IF4071 Pembelajaran Mesin: Pembelajaran Konsep

Sumber utama: Bab 2 Machine Learning (Tom M. Mitchell, 1997)

Masayu Leylia Khodra
S1-IF ITB

Pembelajaran Konsep

- Pembelajaran: specific training data → general konsep
- Konsep:
  - Deskripsi subset, cth: bil.positif $\subset$ bil. bulat
  - **Fungsi boolean yang didefinisikan untuk suatu himpunan, contoh isA(x), x$\in$Himpunan**
  - an abstract or general idea inferred or derived from specific instances (WordNet)
- Pembelajaran konsep:
  - Inferring/approximating suatu fungsi boolean berdasarkan kumpulan anggota dan bukan-anggota dari konsep tersebut.
  - Persoalan pencarian hipotesis yang best fits terhadap training data pada ruang hipotesis.

Training Examples for EnjoySport

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

T*raining experiences:* given instances X (possible days):

<<sky, airtemp, humidity, wind, water, forecast>, enjoysport>

What is the general concept?

Desain Learning System

- *Training experiences:* given instances X (possible days): <<sky, airtemp, humidity, wind, water, forecast>, enjoysport>
- Fungsi target: EnjoySport: X→{0,1}
- Representasi fungsi target: *conjunction of constraints on instance attributes*
- Pemilihan algoritma pembelajaran:
  - Find-S
  - List then eliminate
  - Candidate Elimination Algorithm

Representasi Hipotesis

- H: konjungsi constraints on attributes
- Constraint dapat berupa:
  - a specific value (e.g., "Water= Warm")
  - don't care / any value is acceptable (e.g., "Water = ?")
  - no value allowed (e.g.,"Water = ø")
- Jika instance x satisfy h, h(x)=1

Contoh,

Sky	AirTemp	Humid	Wind	Water	Forecast
<?	Cold	High	?	?	?>

Concept Learning Task

Given:

- Instances X: Possible days, each described by the attributes *Sky, AirTemp, Humidity, Wind, Water, Forecast*
- Target function c: EnjoySport : X → {0,1}
- Hypotheses H: Conjunctions of constraints
- Training examples D: Positive and negative examples of the target function: $< x_1, c(x_1) >, \ldots, < x_n, c(x_n) >$

Determine:

A hypothesis h in H such that h(x) = c(x) for all x in D.

Asumsi Fundamental dalam Pembelajaran Induktif

The inductive learning hypothesis

Any hypothesis found to approximate the target function **well over a sufficiently large set of training examples** will also approximate the target function **well over other unobserved examples.**

Jumlah Instans, Konsep, dan Hipotesis

- Attribute instance:
  - Sky: Sunny, Cloudy, Rainy
  - AirTemp: Warm, Cold
  - Humidity: Normal, High
  - Wind: Strong, Weak
  - Water: Warm, Cold
  - Forecast: Same, Change
- #distinct instances : 3*2*2*2*2*2 = 96
- #distinct concepts : 2^{96}
- #syntactically distinct hypotheses : 5*4*4*4*4*4=5120
- #semantically distinct hypotheses : 1+4*3*3*3*3*3=973

General-to-Specific Ordering

- Most general hypothesis <?,?,?,?,?,?>
- Most specific hypothesis < ø, ø, ø, ø, ø, ø >
- Definition: Let h_j and h_k be boolean-valued functions defined over X then h_j is **more general than or equal to h_k (written $h_j \geq h_k$) if and only if**

$$\forall x \in X : [\,(h_k(x) = 1) \rightarrow (h_j(x) = 1)]$$

- The relation ≥ imposes a partial order over the hypothesis space H that is utilized in many concept learning methods.

