04_Concept_Learning

4.1 A Concept Learning Task

- Given:
  - Variable vector structure $X = \{x_1, x_2, \ldots, x_n\}$
    - A vector, that's a sequence of values, each having its own meaning.
    - e.g., $X = \{Sky, Temp, Humid, Wind, Water, Forecast\}$
  - Target Function $c : X \to \{0, 1\}$
    - e.g., $\{Sky, Temp, Humid, Wind, Water, Forecast\}$
      $\to \{EnjoySpt = Yes, EnjoySpt = No\}$.
  - Hypothesis h:
    - An example input in the format of X.
    - Conjunction of constraints on attributes, where each could be:
      - A specific value (e.g., $Water = Warm$), or
      - Any value (e.g., $Water =?$), or
      - No value (e.g., $Water = \emptyset$).
    - e.g., $h = \{Sky = Sunny, Temp = Warm, Humid = Normal,$
      $Wind = Strong, Water = Warm, Forecast = Same\}$
  - Training examples D:
    - Positive and negative examples of target function
    - $D = \{\langle X_1, c(X_1)\rangle, \langle X_2, c(X_2)\rangle, \ldots, \langle X_m, c(X_m)\rangle\}$
- Do:
  - Derive a general if-then rule to best summarize the training examples.

4.2 Find-S Algorithm 寻详算法

4.2.1 General-to-specific Ordering of Hypothesis

- [DEF] More-General-than-or-Equal-to 详于或等于
  - Let h_j and h_k be bool-valued functions defined over X.
  - Then, h_j is $more\_general\_than\_or\_equal\_to$ h_k ($h_j \geq_g h_k$) iff:
    - $\forall x \in X, h_k(x) = 1 \to h_j(x) = 1$
  - Moreover, h_j is strictly $more\_general$ than h_k iff:
    - $(h_j \geq_g h_k) \wedge (h_k \not\geq_g h_j)$

4.2.2 Find-S Algorithm

- From Specific to General
  Procedure:
- Initialize h to the most specific hypothesis in H.
- for X where $c(X) = 1$:
  - for x_i where $x_i \in X = \{x_1, x_2, \dots\}$:
    - if x_i in h doesn't satisfy x_i in X:
      - Replace x_i in h with a more general constraint that's accepted by X.
- Output: h.

Example:

Training Samples	S	Description
	$S_0 = \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$	
$X_1 = \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle, +$	$S_1 = \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle$	Initialized to be the **most** specific.
$X_2 = \langle Sunny, Warm, High, Strong, Warm, Same \rangle, +$	$S_2 = \langle Sunny, Warm, ?, Strong, Warm, Same \rangle$	$Normal \neq High$, h is not consistent with x_2 at position $Humid$.
$X_3 = \langle Rainy, Cold, High, Strong, Warm, Change \rangle, -$	$S_3 = S_2$	Ignore negative training samples.
$X_4 = \langle Sunny, Warm, High, Strong, Cool, Change \rangle, +$	$S_4 = \langle Sunny, Warm, ?, Strong, ?, ? \rangle$	$Warm \neq Cool, Same \neq Chan$. h is not consistent with x_4 at positions $Water$ and $Forcast$.

4.2.3 Disadvantages of Find-S Algorithm

1. **Unable to determine uniqueness of h.**
   无法确定h的唯一性
   1. It does not inform that whether the derived hypothesis is the **only** one in H that is consistent with D, or there are other hypothesis that are also consistent with D as well.
2. **Picks a maximally specific h.**
   总是选择最详尽的假设h
   1. Find-S algorithm always pick the most specific hypothesis among all the consistent ones.
3. **Does not display the noise in Training data itself.**
   无法指出训练数据D内部本身的不一致性

    1. Occasionally, training examples would contain at least one error, i.e., noise.
    2. That is, there may be two data samples in the training data that are inconsistent with each other.
    3. This kind of inconsistency can severely mislead Find-S algorithm, since it ignores the negative examples.
4. **Doesn't give alternative ways to find h.**
   只能选择多个相同详尽的h中的其中一个.
    1. There might be several hypothesis in H that's equally specific to the selected one.
    2. Therefore, there ought to be different paths to reach those goal.

4.3 List-Then-Eliminate Algorithm 列消算法

4.3.1 Consistency and Version Space 一致性与一致空间

- [DEF] A hypothesis h is ***consistent*** with a set of training examples D of target concept c iff:
    - $Consistent(h, D) \equiv \forall \langle X, c(X) \rangle \in D, h(X) = c(X)$.
    - That is, h produces the same corresponding output facing all input vectors in D.
- [DEF] The ***version space*** $VS_{H,D}$ with respect to:
    - Hypothesis space $H = \{h_1, h_2, \ldots, h_n\}$
        - That is, the set of all the possible hypothesis with respect to the format of X.
    - Training examples $D = \langle X_1, c(X_1) \rangle, \langle X_2, c(X_2) \rangle, \ldots, \langle X_m, c(X_m) \rangle,$
- is defined as:
    - $VS_{H,D} \equiv \{h \in H | Consistent(h, D)\}$
    - That is, a subset
        - from the set of all hypotheses H
        - that's consistent with all training examples in D.

