Regression with Gaussian Processes: Average Case Performance

Manfred Opper

Lehrstuhl fuer Theoretische Physik I, Universitaet Bayreuth, 95440 Bayreuth, Germany

Abstract. Recently, new models for regression and classification have been introduced which may be interpreted as neural networks in the limit of infinitely many parameters. For a regression model, the average case generalization performance is studied using a combination of information theoretic ideas and statistical mechanics methods.

1 Introduction

Artificial neural networks are often regarded as *semiparametric* statistical models. This means, that the number of parameters, say the weights of the network, is not fixed from the beginning, but must be adjusted to the data. Hence, in general, the number of parameters will depend on the number of examples which are used to train the network.

If the number of parameters is chosen too large, the network may tend to overfit, ie. it will have a bad performance on new inputs. Regularization techniques, which e.g. prune unnecessary weights are helpful here.

Recently, alternative models have been discussed, which in principle have infinitely many parameters to be estimated. By construction however, the network will only use an effective number of them which grows with the number of data. Within a Bayesian context, such nonparametric models have been derived for regression problems in [1] and for classification problems in [2]. Simulations indicate, that typically such models behave very well and do not seem to overfit the data.

Hence, it is interesting to study and understand their average case generalization behaviour. While exact asymptotic results can be obtained for networks with finitely many parameters, [3, 4, 5] the nonparametric setting has not been studied sofar (at least to my knowledge). In this contribution I will discuss the simplest case, a model for regression [1]. Based on a recently developped approach [5, 6], which combines information theoretic ideas with techniques from statistical mechanics, exact bounds on the so called Gibbs error will be derived.

2 Regression Problem

In this section, I discuss the well known problem, where an unknown functional relation $y = \theta(x)$ has to be estimated from noisy data. For simplicity, I as-

sume that the input variable x is a scalar quantity, the generalization to higher dimensions is straightforward.

When the inputs x are drawn independently at random from a distribution $f(x)$ and the observations y are corrupted by independent gaussian noise with variance σ^2, then the appropriate stochastic model is given by the probability

$$p_\theta(y|x) = \frac{e^{-\frac{(y-\theta(x))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}.$$

The goal of a learner is to give an estimate of the function $\theta(x)$, which is the unknown parameter of the model, based on a set of observed example data $D_t = ((x_1, y_1), \ldots, (x_t, y_t))$. Somewhat more general, one can say that the learner has to produce a predictive distribution $\hat{p}(y|x, D_t)$ of the dependent variable y, given x, when a set of training data D_t is observed [6].

This goal can be only achieved, when prior assumptions about the class of functions $\theta(x)$ are made. E.g. one could specify, how smooth $\theta(x)$ is expected to be.

3 The Bayesian Approach

A popular way to implement prior knowledge into statistical inference is the Bayes method. Here, a prior probability distribution $\mu(\theta)$ on the parameters of the model is specified. If we do not want to restrict ourselves to finite dimensional parametric classes, like polynomials of fixed degree, we have to specify measures on infinite dimensional function classes. In the approach of [1], Gaussian measures have been used because of their nice computational properties.

Assuming that the gaussian random functions θ have mean zero, their statistics is entirely determined by the covariance kernel K:

$$K(x, x') = \overline{\theta(x)\theta(x')} = \int d\mu(\theta)\ \theta(x)\theta(x'). \tag{1}$$

From the choice of the kernel K, one can model different smoothness properties of the functions $\theta(x)$.

Let me discuss two extreme examples: When $K(x, x') \propto \min(x, x')$, then $\theta(x)$ is a realization of the *Wiener process*, where the random functions θ are continuous but not differentiable. On the other hand, for the choice $K(x, x') \propto e^{-b(x-x')^2}$, which is of the type discussed by [1], all derivatives of $\theta(x)$ will exist. The hyperparameter b, which can be tuned to the data, determines the typical lengthscales over which the random functions vary.

