02/01/2012

RANDOM SEARCH FOR HYPER-PARAMETER OPTIMIZATION

So Lots of people use manual + grid serch when looking to find optimal hyper-paramatet values. They sfart by explaing the optimization problem + Suting that is suffers from the curse of dimentionality when you are trying to solve it using grid search. But they cluim they will show how has the practical advuntages of Random serch (conceptiol simplicity, easy to implement, easy parallelism) + is more efficient in higher Dimentional space.

Take: $f(x,y) = g(x) + h(y) \approx g(x)$

the function has low effective Dimentionality In this case Random serch will work Better. Why Because with grid serch we are Serching every Dimention even those that are not important. Due to the randomness It searches those Relevent Subspaces Just as if it knew to Just Search those. And we cannot know which ones are relevent & which are not since it changes for each dataset (they found).

So they Say they will Redo the results of "Larochelle (2007)" where thire they did grid search & here they will do random search.

In (2007) it was grid search but here they define a distribution to mimic the Space of (2007) & then they Sample from it drawing 256 samples

Now for the results.

They coin this central trade of. Random search. exploration V.S exploitation.

In a restricted space less trials "models to sample + train" are needed to outperform grid search.

But better results can be achieved with a less Restrictive sample space but more trials are needed.

ex. of exploration v.s. exploitation.

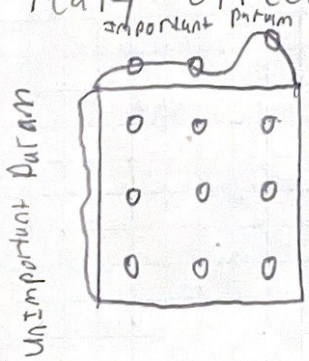
They only needed to do 8 trials to out perform grid search with no preprocessing.

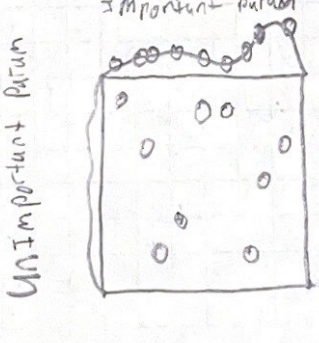
but when they used preprocessing (PCA and Normalization) they needed 32 trials, this indicates that their are many bad ways to preProcess. But When they did 64 trials the results where better than those found in the restricted smaller space (No preprocessing) but those were found faster obviously.

They also found that their are diffrent Shapes to the hyper-Parameter optimization function Ψ for Diffrent datasets.

ex. In the classic MNIST dataset it seems that after 4-8 trials the best model in each of the 4-8 trials (1 experiment) converges to the Same results. Meaning the good Region of Ψ is prob 1/4 to 1/8 of the entire Region. But for a problem like convex or MNIST Rotated with background images even experiments each with 16 to 32 trials had large variation in Results from the Best model. This means that here Ψ has smaller Region of good performance.

Now we start with saying why this Random Sampling works well. The claim this is Because Ψ has a low effective Dimentions. This is Where one or multiple dimentions do not realy effect the output. $f(x,y) = g(x) + h(x) \approx g(x)$





Now they will go on to show why Ψ often has low effective dimentionality.
LOOK AT Figure #7. They show a small fraction of Hyperparamaters matter for anyone dataset. but it differs from dataset to dataset which ones matter.

Then they not that there might Be a better way to Sample point that pure Random and they do it with low-discrepency Sets Sampling methods. Sobol seems to do a little better thur random.

When it comes to Deep Belif Networks (DBN) Random search can be competative but is not consistantly outperforming

In conclusion Random Search is more efficent than grid Search not because all Hyper-params are eqs important but percicly Because some are not. this is Why grid Search gives too many trials to the unimportant dimentions.