Basic Information **Networks for Approximation and Learning**

Author: Tomaso Poggio, Federico Girosi

Proceedings of the IEEE, Sep, 1990. P1481 - 1497 DOI: 10.1109/5.58326

Citation: 4267 (utill 2019 Jan 20, from Google scholar)

Summary This paper gives us the idea that the machine learning problem, to find an input-output mapping from a set of examples, can be considered as hyper-surface reconstruction in approximation theory. To solve this from the approximation point of view, this paper constructs a class of three-layer networks, regularization networks, to deal with nonlinear continuous mappings. That is finding the function $F(W, X)$, where W is the parameter to be learned. Notice that the choice of function F is important and which optimization function to use for a fixed F is also a big concern. The authors mainly focus on these two problems and do not cover the efficiency and parallel-ability.

The scheme went as follow. First, we can consider the function F as a superposition of a class of basic functions. Linear Fs lead to no-hidden-layer networks, linear suitable basses lead to poly nominals, splines and so on, while sigmoid Fs lead to multilayer networks as used in back propagation. Then the authors use function norm to describe the distance $\rho(F(W, X), f(X)) = ||F(W, X) - f(X)||_F$ between the desired solution and the data. Then they use learning algorithms such as gradient descent, conjugate gradient to learn the parameters that minimize the distance: $W^* = arg\min_W \rho(F(W, X), f(X))$.

At this point, we have not proved the existence of an exact representation for a certain kind of continuous functions. Although the Kolmogorov theory does not guarantee an exact representation, approximate representations could exist. However, a good approximate representation has some requirements. The authors claim that priori assumptions are needed, otherwise, we cannot know the number of examples we need for training. By adding a priori information cost, borrowed from standard regularization in approximation problems, the authors managed to keep a balance between the bias of predictions and the priori information. The new loss function is as follow: $\sum_i(\rho(F(W, X_i), f(X_i))) + \lambda||PF||^2$, where P associated with the embedding priori information and λ is a hyper parameter. Usually, we can assume smoothness of the target function, so we will use a norm to describe the priori loss. Then the authors also state the relation between Bayesian estimation. We can consider learning to be finding the conditional probability that minimizes the hypothesis complexity. To sum up, the authors proved that $Spline = StandardRegularization \subset MRFs \subset BayesianEstimator \subset ParallelNetworks$ and the RBFs can also be considered as a special case of the regularization networks.

In the next section, the authors mainly give the mathematical computation of this method. They use partial differential equations to solve the optimization problems. The main idea is to apply the Green functions and compute the integral of the differential equations. Also, they state that the features of priori become characters of the Green functions, such as translational and rotational invariance. And the regularization term can also be considered as reducing the oscillation of the prediction function.

In the IV section, the authors give some extensions of the regularization approach, including moving centers, using different basis functions and multiple scales, using weighted norm to offset the unbalanced different dimensions and modifying the loss to deal with noisy data points.

Strong points

1. The authors give a comprehensive analysis of regularization networks. The theoretical analysis is sound and convincible.

2. This paper generalizes machine learning problems into a larger class of approximation problems. Applying the regularization theory then become natural and easy to understand.

3. This method gives the relations among spline, multi-layer networks, Bayesian estimation, regularization theory, and approximation theory. This helps us understand these methods deeper.

4. The use of regularization enables people to add regularization terms to avoid overfitting, which has a great impact on the development of machine learning.

Weak points

1. The authors do not give too much analysis to the practical part. They mentioned that different optimization algorithms will perform differently for different basic functions. However, they do not give concrete examples and analysis. But I believe this point is important for other researchers to think of when they want to use this algorithm.

2. Although this method looks nice theoretically, computing the Green function and the inverse matrix is complex both in programming and computation perspective.