Using Bayes rule, the predictive distribution in the Bayesian approach is given by

$$\begin{aligned}\hat{p}(y|x, D_t) &= \int d\mu(\theta|D_t)\ p_\theta(y|x) \\ &= \frac{\int d\mu(\theta) \prod_{i=1}^t p_\theta(y_i|x_i)\ p_\theta(y|x)}{\int d\mu(\theta) \prod_{i=1}^t p_\theta(y_i|x_i)}.\end{aligned} \tag{2}$$

Here $d\mu(\theta|D_t)$ is the posterior distribution of θ after the t data have been observed. At first sight, (2) looks like a complicated functional integral. But since the integral depends entirely on the t values $\theta(x_i)$, we have to average over a finite dimensional multivariate Gaussian distribution only. The predictive density (2) comes out to be gaussian with mean

$$\hat{y}(x) = \sum_i K(x, x_i) \sum_j (\{K + \sigma^2 I\}^{-1})_{ij} y_j, \tag{3}$$

which can be used to predict the most likely value for y [1]. Here, $K_{ij} = K(x_i, x_j)$.

4 Measure of Performance

In order to evaluate the average case performance of the Bayesian algorithm, one has to specify, how the success of learning should be measured. Assuming that θ^* is the true function from which the data are produced, we will choose the following information theoretic loss

$$L_t(\theta^*) = \left\langle \left\langle \log \frac{p_{\theta^*}(y_t|x_t)}{\hat{p}(y_t|x_t, D_{t-1})} \right\rangle \right\rangle_{D_t}, \tag{4}$$

where the brackets denote an average over the true distribution of the data. This entropic generalization error is the expected (over all data) relative entropy distance between predictive distribution and true distribution and plays an important role in information theory. The *cumulative* loss which is defined by summing up the losses for all t up to time n

$$R_n(\theta^*) = \sum_{t=1}^{n} L_t(\theta^*) = \left\langle \left\langle \log \frac{\prod_t^n p_{\theta^*}(y_t|x_t)}{\hat{p}(y_1, \ldots, y_n|x_1, \ldots x_n)} \right\rangle \right\rangle_{D_n} \tag{5}$$

can be related to a free energy like quantity, known from statistical mechanics, when the the predictive distribution

$$\hat{p}(y_1, \ldots, y_n|x_1, \ldots x_n) = \prod_{t=1}^{n} \hat{p}(y_t|x_t, D_{t-1}).$$

is given by the Bayesian formula (2). In this case we have

$$\hat{p}(y_1, \ldots, y_n|x_1, \ldots x_n) = \int d\mu(\theta) \frac{e^{-\sum_{i=1}^n \frac{(y_i - \theta(x_i))^2}{2\sigma^2}}}{(2\pi\sigma^2)^{\frac{n}{2}}}, \tag{6}$$

which looks like a partition function in statistical mechanics. Inserting into (5) yields

$$R_n(\theta^*) = -\left\langle \left\langle \log \int d\mu(\theta) \exp\left[-\sum_{i=1}^{n} \frac{\{(y_i - \theta(x_i))^2 - (y_i - \theta^*(x_i))^2\}}{2\sigma^2} \right] \right\rangle \right\rangle_{D_n}. \tag{7}$$

For special distributions of inputs in high dimensional spaces, it is possible to calculate the risk exactly, using the replica method of statistical mechanics. In this contribution I will use another approach, which has been recently developped to get general bounds on the entropic risk [5, 6].

5 Upper Bound on the Risk

Following [5], we can use the simple inequality

$$-\left\langle\left\langle \log \int d\mu(\theta) e^{-\beta H(\theta, D_n)} \right\rangle\right\rangle \leq -\log \int d\mu(\theta) e^{-\beta \langle\langle H(\theta, D_n)\rangle\rangle}. \tag{8}$$

to get an upper bound on the averaged free energy. The right hand side corresponds to the so called high temperature approximation well known in statistical mechanics. Applying this inequality to (7) yields

$$R_n(\theta^*) \leq -\log \int d\mu(\theta) \exp\left[-\frac{n}{2\sigma^2} \int dx f(x) \left(\theta(x) - \theta^*(x)\right)^2\right]. \tag{9}$$

