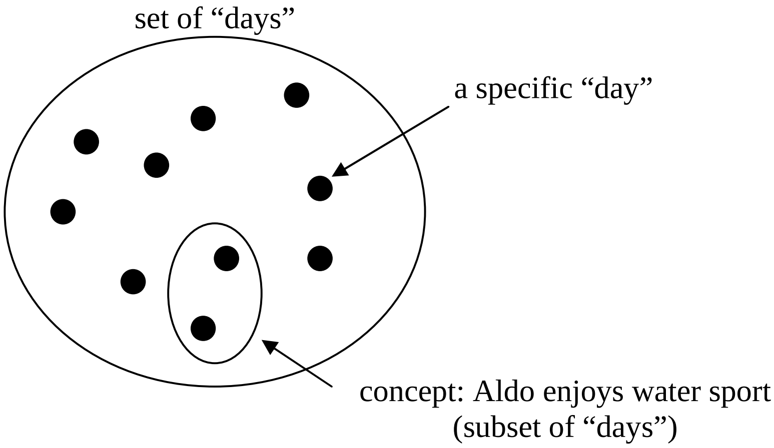
CS 367 Tutorial
25 August 2008
Week 6 (tutorial #4)
Carl Schultz

Material is taken from lecture notes (http://www.cs.auckland.ac.nz/compsci367s2c/lectures/index.html).

NB: recommended text for this part of the course is “Tom M. Mitchell, Machine Learning McGraw-Hill, New York, 1997”

- concept=some ‘interesting’ subset of objects or events
- e.g. “Days Aldo enjoys water sport”



- how to specify a subset? Describe objects using “attributes”, e.g.
    - “day” can have “sunny sky” or “rainy sky”
    - “day” can have “warm temp” or “cold temp”
    - …

attributes

Sky	Temp	Humid	Wind	Water	Forecast
sunny	warm	normal	strong	warm	same
sunny	warm	high	strong	warm	same
rainy	cold	high	strong	warm	change
sunny	warm	high	strong	cool	change

attribute value

distinct “day” events

- can describe a “day” as attribute values, e.g.
  - <sunny,warm,normal,strong,warm,same>= distinct “day”

- so, alternative definition of concept:
  - concept = Boolean-valued function
  - function input =attribute values (Sky=sunny,…)
  - function output =Boolean TRUE, FALSE

Sky	Temp	Humid	Wind	Water	Forecast	Enjoy
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

“day” is in concept?
TRUE / FALSE

attribute-value input

Boolean output

- task: learn Boolean-function from training examples
  - given certain input (Sky=sunny,…) our function will correctly return TRUE (matches concept) or FALSE (not a match)
  - real concept function is called “c”
  - we learn an approximation called “h” (hypothesis)
    - ? = any value acceptable
    - 0 = no value acceptable
  - E.g. h(x) = Sky=sunny AND Temp=warm AND Humidity=? …

The Inductive 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 unobserved examples.

- <u>search problem</u>: **find** best hypothesis out of all possible hypotheses
- e.g. attributes for “days” are

  – Sky (values Sunny, Cloudy, or Rainy)
  – Temp (values Warm or Cold)
  – Humidity (Normal or High)
  – Wind (Strong or Weak)
  – Water (Warm or Cool)
  – Forecast (Same or Change)

- each distinct “day” is a conjunction of attribute values
    - e.g. one distinct “day” has
        - Sky=sunny AND
        - Temp=warm AND
        - Humidity=normal AND
        - Wind=strong AND
        - Water=warm AND
        - Forecast=same
- How many distinct “days” are there?
        - Sky can take 1 of 3 values (sunny, cloudy, rainy)
        - Temp can take 1 of 2 values (warm, cold)
        - …

1	***Sunny***	Warm	Normal	Strong	Warm	Same
2	***Cloudy***	Warm	Normal	Strong	Warm	Same
3	***Rainy***	Warm	Normal	Strong	Warm	Same
4	Sunny	***Cold***	Normal	Strong	Warm	Same
…	…	…				

    - Number of combinations: 3×2×2×2×2×2 = 96 distinct “days”

- How many distinct hypotheses are there? E.g. one distinct hypothesis is

  h(x) = Sky=sunny AND
  Temp=warm AND
  Humidity=? AND
  Wind=strong AND
  Water=warm AND
  Forecast=same

    - for each attribute, hypothesis can put either
        - a particular attribute value
        - ?
        - 0

- number of combinations: 5×4×4×4×4×4 = 5120 **syntactically** distinct hypotheses

- some hypotheses are really saying the same thing, e.g.

$h_1(x)$ = Sky=0 AND
Temp=warm AND
Humidity=? AND
Wind=strong AND
Water=warm AND
Forecast=same

$h_2(x)$ = Sky=sunny AND
Temp=warm AND
Humidity=? AND
Wind=strong AND
Water=0 AND
Forecast=same

- neither of these hypotheses accept any "day", so **semantically** the same
- number of combinations:
  - 1 (hypothesis with one or more 0) +
  - 4×3×3×3×3×3 (add ? to each attribute)
  - = 973 **semantically** distinct hypotheses

[exercise]
Attributes and values for some animals are
Tail (yes, no)
Size (small, medium, large)
Skin (smooth, furry, slimy)
Legs (none, two, four)

a) how many distinct animals are there?
b) how many syntactically distinct hypotheses are there?
c) how many semantically distinct hypotheses are there?

