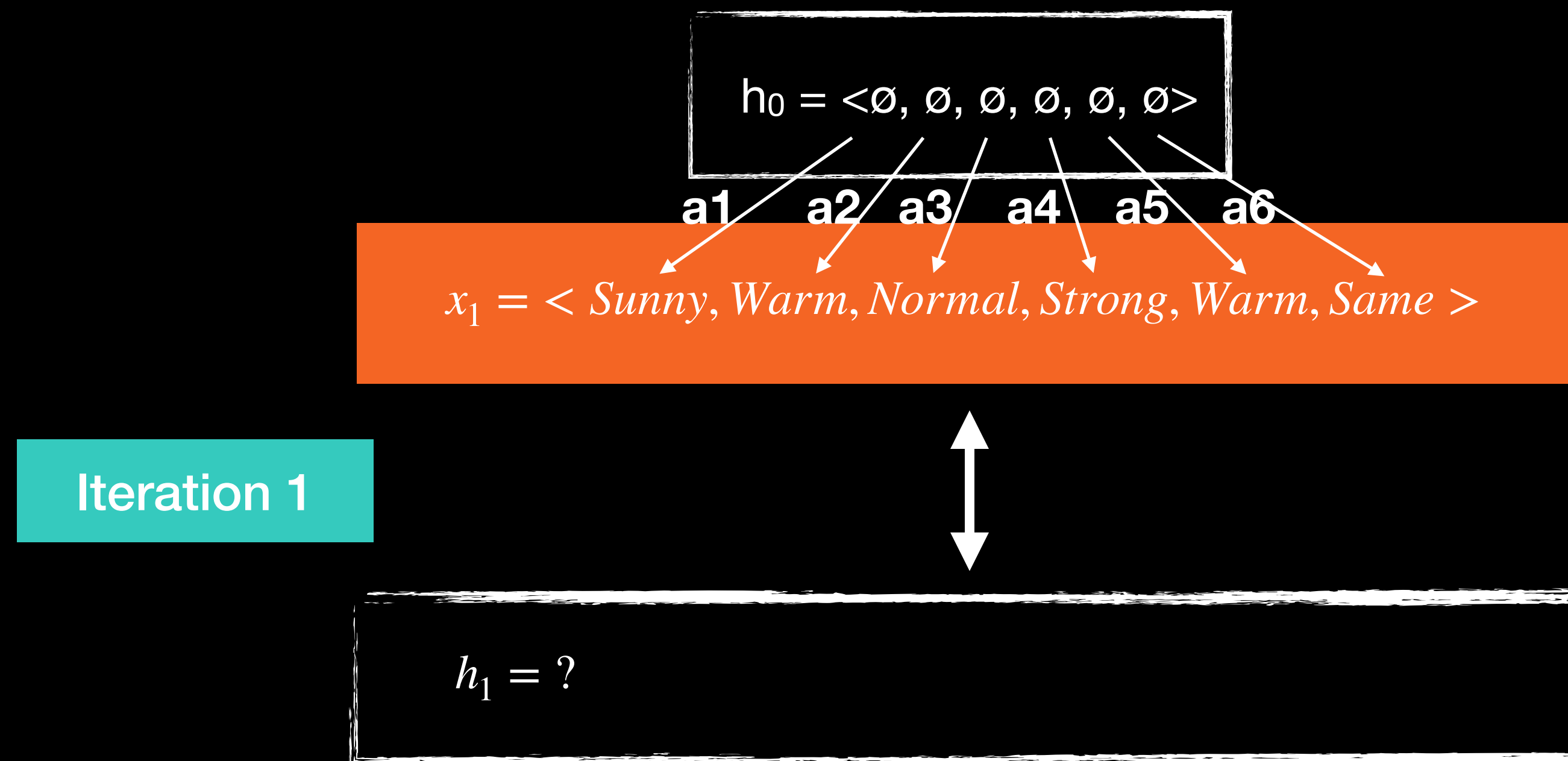
Initialise h to most specific hypothesis in H