6 Evaluation in Terms of Eigenvalues

The Gaussian integral (9) can be most easily evaluated by expanding the random functions as

$$\theta(x) = \sum_k w_k \sqrt{\lambda_k} \phi_k(x) \tag{10}$$

$$\theta^*(x) = \sum_k w_k^* \sqrt{\lambda_k} \phi_k(x) \tag{11}$$

where the w_l are independent standard normal variables which satisfy $\overline{w_k w_l} = \delta_{kl}$. The functions ϕ_k are chosen as normalized eigenfunctions of the integral equation

$$\int K(x, x') \phi_k(x') f(x') dx' = \lambda_k \phi_k(x), \tag{12}$$

so that we have

$$\overline{\theta(x)\theta(x')} = \sum_k \lambda_k \phi_k(x) \phi_k(x') = K(x, x') \tag{13}$$

and

$$\int dx f(x) \phi_k(x) \phi_l(x) = \delta_{kl}.$$

From this, we finally get the expression

$$R_n(\theta^*) \leq \frac{1}{2} \sum_k \log(1 + \frac{n}{2\sigma^2} \lambda_k) + \frac{n}{4\sigma^2} \sum_k \frac{(w_k^*)^2 \lambda_k}{1 + \frac{n}{2\sigma^2} \lambda_k}. \tag{14}$$

In practice, the density f of inputs may not be known. However, since we are evaluating an upper bound, f can be replaced by any function f_0 for which $f \leq f_0$!

7 Relation to Gibbs Error

We consider the following distance between a function $\theta(x)$ and the true function $\theta^*(x)$

$$\varepsilon_1(\theta, \theta^*) = \left\{ 1 - e^{\left[-\frac{1}{4\sigma^2} \int dx f(x) (\theta(x) - \theta^*(x))^2 \right]} \right\}. \tag{15}$$

This is related to the expected Hellinger distance [5] between the distributions $p_\theta(y|x)$ and $p_{\theta^*}(y|x)$. It can be shown [5] that

$$\begin{aligned} \varepsilon_1(n) &\doteq \int d\mu(\theta^*) \left\langle \left\langle \int d\mu(\theta|D_n) \varepsilon_1(\theta, \theta^*) \right\rangle \right\rangle_{D_n} \\ &\leq \int d\mu(\theta^*)\ L_n(\theta^*) \leq \frac{1}{n} \int d\mu(\theta^*)\ R_n(\theta^*). \end{aligned} \tag{16}$$

The expression on the left measures the average distance between the true unknown function $\theta^*(x)$ and an estimate $\theta(x)$, when the estimate is chosen at random from the posterior distribution $\mu(\theta|D_n)$. This random choice is known as the *Gibbs algorithm* in the Statistical Mechanics of learning [7]. The last inequality follows from the fact that the expected loss is nondecreasing. Note, that also an average over the prior distribution of true functions has to be included.

8 Lower Bound on the Gibbs Error

I will not explore the lower bound on the entropic risk given in [5, 6] but rather mention a direct lower bound on a related average error measure for the Gibbs algorithm. This measure is defined by

$$\varepsilon_2(\theta, \theta^*) = \left[\frac{1}{2\sigma^2} \int dx f(x) \left(\theta(x) - \theta^*(x) \right)^2 \right]. \tag{17}$$

Using a maximum entropy argument to be explained elsewhere, it is possible to show that

$$\begin{aligned} \varepsilon_2(n, \theta^*) &\doteq \left\langle \left\langle \int d\mu(\theta|D_t)\ \varepsilon_2(\theta, \theta^*) \right\rangle \right\rangle_{D_n} \geq \\ -\frac{1}{4} \frac{\partial}{\partial n} \log \int d\mu(\theta) \exp &\left[-\frac{n}{\sigma^2} \int dx f(x) \left(\theta(x) - \theta^*(x) \right)^2 \right]. \end{aligned} \tag{18}$$

This gaussian functional integral is of the same form as (9) and can be evaluated in the same way. We will also define

$$\varepsilon_2(n) = \int d\mu(\theta^*) \varepsilon_2(n, \theta^*).$$

Since for $\theta - \theta^*$ small, $2\varepsilon_1(\theta, \theta^*) \approx \varepsilon_2(\theta, \theta^*)$, we can expect that $\varepsilon_1(n)$ and $\varepsilon_2(n)$ will be of the same order of magnitude as $n \to \infty$.