- general vs. specific hypotheses

h_1=<sunny,?,?,strong,?,?>
h_2=<sunny,?,?,?,?,?>

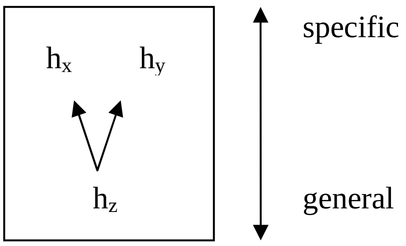
- h_2 is **more general** than h_1 because
  - whenever h_1 is TRUE, h_2 is also TRUE
  - and sometimes when h_2 is TRUE, h_1 is *not* TRUE
  - e.g. <**sunny**, warm, normal, **weak**, warm, same>
    - h_2 says TRUE but h_1 says FALSE

- the most general hypothesis is <?,?,?,?,?,?> …this is *always* TRUE
- the most specific hypothesis is <0,0,0,0,0,0> …this is *always* FALSE

[exercise]
Arrange the following hypotheses in order of generality

h_a=<sunny, warm, ?, strong, cool, same>
h_b=<sunny, ?, ?, strong, ?, ?>
h_c=<sunny, warm, ?, strong, ?, same>
h_d=<sunny, ?, ?, ?, ?, ?>
h_e=<sunny, warm, high, strong, cool, same>
h_f=<sunny, warm, ?, strong, ?, ?>
h_g=<?, ?, ?, ?, ?, ?>

- hypotheses only in a **partial** ordering
  - is h_x=<sunny, ?, ?, ?, ?, ?> more general than h_y=<rainy, warm, ?, ?, ?, ?>…?
  - no, because <rainy, warm, …> is TRUE for h_y and FALSE for h_x
  - h_z=<?, ?, ?, ?, ?, ?> is still **more general** than both h_x and h_y



[exercise]
Draw a graph of generality (partial order) for the following hypotheses. *Hint*: start with the most general and the most specific then fill in the gaps.

h_a=<sunny, warm, ?, ?, ?, ?>
h_b=<?, warm, ?, ?, ?, ?>
h_c=<sunny, ?, ?, ?, ?, ?>
h_d=<?, warm, ?, strong, ?, ?>
h_e=<rainy, warm, ?, strong, ?, ?>
h_f=<sunny, ?, ?, strong, ?, ?>
h_g=<sunny, warm, ?, strong, ?, ?>

- learning – finding the maximally specific hypothesis: “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 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

[exercise] ***Please note, the class handouts were wrong – this is the correct version***
Attributes and values for some animals are

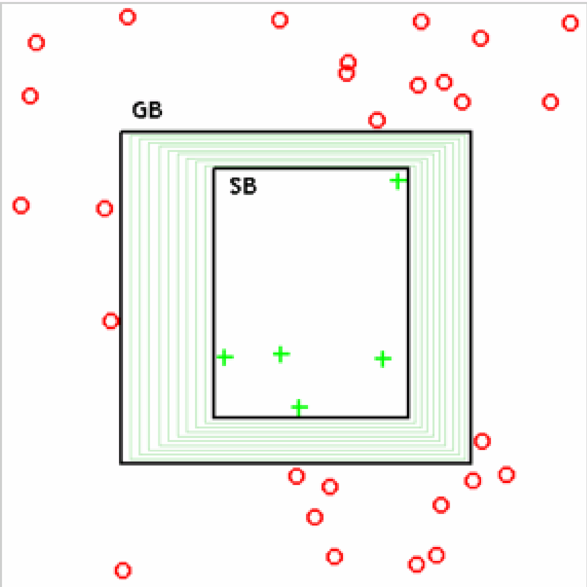
Tail (yes, no)
Size (small, large)
Skin (furry, slimy)
Legs (two, four)

Perform the “Find-S” algorithm to determine the maximally specific hypothesis for the following training data

1. <yes, small, slimy, four>, +
2. <no, small, slimy, four>, -
3. <yes, large, slimy, four>, +
4. <yes, small, furry, four>, +

- more than one hypothesis can match the training data
- version space: subset of hypotheses that are consistent with training examples
  - **general boundary**: set of hypotheses consistent with training examples that are *maximally* general
  - **specific boundary**: set of hypotheses consistent with training examples that are *minimally* general

The following image is from Wikipedia at
http://en.wikipedia.org/wiki/Version_space



Version space for a "rectangle" hypothesis language in two dimensions. Green pluses are positive examples, and red circles are negative examples. GB is the maximally **general** positive hypothesis boundary, and SB is the maximally **specific** positive hypothesis boundary. The intermediate (thin) rectangles represent the hypotheses in the version space.

- “Candidate Elimination” algorithm
  - **positive** examples → relax (generalise) **specific** boundary to accommodate
    - prune (remove) inconsistent hypotheses in general boundary
  - **negative** examples → tighten (specialise) **general** boundary to eliminate
    - prune (remove) inconsistent hypotheses in specific boundary

- good example:
http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/vspace/3_vspace.html