Fix $k \in \mathbb{N}$. Let $\mathcal{L}_e = \Sigma^*$ and $\mathcal{L}_h = \wp\big(\texttt{factor}_k(\rtimes \cdot \Sigma^* \cdot \ltimes)\big)$. We will interpret hypotheses as sets of forbidden factors. Formally,

$$L(h) = \{w \in \Sigma^* \mid \texttt{factor}_k(\rtimes w \ltimes) \cap h = \emptyset\} \ .$$

Note we interpret $\texttt{factor}_k(w)$ as a function which returns *all* factors in w *up to* size k. Note also, if u is a factor of w we write $u \sqsubseteq w$.

For concreteness, in what follows we fix $k = 2$ and $\Sigma = \{a, b, c\}$.

Example 1. Let $h = \{c, bb\}$. Then $L(h)$ includes words *abab, aaa* and excludes words like *aca, abba*.

Remark 1. For all $g, s \in \mathcal{L}_h$, $g \preceq s$ whenever $L(s) \subseteq L(g)$. The "language of the grammar" corresponds exactly to the "cover" relation defined in (De Raedt, 2008, chap. 03).

Example 2.

1. Consider h_1, h_2 such that $h_1 \subseteq h_2$. So h_2 forbids all the factors h_1 does and maybe more. It follows that $L(h_2) \subseteq L(h_1)$ so $h_1 \preceq h_2$. In words, whenever h_2 forbids at least the same factors as h_1, then h_1 is more general than h_2.
2. Consider $h = \emptyset$. This hypothesis contains no forbidden factors. Thus, $L(h) = \Sigma^*$. It follows that for all $h' \in \mathcal{L}_h$ we have $h \preceq h'$. In other words, h is the most general hypothesis in $\mathcal{L}_h$.

Example 3. Define $\mathcal{Q}(h, D) = 1$ whenever $D \subseteq L(h)$ otherwise $\mathcal{Q}(h, D) = 0$. In other words, this quality criterion requires 100% success on the data sample.

- Observe this $\mathcal{Q}$ is a local quality criterion.
- Observe that $\mathcal{Q}$ is anti-monotonic. **Proof:** Suppose $g \preceq s$ and $\mathcal{Q}(s, D) = 1$. Then:

  1. $D \subseteq L(s)$ by definition of $\mathcal{Q}$.
  2. $L(s) \subseteq L(g)$ by definition of $g \preceq s$.

  By (1) and (2) above and transitivity of the subset relation, it follows $D \subseteq L(g)$ and hence $\mathcal{Q}(g, D) = 1$.

Remark 2. Consider $h = \{c\}$ and $h' = \{ac, bc, ca, cb, cc, \rtimes c, c \ltimes\}$. So h forbids c and h' forbids 2-factors containing c. It follows both grammars recognize all strings which do not contain a c; formally, $L(h) = L(h')$.

Example 4. Define $\mathcal{Q}'(h, D, \mathcal{L}_h) = 1$ iff $\mathcal{Q}(h, D) = 1 \wedge (\forall h' \in \mathcal{L}_h)[\mathcal{Q}(h', D) = 1 \rightarrow h' \preceq h]$; otherwise it equals 0. In words, h has to

1. cover the data, and
2. it has to be *more specific* than all the other hypotheses that cover the data.

$\mathcal{Q}'$ is a global quality criterion.

Example 5. Suppose $D = \{ababab, bababa, aaaa\}$. Consider the following hypotheses.

- $h_1 = \emptyset$. Clearly $\mathcal{Q}(h_1, D) = 1$.
- $h_2 = \{c\}$. Since $\mathcal{Q}(h_2, D) = 1$ and $h_1 \preceq h_2$ it follows h_1 does not satisfy $\mathcal{Q}'$.
- $h_3 = \{c, bb\}$. Since $\mathcal{Q}(h_3, D) = 1$ and $h_2 \preceq h_3$ it follows h_2 does not satisfy $\mathcal{Q}'$.
- $h_4 = \{ac, bc, ca, cb, cc, \rtimes c, c\ltimes, bb\}$. $\mathcal{Q}(h_4, D) = 1$.

I think both h_3 and h_4 satisfy $\mathcal{Q}'$.

The queston is: On what grounds can we prefer h_3 over h_4?

Example 6. Define $\mathcal{Q}''(h, D, \mathcal{L}_h) = 1$ iff

$$\mathcal{Q}'(h, D, \mathcal{L}_h) = 1 \wedge (\forall h' \in \mathcal{L}_h)\big[(\mathcal{Q}'(h', D, \mathcal{L}_h) = 1 \wedge f' \in h' \rightarrow (\exists f \in h)[f \sqsubseteq f'])\big]$$

In words, h has to

1. cover the data, and
2. it has to be *more specific* than all the other hypotheses that cover the data, and
3. contain factors that are contained in the factors of the other hypothesis that also cover the data and are more specific than the other hypotheses.

$\mathcal{Q}''$ is also global quality criterion. In Example 5, h_3 satisfies $\mathcal{Q}''$ but h_4 does not.

Now, I think we are in position to state a learning problem!

A Learning Problem

Given $\mathcal{L}_e, \mathcal{L}_h$ and $L(h)$ as defined above, and given some finite set $D \subsetneq \Sigma^*$, how can we find the hypotheses that satisfy $\mathcal{Q}''$?

De Raedt discusses "general-to-specific" and "specfic-to-general" approaches. The approach that string extension learning uses is specific-to-general. So the specific-to-general approach may identify which hypotheses satisfy $\mathcal{Q}'$ and then apply it again within the space of factors?

What the criterion $\mathcal{Q}''$ offers over $\mathcal{Q}'$ is to pick the most general way to state extensionally-equivalent hypotheses. In other words, while both "$\neg$np $\wedge$ $\neg$nt $\wedge$ $\neg$nk $\wedge$ $\neg$mp $\wedge$ $\neg$mt $\wedge$ $\neg$mk" and "$\neg$[`nasal`][$-$`sonorant`]" may satisfy $\mathcal{Q}'$ for some D, only the latter may satisfy $\mathcal{Q}''$.

Some Lessons

1. Being clear about the problem to be solved is important.
2. Being clear about the "spaces" is important. Here there are four: the space of examples ($\mathcal{L}_e$), the space of hypotheses ($\mathcal{L}_h$), the space of languages ($\wp(\mathcal{L}_e)$), and the space of factors!
3. Being clear about the orderings and relations among these spaces is important.

References

De Raedt, Luc. 2008. *Logical and Relational Learning*. Springer-Verlag Berlin Heidelberg.