# TP 3
# Techniques of AI [INFO-H-410]
# Correction
v1.0.0

### Calculating the size of the hypothesis space

**Question 1.** Suppose there are m attributes in a learning task and that every attribute $i$ can take $k_i$ possible values. What will be the size of the hypothesis space?

**Answer:** syntactically different: $(k_i + 2)^m$ (+2 accounts for empty and don't care), semantically different: $1 + (k_i + 1)^m$ (full empty attribute + attributes with don't care)

### Order of training instances

**Question 2.** In candidate elimination, suppose you have $n$ training instances $T_1...T_n$. After the $n_{th}$ training instance, candidate elimination learned the boundaries $S$ and $G$. Will $S$ and $G$ differ or not when providing the training instances in reverse order: $T_n...T_1$? Explain why (not).

**Answer:** The concept of version space aims at invariance to instance order, keeping not a single concept description but a set of possible descriptions that evolves as new instances are presented, so order does not matter.

**Question 3.** What is the version space while tracing the candidate elimination algorithm with the following examples?

$Architecture \in \{Gothic, Romanesque\}$
$Size \in \{Small, Large\}$
$Steeples \in \{Zero, One, Two\}$

| Architecture | Size | Steeples | Classified ? |
|---|---|---|---|
| G | S | 2 | True |
| R | S | 2 | False |
| G | L | 2 | True |
| G | S | 0 | False |
| R | L | 2 | True |

**Answer:** $S_0 = \{\emptyset, \emptyset, \emptyset\}$ and $G0 = \{?, ?, ?\}$
$S_1 = \{G, S, 2\}$ and $G_1 = \{?, ?, ?\}$
$S_2 = \{G, S, 2\}$ and $G_2 = \{G, ?, ?\}$
$S_3 = \{G, ?, 2\}$ and $G_3 = \{G, ?, ?\}$
$S_4 = \{G, ?, 2\}$ and $S_4 = \{G, ?, 2\}$
$S_5 = G_5 = \emptyset$

**Rectangular version spaces and candidate elimination**

**Question 4.** Consider the instance space consisting of integer points in the $x, y$ plane and the set of hypotheses H consisting of rectangles. More precisely, hypotheses are of the form $a \leq x \leq b$, $c \leq y \leq d$, where $a$, $b$, $c$ and $d$ can be any integers. Consider the version space with respect to the set of positive (+) and negative (-) training examples:

| | |
|---|---|
| $-(1,3)$ | $-(2,6)$ |
| $+(6,5)$ | $+(5,3)$ |
| $-(9,4)$ | $-(5,1)$ |
| $+(4,4)$ | $-(5,8)$ |

a) What is the S boundary of the version space in this case? Write a diagram with the training data and the S boundary.

b) What is the G boundary of this version space? Draw that in the diagram as well.

c) Use python to find the S and the G boundary and plot them (you can use the template code).

d) Suppose the learner may suggest a new $x, y$ instance and ask the trainer for its classification. Suggest a query guaranteed to reduce the size of the version space, regardless how the trainer classifies it. Suggest one that will not.

e) Now assume you are the teacher, attempting to reach a particular target concept, $3 \leq x \leq 5, 2 \leq y \leq 9$. What is the smallest number of training examples you can provide so that the Candidate- Elimination algorithm will perfectly learn the concept?

**Answer:** (a) $S = \{4, 6, 3, 5\}$
(b) $G = \{3, 8, 2, 7\}$
(c) To reduce the VS: $(4, 6)$ or $(7, 3)$. Instances with no impact on the VS: $(5, 4)$ or $(3, 9)$.
(d) 6 points: +$(3, 9)$, +$(5, 2)$, -$(2, 5)$, -$(4, 1)$, -$(6, 5)$, -$(4, 10)$.

**Finding a maximally specific consistent hypothesis**

**Question 5.** Consider a concept learning problem in which each instance is a real number and in which each hypothesis is an interval over the reals. More precisely, each hypothesis in H is of the form: $a < x < b$ as in $4.5 < x < 6.1$, meaning that all real numbers between 4.5 and 6.1 are classified as positive examples and all others are classified as negative examples.

a) Explain informally why there cannot be a maximally specific consistent hypothesis for any set of positive training examples.

b) Suggest a modification to the hypothesis representation so this will not happen.

**Answer:** (a) $a < x < b$ is not a well defined hypothesis representation, (b) $a \leq x \leq b$