VC-dimension

Exercises!

- Instance Space $X = \mathbb{R}^2$; Hypotesis Space $\rightarrow$ quadrilaterals in $\mathbb{R}^2$ with edges that are parallel to the axes:
  $\mathcal{H} = \{f_{a,b,c,d}(\vec{y}) | f_{a,b,c,d}(\vec{y}) =$
  1 if $(a \leq y_1 \leq b)$ and $(c \leq y_2 \leq d)$; -1 otherwise$\}$
  VC($\mathcal{H}$) ?
- Find $\mathcal{H}$ with a single parameter such that VC($\mathcal{H}$)=∞

Example of Inductive Bias for Concept Learning

Definition: A concept on an Instance Space X is defined as a boolean function on X.

Definition: An example of a concept c on the Instance Space X is defined as a couple $(x, c(x))$, where $x \in X$ and $c()$ is a boolean function.

Definition: Let h be a boolean function defined on the Instance Space X. We say that h satisfies $x \in X$ if $h(x) = 1$ (*true*).

Definition: Let h be a boolean function defined on the Instance Space X and $(x, c(x))$ an example of $c()$. We say that h is consistent with the example if $h(x) = c(x)$. Moreover we say that h is consistent with a set of examples Tr if h is consistent with every example in Tr.

Hypothesis Space: partial order

Definition: Let h_i and h_j be boolean functions defined on an Instance Space X. We say that h_i is more general than or equivalent to h_j ($h_i \geq_g h_j$) if and only if

$$(\forall x \in X)[(h_j(x) = 1) \rightarrow (h_i(x) = 1)]$$

Examples

- $l_1 \geq_g (l_1 \wedge l_2)$
- $l_2 \geq_g (l_1 \wedge l_2)$
- $l_1 \not\geq_g l_2$ e $l_2 \not\geq_g l_1$ (not comparable)

Exercise: learning of conjunctions of literals

Find-S Algorithm
/* it finds the more specific hypothesis which is consistent with the training set */

- input: training set Tr
- initialize h to the most specific
  $h \equiv l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \wedge \cdots \wedge l_m \wedge \neg l_m$
- for each positive training instance $(x, true) \in Tr$
  - remove from h any literal which is not satisfied by x
- returns h

Example of application: $m = 5$

(positive) Example	current hypothesis
	$h_0 \equiv l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \wedge l_3 \wedge \neg l_3 \wedge l_4 \wedge \neg l_4 \wedge l_5 \wedge \neg l_5$
1 1 0 1 0	$h_1 \equiv l_1 \wedge l_2 \wedge \neg l_3 \wedge l_4 \wedge \neg l_5$
1 0 0 1 0	$h_2 \equiv l_1 \wedge \neg l_3 \wedge l_4 \wedge \neg l_5$
1 0 1 1 0	$h_3 \equiv l_1 \wedge l_4 \wedge \neg l_5$
1 0 1 0 0	$h_4 \equiv l_1 \wedge \neg l_5$
0 0 1 0 0	$h_5 \equiv \neg l_5$

Notice that $h_0 \leq_g h_1 \leq_g h_2 \leq_g h_3 \leq_g h_4 \leq_g h_5$

Moreover, at every step the current hypothesis h_i is substituted by hypothesis h_{i+1} which constitutes a *minimal generalization* of h_i consistent with the current example.

Thus **Find-S** returns the most specific hypothesis which is consistent with *Tr*

Observations on **Find-S**

Find-S actually can be adapted to several and different Instance and Hypothesis Spaces.

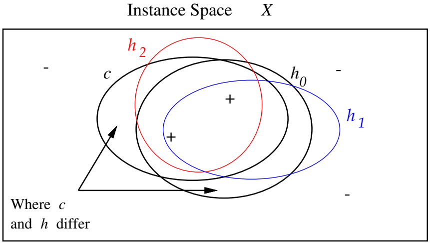
The basic idea of the algorithm is to compute a *minimal generalization* of the current hypothesis when this is not consistent with the current example.

Notice that every time the current hypothesis h is *generalized* leading to a new hypothesis h' ($h' \geq_g h$), all the positive examples seen in the past are satisfied by the new hypothesis h' (in fact, since $h' \geq_g h$, we have that $\forall x \in X,\ (h(x) = 1) \rightarrow (h'(x) = 1)$)

Finally, if the concept to be learned is contained in $\mathcal{H}$, all the negative examples (i.e., $c(x) = 0$) are automatically satisfied by the hypothesis returned by **Find-S** since that hypothesis is the most specific consistent hypothesis, i.e., the one that assigns the smallest number of 1's to instances in X.

Is there any valid motivation to prefer the most specific consistent hypothesis ?

Empirical Error

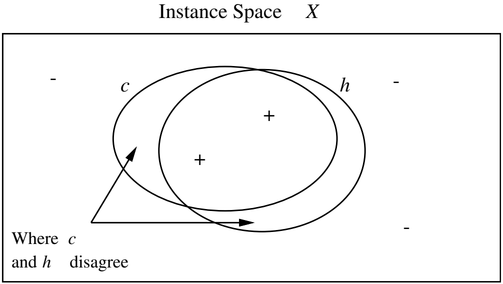


Def: The **Empirical Error** ($error_{Tr}(h)$) of hypothesis h with respect to Tr is the number of examples that h misclassifies:

$$error_{Tr}(h) \equiv \#\{(x, f(x)) \in Tr | f(x) \neq h(x)\}$$

Def: $h \in \mathcal{H}$ overfits Tr if $\exists h' \in \mathcal{H}$ such that $error_{Tr}(h) < error_{Tr}(h')$, but $error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$.

True Error



Def: The **True Error** ($error_{\mathcal{D}}(h)$) of hypothesis h with respect to target concept c and distribution $\mathcal{D}$ (to observe an input instance $x \in X$) is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$:

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$