$h_0 = \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$

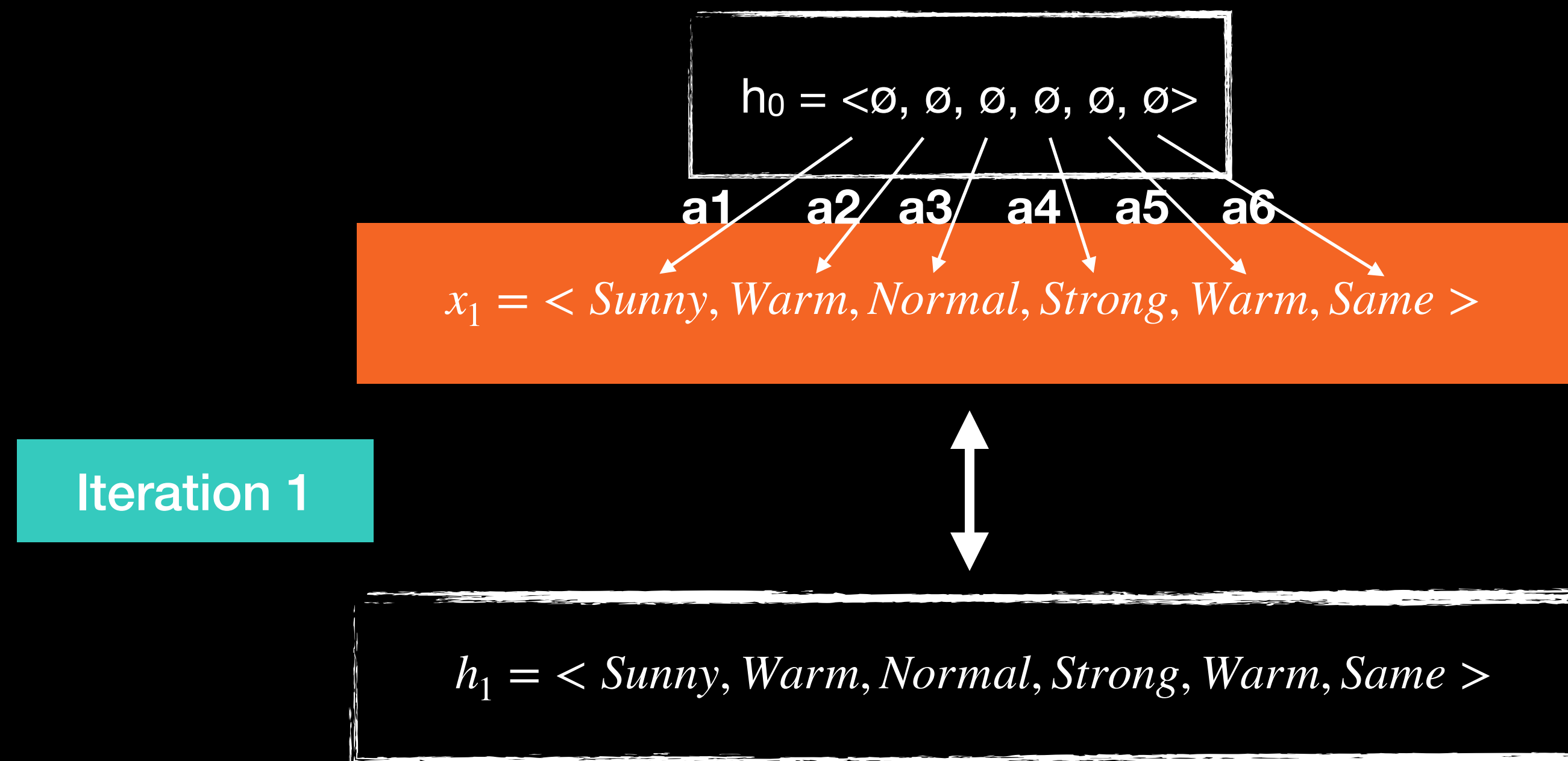
Simona Halep Data

Day_ID	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoyMatch
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

Find-S: Step 1



Find-S: Step 1



Find-S: Step 2

$h_1 = \langle \text{Sunny}, \text{Warm}, \text{Normal}, \text{Strong}, \text{Warm}, \text{Same} \rangle$

