John Langford 2/3/98

I find the notation used in theoretical learning to learn rather ad hoc and cumbersome. There are too many versions h, h, h_i, H, *H*, $\widehat{H}$, etc..(hmm.. I really don't know how to make all the different h's in latex yet). The only consistency I can find, is that the more esoteric the 'h', the more abstract the concept it represents. I can't see a nice way to extend this notation into the staged learning which interests me (learning to learn to ... to learn). I've been trying to figure out a better notation, which I'll show below. The notation draws from elements of GR and partial function application. I would really like your input on what you think is good or bad about the notation.

First note: an element in a range is specified by a lower case letter i.e. i, while the range is specified by an uppercase letter i.e. I. GR's implicit notation requires some plasticity of thought about the meaning of i.

I think it's easiest to specify the language I'm thinking of by example.

- $z : x \to y$ a labeled example
- $h : z \to [0, 1]$ the loss of an hypothesis on a labeled example
- z_I a vector of examples
- z_i an element from a vector of examples
- $h_K : z_{_I} \to h$ a hypothesis set. This can also be though of (implicitly) as a function which when given a vector of examples gives produces a hypothesis which when given an example, produces a loss. In other words: $h_K : z_I \to z \to [0, 1]$ or $h_K : z_I \to (z \to [0, 1])$
- h_k an element of a hypothesis set.
- $z_{IJ} : Z^{I^J}$ a vector of vector of examples.
- z_{ij} an element from a vector of vector of examples.
- $h_{KL} : z_{IJ} \to h_K$ a set of hypothesis sets. Or a set of hypothesis spaces. Or a function from a vector of tasks to a hypothesis space.
- h_{kl} an element of a vector of hypothesis vectors

Make sense? The only technical note is that I chose the innermost index to be the least abstract (always a vector of examples) and the outermost index to be most abstract (the vector of vector of ....).

Now here's some more radical extensions.

First, function application:

$h_{kl} z_{ij} = h_{kl}(z_{ij})$ for some element in h_{kl} and some element in z_{ij}. There is some trickery here, because the notation $h_{kl} z_{ij}$ is very suggestive that h_{kl} is dual to z_{ij} in the mathematical sense. Does this seem legitimate to you?

Next, the δ_k operator

- $h_{kl} z_{ij} \delta_j = \frac{1}{J} \sum_{j=1}^{J} h_{kl}(z_{ij})$ = a hypothesis from each hypothesis space applied to some example in each task space. This should correspond to your definition of $h_1 \bigoplus ... \bigoplus h_J(\overrightarrow{z})$ in definition 5.

- $h_{kl} z_{ij} \delta_j \delta_i = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{J} \sum_{j=1}^{J} h_{kl}(z_{ij})$ = the average loss on a vector of tasks of an $h_1 \bigoplus ... \bigoplus h_J(\overrightarrow{z})$. This should correspond to $\hat{er}_Z(h)$ in definition 5.

In general, δ_k means do an empirically weighted sum over the index.

It's also important to consider the true underlying distribution instead of the empirical distribution. We can do this with another operator, d_k. which is essentially the integral version of δ_k. Examples:

- $h_{kl} z_{ij} \delta_j d_i$ corresponds to $er_P(h)$ in definition 5.

- $d_P(h, h') = |(h_{kl} - h'_{kl}) z_{ij} \delta_j| d_i$

The Note that it is only valid to δ_k or d_k where k ranges over example indexes because we make no assumptions about the distribution of the hypotheses.

Do you see problems with this notation? Ambiguities which get introduced? The notation seems like it is more extensible, but I haven't tried writing out a proof in it.