Assessing Generalization of Feedforward Neural Networks

Michael J. Turmon and Terrence L. Fine

School of Electrical Engineering, Cornell University, Ithaca, NY 14853

I. Introduction

Neural networks have been used to tackle what might be termed 'empirical regression' problems. Given independent samples of input/output pairs (x_i, y_i), we wish to estimate $f(x) = E[Y \mid X = x]$. The approach taken is to choose an approximating class of networks $\mathcal{N} = \{\eta(x; w)\}_{w \in \mathcal{W}}$ and within that class, by an often complex procedure, choose an approximating network $\eta(\cdot; w^*)$. The distance (in mean squared error) of this network from f can be separated into two terms: one for approximation or bias — choosing $\mathcal{N}$ large enough so that some $\eta(\cdot; w^0)$, say, models f well — and one for estimation or variance — how well the chosen $\eta(\cdot; w^*)$ performs relative to $\eta(\cdot; w^0)$. We address the latter term.

II. Problem Statement

Networks are parameterized by weight vectors $w \in \mathcal{W} \subseteq R^d$ and take inputs $x \in R^k$. In classification, network output is restricted to $\{0, 1\}$ while for regression it may be any real number. The complexity of the architecture $\mathcal{N}$ may be measured by the number of weights d or by its Vapnik-Chervonenkis (VC) dimension v. Performance of a network is measured by $\mathcal{E}(w) = E\left(\eta(x; w) - y\right)^2$ and the optimal net w^0 minimizes this. In practice, the law P is unknown so weights $w^* \in \mathcal{W}$ are chosen using the training set $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$ by minimizing $\nu_\mathcal{T}(w) = \frac{1}{n}\sum_{i=1}^n \big(\eta(x_i; w) - y_i\big)^2$.

The question of determining the relation between architecture complexity, estimation error, and training set size comes down to finding n large enough so that for a given d (or v), $\mathcal{E}(w^*) - \mathcal{E}(w^0) < \epsilon$ with high probability. We adopt this as a definition of reliable generalization. We can avoid dealing directly with the stochastically chosen network w^* by noting the triangle equality implies

$$0 \leq \mathcal{E}(w^*) - \mathcal{E}(w^0) \leq 2 \sup_{w \in \mathcal{W}} |\nu_\mathcal{T}(w) - \mathcal{E}(w)| \quad .$$

Vapnik [1] shows that $n = (9.2v/\epsilon^2)\log(8/\epsilon)$ is sufficient for reliable generalization. In cases where $\nu_\mathcal{T}(w^*) = 0$, this can be lowered [2] to $n = (5.8v/\epsilon)\log(12/\epsilon)$, but both are orders of magnitude higher than practice indicates.

III. Approximations via Poisson Clumping

For the large n we anticipate, the central limit theorem leads us to replace the original empirical process $\nu_\mathcal{T}(w) - \mathcal{E}(w)$ with the corresponding zero-mean Gaussian process $Z(w)$:

$$P(\| \, |\nu_\mathcal{T}(w) - \mathcal{E}(w)| \, \| > \epsilon) \simeq P(\| \, |Z(w)| \, \| > b)$$

where we have set $b = \epsilon\sqrt{n}$ and used the notation $\| \cdot \|$ for supremum over weights.

The Poisson clumping heuristic (PCH) [3] is a recently introduced tool for finding such exceedance probabilities. The PCH tells us that the region of weight space where $Z(w)$ exceeds level b is a group of clumps. The clump centers fall according to a Poisson process and the size $C_b(w)$ of a clump centered at w is chosen independently of all other clumps. The PCH leads to

$$P(\|Z(w)\| > b) \simeq \int_\mathcal{W} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)}\, dw \tag{1}$$

where $\bar{\Phi}$ is the complementary cdf of $N(0, 1)$ and $\sigma^2(w)$ is the variance of $Z(w)$. Loosely, the overall exceedance probability is a sum (integral) of the point exceedance probabilities, each scaled according to the number of weights that have exceedances with it.

This provides a means to get accurate approximations for the exceedance probabilities when the level b is large. For example, if network activation functions are twice differentiable and the variance has a unique maximum $\bar{\sigma}^2$ at $\bar{w} \in \mathcal{W}$, then $n = d\bar{\sigma}^2 K/\epsilon^2$ samples are sufficient for reliable generalization, where K is determined by P and $\mathcal{N}$. Explicit results for the problems of recognizing rectangles and halfspaces in R^k can also be obtained. These are again of order d/ϵ^2 but with constants far lower than previous upper bounds.

IV. Lower Bounds

These PCH-based estimates are of theoretical interest, but in practice evaluation of the constants is not possible due to ignorance of P. Now consider the following related tool for obtaining rigorous lower bounds to exceedance probabilities of $Z(w)$, where for simplicity we normalize $Z(w)$ by its standard deviation $\sigma = \sigma(w)$.

$$\begin{aligned} P(\|Z(w)/\sigma(w)\| > b) &= \int_\mathcal{W} \frac{\bar{\Phi}(b)}{E[D_b^{-1}|Z(w)/\sigma > b]^{-1}}\, dw \\ &\geq \bar{\Phi}(b) \int_\mathcal{W} \frac{1}{E[D_b|Z(w)/\sigma > b]}\, dw \end{aligned}$$

where D_b is the volume of $\{w : Z(w)/\sigma(w) > b\}$. Simple computations link this to the correlation $\rho = \rho(w, w')$ via

$$E[D_b|Z(w)/\sigma > b] \simeq \int_\mathcal{W} \bar{\Phi}((b/\sigma)\,\zeta)\, dw' \tag{2}$$

with $\zeta = \zeta(w, w') = \big((1-\rho)/(1+\rho)\big)^{1/2}$.

This link provides the basis for estimating the exceedance probability empirically, without knowledge of P. Using the training set, compute $(y_i - \eta(x_i; w))^2$ at w and w' for all n points. This yields an estimate of ρ and in turn an estimate of ζ which can be used to compute the integral (2). Simulations for the examples of recognizing rectangles and halfspaces show that reasonable estimates of sample size can be obtained in the absence of analytical information about P and $\mathcal{N}$.

References

[1] V. Vapnik, *Estimation of Dependences Based on Empirical Data.* Springer, 1982.

[2] A. Blumer et al., "Learnability and the Vapnik-Chervonenkis dimension," *Jour. Assoc. Comp. Mach.*, 36(4):929–965, 1989.

[3] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic.* Springer, 1989.