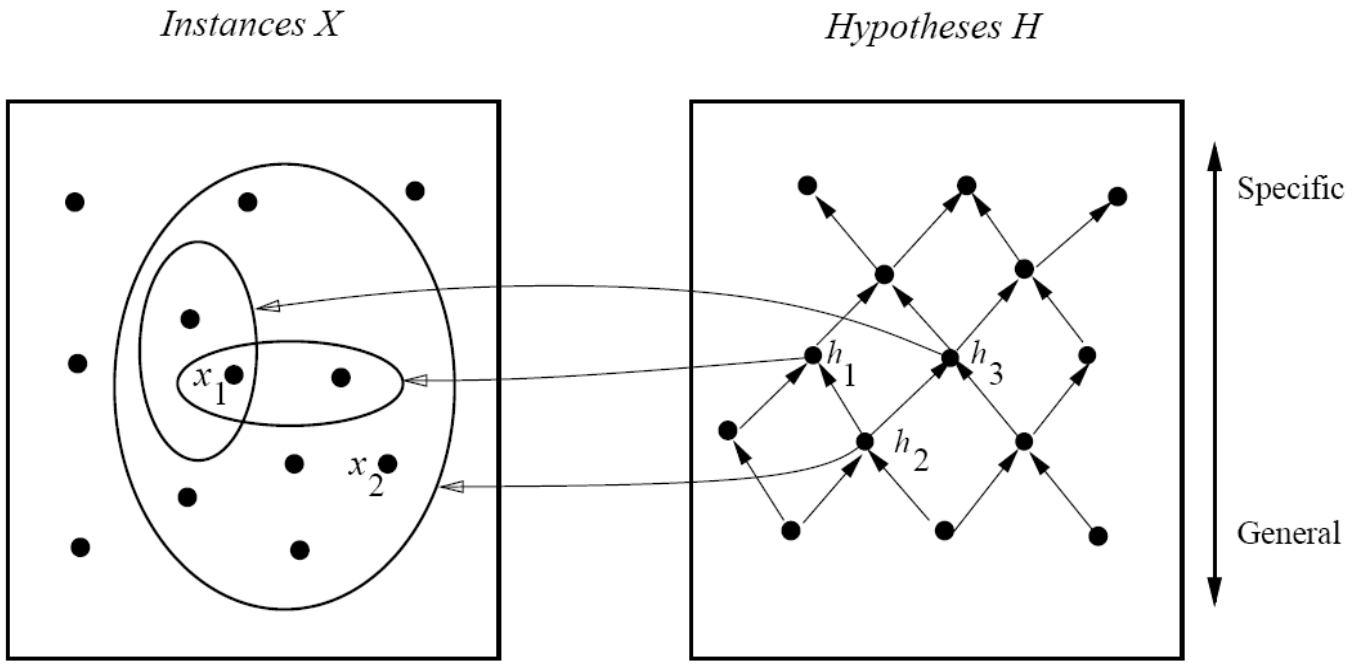
Hint: Organize the search to take advantage of the structure of the hypothesis space to improve running time

General-to-Specific Ordering

- Consider two hypotheses:
  - h_1=< Sunny,?,?,Strong,?,?>
  - h_2=< Sunny,?,?,?,?,?>
- Set of instances covered by h1 and h2:
  - h_2 memiliki batasan yang lebih sedikit dibanding h_1, sehingga akan mengklasifikasikan lebih banyak instans x bernilai postif atau h(x)=1.
    - h_2 merupakan konsep yang lebih general.

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Instance, Hypotheses, and More-General-Than



x_1= <Sunny, Warm, High, Strong, Cool, Same>

x_2= <Sunny, Warm, High, Light, Warm, Same>

h_1= <Sunny, ?, ?, Strong, ?, ?>

h_2= <Sunny, ?, ?, ?, ?, ?>

h_3= <Sunny, ?, ?, ?, Cool, ?>

- h2 is more general than h1
- h2 is more general than h3
- Neither h1 nor h3 is more general than the other
- More-general-than-or-equal-to defines a partial order over the hypothesis space H

Find-S: Find Maximally Specific Hypothesis

- Begin with the most specific possible hypothesis in H, generalize this hypothesis each time it fails to cover an observed positive training example.
  - most specific possible hypothesis: conjunction of ø
  - Generalize: ø → attribute value → ?
- Find-S algorithm ignores negative examples

