I have a question about choosing the best learning model for the dataset with a \*stochastic noise\*. Let’s say I have the input data, which is a third degree polynomial + noise:

- $y = ax^3 + bx^2 + cx + d + e(x), \space \forall x\in X=[-1; 1]$

- $e(x)$ is a Gaussian noise: $\mu(e(x)) = 0, \space D(e(x)) = \sigma^2$

- dataset size is big enough: (for example, $N\_{in}>1000$ and we have another $N\_{out}>1000$ different points to estimate out-of-sample error $E\_{out}$.

And let the stochastic noise has the specific energy: it’s not too big and not too small so that after learning the best hypothesis set will be $H\_{sin} = \left\{ h(x):= a \* \sin (b\*x) + c \space |\space a, b, c \in \Re \right\}$, which contains the best hypothesis $h\_{sin \space best}(x)$ that gives the lowest $E\_{out}$ error even if I knew what was the generator of input data – third degree polynomial, - where the only unknown parameter is $\sigma$ (pertains to this stochastic noise).

The question is should I take $h\_{sin \space best}(x)$ as a winner for the given dataset? If the answer is YES, why not to \*walk through all possible hypothesis sets\* (see below) each time during the learning at all? I mean to keep this idea \*each time\* we must fit any type of data to pick up the best of the best learning hypothesis from:

- $H\_{sign}$

- $H\_{polynomial}$

- $H\_{sin}$

- $H\_{exp}$

- $H\_{log}$

- $H\_{hyperbola}$

… and another exotic types of functions.

For clarification, I don’t care about the running time (in fitting and testing). The only thing which is matter is the \*hypothesis accuracy\* (the lowest Eout).

As I think, if the answer to the first question is YES, the answer to the second one is also YES!

I have a question about choosing the best learning model for the dataset with a stochastic noise. Let’s say I have the next input data:

y = ax^3 + bx^2 + cx + d + e(x) for each x lies in [-1; 1]

e(x) is a Gaussian noise: mean(e(x)) = 0, deviation(e(x)) = sigma^2

dataset size is big enough (for example, Nin > 1000)

and we have another Nout > 1000 different points to estimate out-of-sample error Eout.

And let the stochastic noise has the specific energy: it’s not too big and not too small so that after learning the best hypothesis set will be Hsin = {h(x) = a\*sin(b\*x) + c; a,b,c - real}, which contains the best hypothesis hsin\_best(x), that gives the lowest Eout error even if I knew what was the generator of input data – third polynomial function, where the only unknown parameter is sigma (pertains to this stochastic noise).

The question is should I take hsin\_best(x) as the winner for the given dataset? If the answer is YES, why not to walk through all possible hypothesis sets (see below) each time during the learning at all? I mean to keep this idea each time we must fit any type of data to pick up the best of the best learning hypothesis.

Hsin =

Hexp =

Hlog =

Hpolynom =

Hhyperb =

… and another exotic types of functions.

For clarification, I don’t care about the running time (in fitting and testing). The only thing which is matter is the hypothesis accuracy (the lowest Eout).