a1

a2

a3

a4

a5

a6

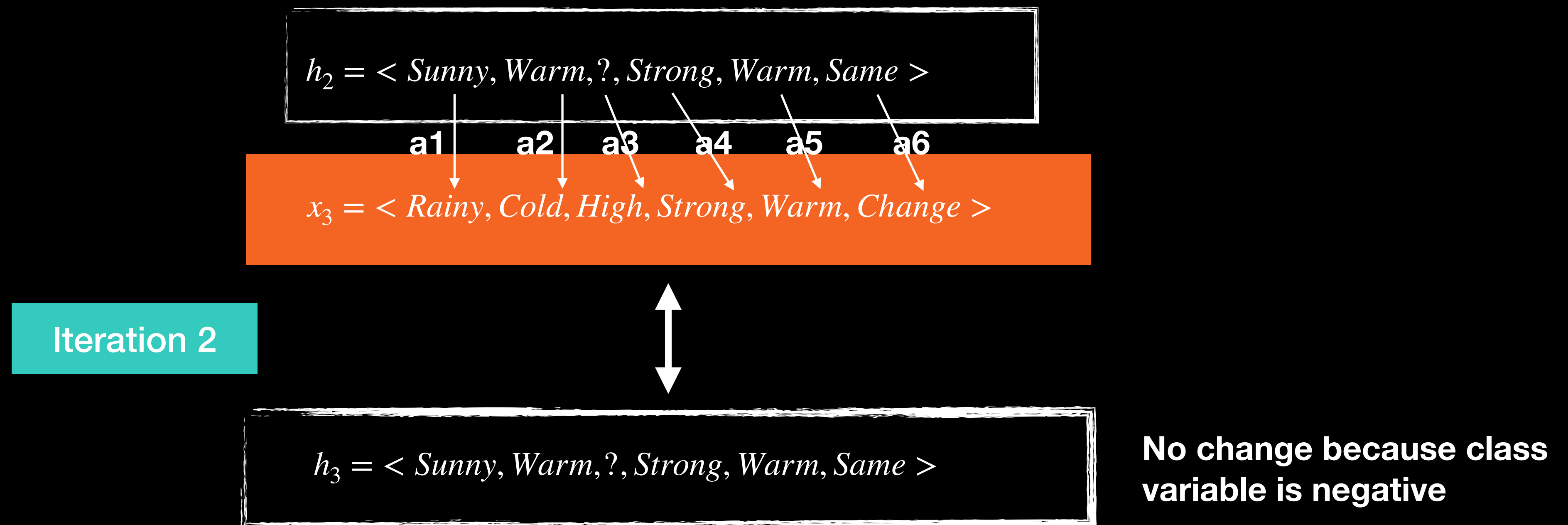
$x_2 = \langle \text{Sunny}, \text{Warm}, \text{High}, \text{Strong}, \text{Warm}, \text{Same} \rangle$

Iteration 2

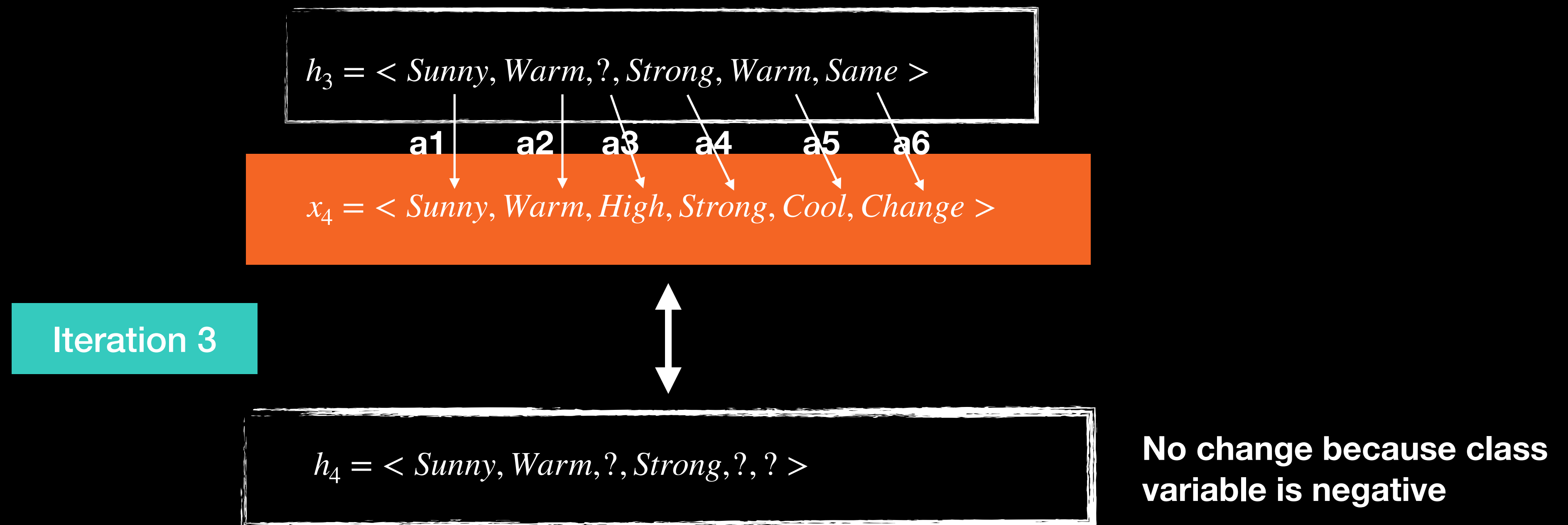


$h_2 = \langle \text{Sunny}, \text{Warm}, ?, \text{Strong}, \text{Warm}, \text{Same} \rangle$