Find-S Algorithm

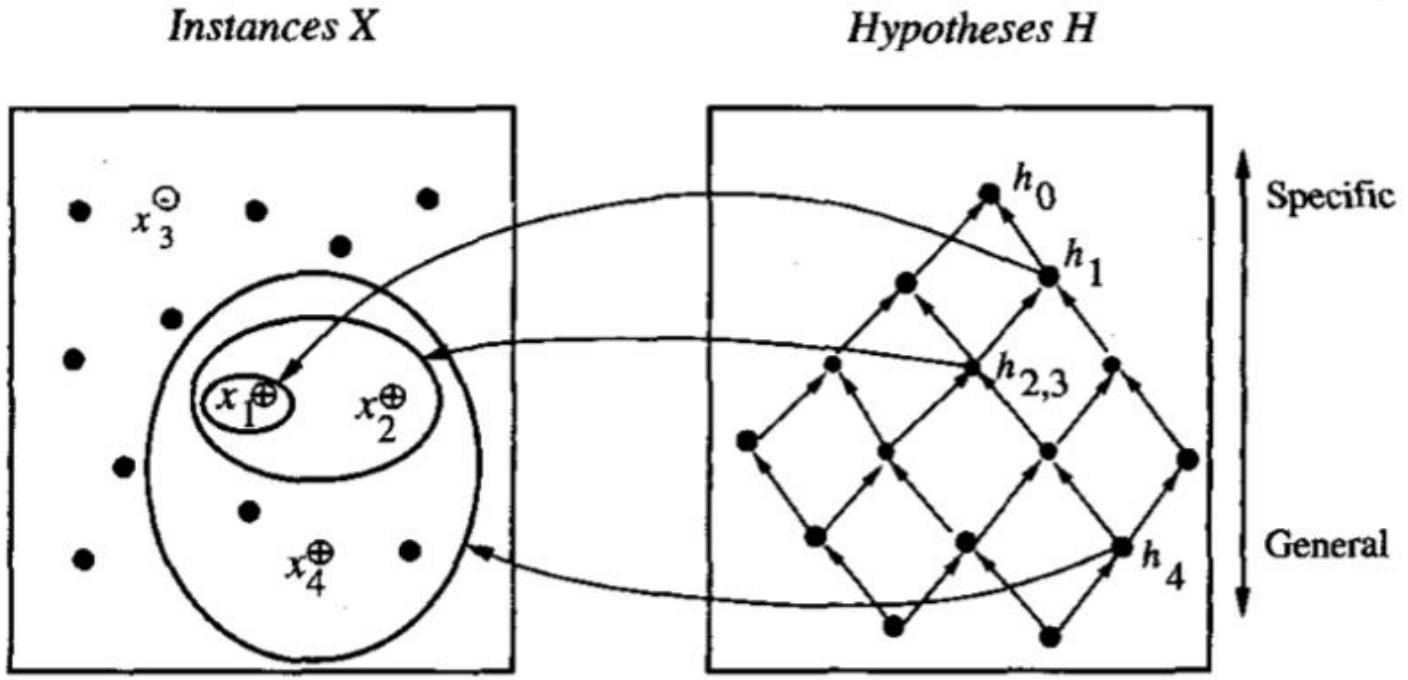
1. Initialize h to the most specific hypothesis in H
2. For each positive training instance x
   - For each attribute constraint a_i in h

   If the constraint a_i in h is satisfied by x

   Then do nothing

   Else replace a_i in h by the next more general constraint that is satisfied by x
3. Output hypothesis h