4.3.2 Representation of Version Spaces

- [DEF] **Specific Boundary** (S) of $VS_{H,D}$
  一致空间的详尽边界
    - The set of $VS_{H,D}$'s **maximally specific** members of H that's consistent with D.
    - $S \equiv \{s \in H | Consistent(s, D)$
      $\wedge (\neg \exists s' \in H)[(s >_g s') \wedge Consistent(s', D)]\}$
        - s is consistent with D.
        - Doesn't exist s' who is consistent with D and also less general than the s.

        - 一致空间中没有比集合S中的假设更加**详尽**的假设了。
- [DEF] **General Boundary** (G) of $VS_{H,D}$
  一致空间的通用边界
    - The set of $VS_{H,D}$'s **maximally general** members of H that's consistent with D.
    - $G \equiv \{g \in H | Consistent(sg, D)$
      $\wedge (\neg \exists g' \in H)[(g' >_g g) \wedge Consistent(g', D)]\}$
        - g is consistent with D.
        - Doesn't exist g' who is consistent with D and also more general than g.
        - 一致空间中没有比集合G中的假设更加**通用**的假设了。

4.3.3 List-Then-Eliminate Algorithm

Procedure:

- For each training example $\langle x, c(x) \rangle \in D$ do:
    - if$c(x) = Positive$ do:
        - Remove from G any hypothesis inconsistent with d.
        - For each hypothesis $s \in S$ that's inconsistent with d:
            - Remove s from S.
            - Add to S a small minimal generalizations h of s s.t.,
                - h is consistent with d, and some member of G is more general than h.
            - Remove from S any hypothesis that
                - is more general than another hypothesis in S.
    - if$c(x) = Negative$ do:
        - Remove from S any hypothesis inconsistent with d.
        - For each hypothesis $g \in G$ that's inconsistent with d:
            - Remove g from G.
            - Add to G a small minimal specializations h of g s.t.,
                - h is consistent with d, and some other member of S is more specific than h.
            - Remove from G any hypothesis that
                - is less general than another hypothesis in G.

Example

Training Samples	S - Specific Bound	G - General Bound	Descriptio
	$S_0 = \{\langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\}$	$G_0 = \{\langle ?, ?, ?, ?, ?, ? \rangle\}$	

Training Samples	S - Specific Bound	G - General Bound	Descriptio
$X_1 = \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle, +$	$S_1 = \{ \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle \}$	$G_1 = G_0$	
$X_2 = \langle Sunny, Warm, High, Strong, Warm, Same \rangle, +$	$S_1 = \{ \langle Sunny, Warm, ?, Strong, Warm, Same \rangle \}$	$G_2 = G_1$	
$X_3 = \langle Rainy, Cold, High, Strong, Warm, Change \rangle, -$	$S_3 = S_2$	$G_3^{before} = \{ \langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle, \langle ?, ?, ?, ?, ?, ? \rangle, \langle ?, ?, ?, Strong, ? \rangle, \langle ?, ?, ?, Warm, ? \rangle, \langle ?, ?, ?, ?, Same \rangle, \}$ $\downarrow$ $G_3^{after} = \{ \langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle, \langle ?, ?, ?, ?, Same \rangle, \}$	Remove inconsisten hypothesis Here, the classificatic "-". Therefc to be "consistent the target v should be different frc the original
$X_4 = \langle Sunny, Warm, High, Strong, Cool, Change \rangle, +$	$S_4 = \{ \langle Sunny, Warm, ?, Strong, ?, ? \rangle \}$	$G_4 = \{ \langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle, \}$	$Same \neq Cl$, inconsiste

4.4 Summary

The relationships among "Spaces" could be defined as:

$$Hypothesis\ Space > Version\ Space > S \cup G$$

- Hypothesis Space
  - is the "Union" set with respect to X, the vector structure.
  - It contains all the possible hypothesis under the structure of X.
- Version Space
  - is a subset of hypothesis space.

    - Contains all the special hypothesis in hypothesis space, which are consistent to the given training data.
- $S \cup G$:
    - S and G are the most extreme hypothesis within a given version space, i.e., the boundaries.
        - S contains the most specific hypothesis in version space.
        - G contains the most general hypothesis in version space.
    - Any other hypothesis in version space can not be more general than the ones in G, and also can not be more specific than the ones in S.