Find-S: Step 3



Find-S: Step 4





Find-S Shortfalls

- Has the learner converge to the correct target concept
- Why prefer the most specific hypothesis
- Are the training example consistent and void of noise
- What if there are several maximally specific consistent hypothesis



List-then-Eliminate Algorithm

- Version Space \leftarrow a list containing every hypothesis in H
- For each training example, $(x, c(x))$
 - Remove from Version Space any hypothesis h for which
$$h(x) \neq c(x)$$
- Output the list of hypotheses in VersionSpace

L-t-E: Step 0

$$G0 \Leftarrow \langle ?, ?, ?, ?, ?, ? \rangle$$

Initialisation

$$S0 \Leftarrow \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$$

Simona Halep Data							
Day_ID	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoyMatch
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

L-t-E: Step I

$x_1 = \langle \text{Sunny, Warm, Normal, Strong, Warm, Same} \rangle$

Iteration 1

$G1 \Leftarrow \{ \langle ?, ?, ?, ?, ?, ? \rangle \}$

$S1 \Leftarrow \{ \langle \text{Sunny, Warm, Normal, Strong, Warm, Same} \rangle \}$

Remove from G any hypothesis inconsistent with x

For each hypothesis s in S not consistent with x

Remove s from S

Add to S all minimal generalisations h of s such that

h is consistent with x and some member of G is more general than h

Remove from S any hypothesis more general than another hypothesis in S

L-t-E: Step 2

$x_2 = \langle \text{Sunny, Warm, High, Strong, Warm, Same} \rangle$

Iteration 2

$G2 \Leftarrow \{ \langle ?, ?, ?, ?, ?, ? \rangle \}$

$S2 \Leftarrow \{ \langle \text{Sunny, Warm, ?, Strong, Warm, Same} \rangle \}$

Remove from G any hypothesis inconsistent with x

For each hypothesis s in S not consistent with x

Remove s from S

Add to S all minimal generalisations h of s such that

h is consistent with x and some member of G is more general than h

Remove from S any hypothesis more general than another hypothesis in S

L-t-E: Step 3

$x_3 = \langle \text{Rainy}, \text{Cold}, \text{High}, \text{Strong}, \text{Warm}, \text{Change} \rangle$

Iteration 3

$G3 \Leftarrow \{ \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle, \langle ?, \text{Warm}, ?, ?, ?, ? \rangle, \langle ?, ?, ?, ?, ?, \text{Same} \rangle \}$

$S3 \Leftarrow \{ \langle \text{Sunny}, \text{Warm}, ?, \text{Strong}, \text{Warm}, \text{Same} \rangle \}$

Remove from G any hypothesis inconsistent with x

For each hypothesis s in S not consistent with x

Remove s from S

Add to S all minimal generalisations h of s such that

h is consistent with x and some member of G is more general than h

Remove from S any hypothesis more general than another hypothesis in S

L-t-E: Step 4

$x_4 = \langle \text{Sunny, Warm, High, Strong, Cool, Change} \rangle$

Iteration 3

$G_4 \Leftarrow \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle, \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

$S_4 \Leftarrow \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$

Remove from G any hypothesis inconsistent with x

For each hypothesis s in S not consistent with x

Remove s from S

Add to S all minimal generalisations h of s such that

h is consistent with x and some member of G is more general than h

Remove from S any hypothesis more general than another hypothesis in S



Convergence Condition

- No errors in training examples D .
- There is some hypothesis in H that correctly describes target concept.
- Target concept exactly learned when S and G boundary sets converge to single, identical hypothesis.



Practice Lab

Implement a find-S algorithm in R

Use the following Instructions:

- Get Simona Halep Data into R
- Write a user defined function that implements the Find-S algorithm
- Test on sample data from the class exercise



Recap/Summary

At the end of this Module, you should understand;

- Get an overview of classification data mining process
- Understand the basis of concept learning as a data mining technique
- Be introduced to 2 search-based algorithms (Find-S & List-then-Eliminate)



Suggested Material

- Machine Learning by Tom Mitchell Pages 20 - 28
- Machine Learning by Tom Mitchell Pages 29 - 47
- <https://rpubs.com/anablake/concept-learning>