Hypothesis Space Search by Find-S



x_1 = <*Sunny Warm Normal Strong Warm Same*>, +

x_2 = <*Sunny Warm High Strong Warm Same*>, +

x_3 = <*Rainy Cold High Strong Warm Change*>, -

x_4 = <*Sunny Warm High Strong Cool Change*>, +

h_0 = <∅, ∅, ∅, ∅, ∅, ∅>

h_1 = <*Sunny Warm Normal Strong Warm Same*>

h_2 = <*Sunny Warm ? Strong Warm Same*>

h_3 = <*Sunny Warm ? Strong Warm Same*>

h_4 = <*Sunny Warm ? Strong ? ?*>

Find-S

- Pencarian hipotesis yang lebih general secara progresif dari hipotesis yang paling spesifik
- Find-S menjamin menghasilkan:
  - Hipotesis yang konsisten dengan contoh-contoh positif, dan tetap konsisten dengan contoh-contoh negatif
- Persoalan yang belum terjawab:
  1. Hipotesis yang paling spesifik yang konsisten dengan contoh-contoh pelatihan → cukup general ?
  2. Telah menghasilkan konsep target yang tepat ?
  3. Inconsisten training data → mislead Find-S
  4. Satu hipotesis unik yang paling spesifik → beberapa ?

Candidate-Elimination Algorithm

- Find-S: satu hipotesis konsisten dgn training data
  - Satu dari beberapa hipotesis yang konsisten
- List-then-eliminate: set semua hipotesis konsisten (version space)
  - Dengan enumerasi secara eksplisit semua hipotesis
- Candidate-Elimination: set semua hipotesis konsisten (version space)
  - Tanpa enumerasi secara eksplisit semua hipotesis
- Noisy training data → belum dapat ditangani

Version Spaces

The **version space**, $VS_{H,D}$, with respect to hypothesis space H and training examples D, is the subset of hypotheses from H consistent with all training examples in D.

$$VS_{H,D} \equiv \{h \in H | Consistent(h, D)\}$$

A hypothesis h is **consistent** with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\langle x, c(x)\rangle$ in D.

$$Consistent(h, D) \equiv (\forall \langle x, c(x)\rangle \in D)\ h(x) = c(x)$$

Version Space by List-Then-Eliminate

Algoritma brute-force:

1. VersionSpace ← a list containing every hypothesis in H
2. For each training example, <x, c(x)>

   remove from VersionSpace any hypothesis h for which h(x) ≠ c(x)
3. Output the list of hypotheses in VersionSpace

Batasan:

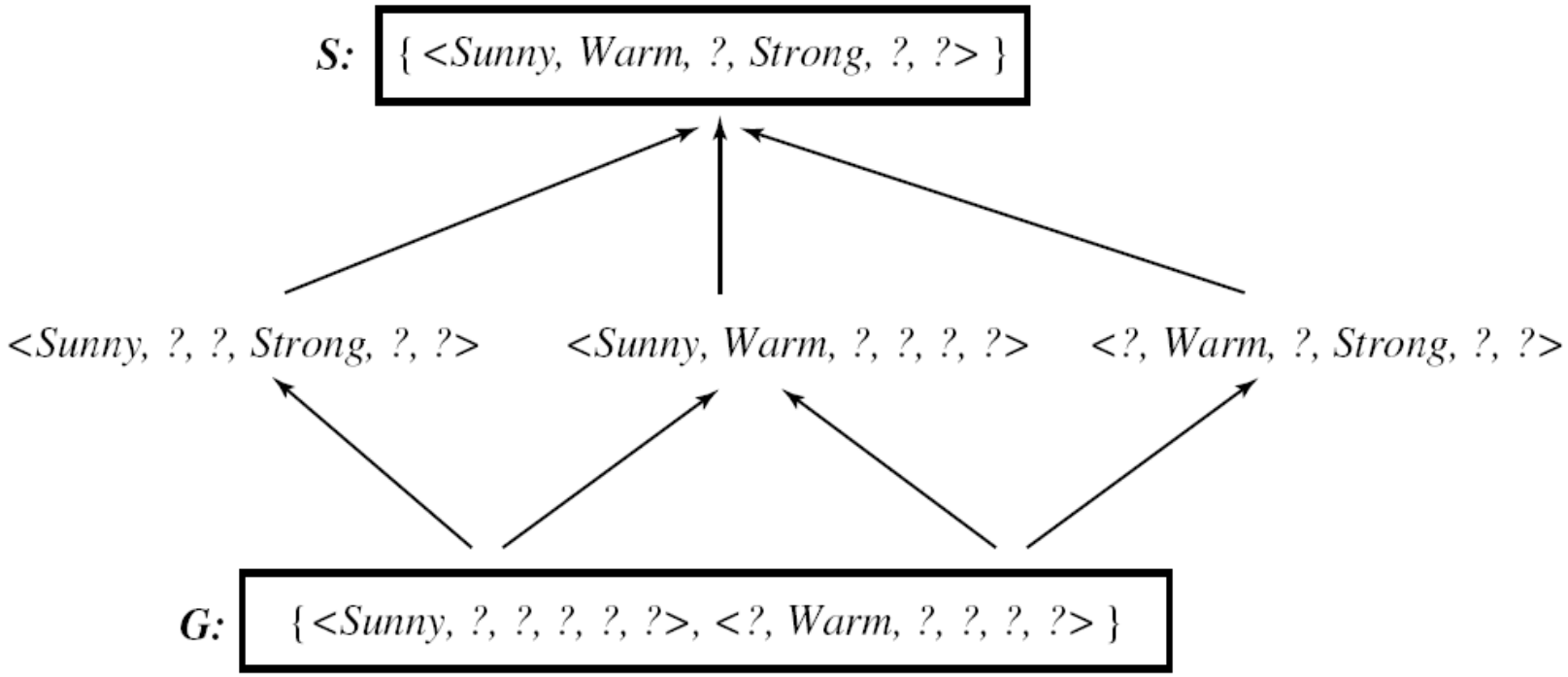
- Jumlah hipotesis terbatas
- Inisialisasi dengan exhaustive search → tidak realistik dgn fitur yang banyak

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Sky: Sunny, Cloudy, Rainy
AirTemp: Warm, Cold
Humidity: Normal, High
Wind: Strong, Weak
Water: Warm, Cold
Forecast: Same, Change

- ~~<ø, ø, ø, ø, ø, ø>~~
- ~~<S, ø, ø, ø, ø, ø>~~
- ~~..~~
- ~~< ø, W, ø, ø, ø, ø, >~~
- ...
- ~~<S,W,N,S,W,S>~~ data 2
- ~~<S,W,H,S,W,S>~~ data 1
- ...
- ~~<S,W,?,S,W,S>~~ data 4
- ...
- <S,W,?,S,?,?>
- ~~<?, ?, ?, ?, ?, ?>~~ data 3

Example Version Space



Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

S: most specific hypotheses
G: most general hypotheses

Representing Version Spaces

The General boundary, G, of version space $VS_{H,D}$ is the set of its maximally general members

The Specific boundary, S, of version space $VS_{H,D}$ is the set of its maximally specific members

Every member of the version space lies between these boundaries

$$VS_{H,D} = \{h \in H | (\exists s \in S)(\exists g \in G)(g \geq h \geq s)\}$$

where x ≥ y means x is more general or equal to y

Candidate Elimination Algorithm

G ← maximally general hypotheses in H

S ← maximally specific hypotheses in H

For each training example d, do

- If d is a positive example
  - Remove from G any hypothesis inconsistent with d
  - For each hypothesis s in S that is not consistent with d
    - Remove s from S
    - Add to S all minimal generalizations h of s such that
      1. h is consistent with d, and
      2. some member of G is more general than h
    - Remove from S any hypothesis that is more general than another hypothesis in S

Candidate Elimination Algorithm (2)

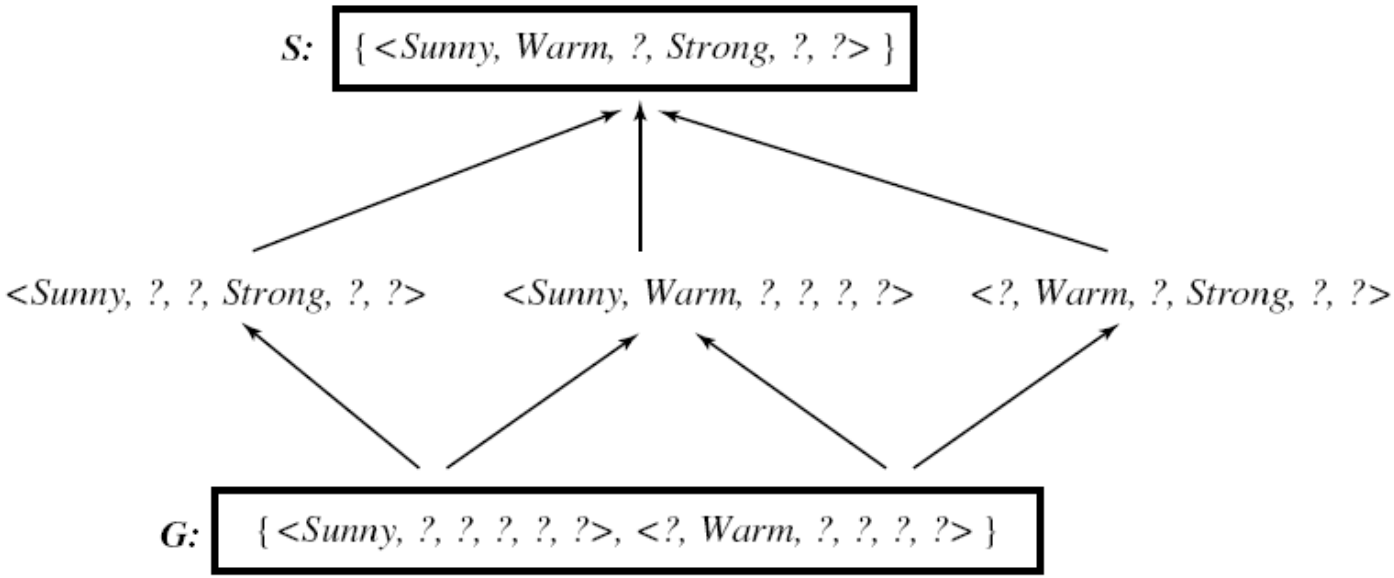
- If d is a negative example
  - Remove from S any hypothesis inconsistent with d
  - For each hypothesis g in G that is not consistent with d
    - Remove g from G
    - Add to G all minimal specializations h of g such that
    1. h is consistent with d, and
    2. some member of S is more specific than h
    - Remove from G any hypothesis that is less general than another hypothesis in G

Version Spaces for EnjoySport

	G	S
Inisialisasi	{<?,?,?,?,?,?>}	{<∅, ∅, ∅, ∅, ∅, ∅>}
<sunny, warm, normal, strong, warm, same, yes>	{<?,?,?,?,?,?>}	{~~<∅, ∅, ∅, ∅, ∅, ∅>,~~<sunny, warm, normal, strong, warm, same>}
<sunny, warm, high, strong, warm, same, yes>	{<?,?,?,?,?,?>}	{<sunny, warm, ?, strong, warm, same>}
<rainy, cold, high, strong, warm, change, no>	{<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, <?,?,?,?,?,same>}	{<sunny, warm, ?, strong, warm, same>}
<sunny, warm, high, strong, cool, change, yes>	{<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, ~~<?,?,?,?,?,same>~~}	{<sunny, warm, ?, strong, ?, ? >}

$VS_{H,D}$= {<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, <sunny, warm, ?, strong, ?, ? >, <sunny, warm,?,?,?,?>, <sunny, ?,?,strong,?,?>, <?, warm,?,strong,?,?>}

How Should These Be Classified



$\langle Sunny\ Warm\ Normal\ Strong\ Cool\ Change \rangle$

$\langle Rainy\ Cool\ Normal\ Light\ Warm\ Same \rangle$

$\langle Sunny\ Warm\ Normal\ Light\ Warm\ Same \rangle$

Classify Instance

	<S,W,N,S,C,C>	<R,C,N,L,W,S>	<S,W,N,L,W,S>	<S,C,N,S,W,S>
<S,?,?,?,?,?>	✓	×	✓	✓
<?,W,?,?,?,?>	✓	×	✓	×
<S,?,?,S,?,?>	✓	×	×	✓
<S,W,?,?,?,?>	✓	×	✓	×
<?,W,?,S,?,?>	✓	×	×	×
<S,W,?,S,?,?>	✓	×	×	×
confidence	6/6=2	0/6=0	3/6=0.5	2/6=0.33
Prediksi kelas	+	-	unknown	- (voting)

We need not enumerate every hypothesis in the version space in order to test whether each classifies the instance as positive or negative.

-If the new instance satisfies all members of S, it must also satisfy each of these more general hypothesis → predict +

-If the new instance satisfies none of the members of G → predict -

Find-S vs Candidate Elimination

- Hipotesis
  - CE: version space (set hipotesis yang spesifik sampai general) yang konsisten
  - Find-S: satu hipotesis yang paling spesifik yang konsisten
- Inconsistent training data: tidak ada perbaikan
  - Asumsi CE & Find-S:
    - training data akurat,
    - ruang hipotesis memiliki hipotesis target (inductive bias)

Inductive Bias

- Prior/implicit assumption
- Fundamental prop
- CE: target concept could be represented by a conjunction of attribute values.
- If assumption is correct: its classification of new instances will also be correct
- If assumption is incorrect: misclassify

Biased Hypothesis Space

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Cool	Change	Yes
Cloudy	Warm	Normal	Strong	Cool	Change	Yes
Rainy	Warm	Normal	Strong	Cool	Change	No

- Find-S: <?,W,N,S,C,C> → overly general. It incorrectly covers the third negative instance.
- CE: no hypothesis consistent
  S1,2: {<?,W,N,S,C,C>} → S3: {}
  G0,1,2: {<?,?,?,?,?,?>} → G3: {}
- Require a more expressive hypothesis space than conjunctive hypothesis: Sky=Sunny or Sky=Cloudy

Very Biased Hypothesis Space

- Very biased hypothesis space:
  - Ruang hipotesis: konjungsi nilai atribut (max 973 hipotesis)
  - 96 instances → possible dataset (subset): $2^{96} \approx 10^{28}$ → 10^{28} konsep target
- Desain ruang hipotesis yang lebih ekspresif: konjungsi + disjungsi + negasi
  - Sky=Sunny or Sky=Cloudy: <S,?,?,?,?,?,?> v <C,?,?,?,?,?,?>

CE: Minimal Specialization: G

	G	S
Inisialisasi	{<?,?,?,?,?,?>}	{<∅, ∅, ∅, ∅, ∅, ∅>}
<sunny, warm, normal, strong, warm, same, yes>	{<?,?,?,?,?,?>}	{~~<∅, ∅, ∅, ∅, ∅, ∅>,~~<sunny, warm, normal, strong, warm, same>}
<sunny, warm, high, strong, warm, same, yes>	{<?,?,?,?,?,?>}	{<sunny, warm, ?, strong, warm, same>}
<rainy, cold, high, strong, warm, change, no>	{<sunny, ?,?,?,?,?>, ~~<cloudy, ?,?,?,?,?>,~~ <?,warm,?,?,?,?>, ~~<? ,?,normal,?,?,?>,~~ ~~<? ,?,?,weak,?,?>,~~ <?,?,?,?,?,same>}	{<sunny, warm, ?, strong, warm, same>}
<sunny, warm, high, strong, cool, change, yes>	{<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, ~~<?,?,?,?,?,same>~~}	{<sunny, warm, ?, strong, ?, ? >}

$VS_{H,D}$= {<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, <sunny, warm, ?, strong, ?, ? >, <sunny, warm,?,?,?,?>, <sunny, ?,?,strong,?,?>, <?, warm,?,strong,?,?>}

CE: Noisy Training Data

	G	S
Inisialisasi	{<?,?,?,?,?,?>}	{<∅, ∅, ∅, ∅, ∅, ∅>}
<sunny, warm, normal, strong, warm, same, yes>	{<?,?,?,?,?,?>}	~~{<∅, ∅, ∅, ∅, ∅, ∅>,~~<sunny, warm, normal, strong, warm, same>}
<sunny, warm, high, strong, warm, same, ~~yes~~ no>	{~~<rainy, ?,?,?,?,?>, <cloudy, ?,?,?,?,?>, <?,cold,?,?,?,?>,~~ <?,?,normal,?,?,?>, ~~<?,?,?,weak,?,?>, <?,?,?,?,cool,?>, <?,?,?,?,?,change>~~}	{<sunny, warm, normal, strong, warm, same>}
<rainy, cold, high, strong, warm, change, no>	{<?,?,normal,?,?,?>}	{<sunny, warm, normal, strong, warm, same>}
<sunny, warm, high, strong, cool, change, yes>	~~{<?,?,normal,?,?,?>}~~	~~{<sunny, warm, ?, strong, ?, ?>}~~

$VS_{H,D}$= {}

CE: 1st Instance is Negative

	G	S
Inisialisasi	{<?,?,?,?,?,?>}	{<∅, ∅, ∅, ∅, ∅, ∅>}
<rainy, cold, high, strong, warm, change, no>	{<sunny, ?,?,?,?,?>, <cloudy, ?,?,?,?,?>, <?,warm,?,?,?,?>, <?,?,normal,?,?,?>, <?,?,?,weak,?,?>, <?,?,?,?,cool,?>, <?,?,?,?,?,same>}	{<∅, ∅, ∅, ∅, ∅, ∅>}
<sunny, warm, normal, strong, warm, same, yes>	{<sunny, ?,?,?,?,?>, ~~<cloudy, ?,?,?,?,?>,~~ <?,warm,?,?,?,?>, <?,?,normal,?,?,?>, ~~<?,?,?,weak,?,?>,~~ ~~<?,?,?,?,cool,?>,~~ <?,?,?,?,?,same>}	{~~<∅, ∅, ∅, ∅, ∅, ∅>,~~ <sunny,warm, normal, strong, warm,same>}
<sunny, warm, high, strong, warm, same, yes>	{<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, <?,?,?,?,?,same>}	{<sunny, warm, ?, strong, warm, same>}
<sunny, warm, high, strong, cool, change, yes>	{<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, ~~<?,?,?,?,?,same>~~}	{<sunny, warm, ?, strong, ?, ? >}

$VS_{H,D}$= {<sunny, ?,?,?,?,?>, <?,warm,?,?,?,?>, <sunny, warm, ?, strong, ?, ? >, <sunny, warm,?,?,?,?>, <sunny, ?,?,strong,?,?>, <?, warm,?,strong,?,?>}

Latihan 1:
Pembelajaran konsep “reading article”

article	crime	academic	local	music	reads
A1	true	False	false	true	true
A2	true	false	false	false	true
A3	false	false	false	false	false
A4	false	false	true	false	false
A5	true	true	false	false	true

Solusi

	G	S
Inisialisasi	{<?,?,?,?,?,?>}	{<∅, ∅, ∅, ∅, ∅,>}
<a1,true,false, false, true,+>	{<?,?,?,?,?,?>}	<a1,true,false, false, true>
<a2,true,false, false, false,+>	{<?,?,?,?,?,?>}	<?,true,false, false, ?>
<a3,false,false,false,false,->	{<?,true,?,?,?>}	<?,true,false, false, ?>
<a4,false,false,true,false,->	{<?,true,?,?,?>}	<?,true,false, false, ?>
<a5,true,true,false,false,+>	{<?,true,?,?,?>}	<?,true,?, false, ?>

$VS_{H,D}$= {crime, crime^~local}

THANK YOU