9 Examples

a) For the Wiener process with $K(x,x') = \min(x,x')$ under the constant input density $f(x) = 1$ in the interval $[0,1]$, the solution to the integral equation (12) is well known and one finds

$$\lambda_k = \frac{1}{\pi^2(k+\frac{1}{2})^2},$$

for $k = 0, 1, 2 \ldots$. If we concentrate on the risk averaged over the prior, the sums over eigenvalues can be calculated analytically and one obtains

$$\int d\mu(\theta^*) R_n(\theta^*) \leq \frac{1}{2} \log\cosh\sqrt{n/(2\sigma^2)} + \frac{\sqrt{n/(2\sigma^2)}}{4} \tanh\sqrt{n/(2\sigma^2)} = \mathcal{O}(\sqrt{n}). \tag{19}$$

We find for the asymptotic decrease of the Gibbs errors

$$\varepsilon_1(n) \leq \mathcal{O}(n^{-\frac{1}{2}}) \tag{20}$$

and

$$\varepsilon_2(n) \geq \mathcal{O}(n^{-\frac{1}{2}}). \tag{21}$$

For the second case, we consider
b) smooth gaussian functions $\theta(x)$ with kernel $K(x,x') \propto e^{-(x-x')^2}$ and input density $f(x) = e^{-x^2}$. In this case, the integral equation (12) can be solved in terms of Hermite polynomials [9] and one gets

$$\lambda_k = \frac{\sqrt{\pi}}{\sqrt{2+\sqrt{3}}} \left(\frac{1}{\sqrt{2+\sqrt{3}}}\right)^k$$

so that

$$\int d\mu(\theta*) R_n(\theta*) \leq \mathcal{O}(\log^2(n)), \qquad n \to \infty \tag{22}$$

Similarly, the asymptotic decreases of the Gibbs errors are bounded by

$$\varepsilon_1(n) \leq \mathcal{O}(\frac{\log^2}{n}) \tag{23}$$

$$\varepsilon_2(n) \geq \mathcal{O}(\frac{\log}{n}). \tag{24}$$

It is interesting to compare (23) with the standard decrease of Gibbs errors obtained for network models with finitely many parameters D. In the realizable case, where the network is able to realize the unknown stochastic model perfectly, an asymptotic scaling $\varepsilon(n) = \mathcal{O}(\frac{D}{n})$ can be expected [3, 4, 5]. The result for the infinite dimensional case discussed in this contribution maybe interpreted in terms of a number D_{eff} of effective degrees of freedom which are explored by the algorithm. For the case of learning smooth functions $\theta(x)$, this number grows at least like $\mathcal{O}(\frac{\log}{n})$ but not faster than $\mathcal{O}(\frac{\log^2}{n})$ with the number of data.

10 Outlook

It will be interesting to generalize the present results from the regression problems to classification problems as discussed by [2]. For a two class problem, the probabilistic model is defined by

$$p_\theta(y = 1|x) = sigmo(\theta(x)),$$

which gives the probability that the input x belongs to class 1. *sigmo* is a sigmoidal function and $\theta(x)$ is again a gaussian process.

Finally, we can argue that these models have also a relationship with Vapnik's *Supportvector- Machines* [8], which for the case of a noisefree classification problem compute the output

$$y = \text{sign}(\theta(x)).$$

$\theta(x)$ has an expansion

$$\theta(x) = \sum_k w_k \sqrt{\lambda_k}\phi_k(x)$$

in terms of eigenfunctions of a positive definite kernel $K(x, x')$ like in (10). In contrast to the Bayesian case, the coefficients w_k are not gaussian random variables, but are chosen deterministically such that the stability

$$\kappa \doteq [\min_i \theta(x_i)y_i]$$

is maximal under the condition that $\sum_k w_k^2 = 1$. It is a challenge to extend our methods to this problem.

Acknowlegement
I thank Chris Williams for the solution of the integral equation (12).

References

1. Williams, C. K. I. and Rasmussen C.E.: Gaussian processes for regression, in: Advances in Neural Information Precessing Systems 8, ed. by D. S. Touretzky, M. C. Moser, and M. E. Hasselmo, published by MIT Press (1996).
2. Barber, D. and Williams, C. K. I.: Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo, in: Advances in Neural Information Precessing Systems 9, ed. by M.C. Mozer, M. I. Jordan and T. Petsche, published by MIT Press (1997).
3. Seung H., Sompolinsky H., and Tishby N. Physical Review A **45** (1992) 6056.
4. Amari, S. and Murata, N.; Neural Computation **5** (1993) 140.
5. Opper, M. and Haussler, D.; Phys. Rev. Lett. **75** (1995) 3772 .
6. Haussler, D. and Opper, M.: Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk, to be published by Annals of Statistics (1997).
7. Opper, M. and Kinzel, W.: Statistical Mechanics of Generalization in: Physics of Neural Networks ed. by J. L. van Hemmen, E. Domany and K. Schulten. Springer Verlag, Berlin (1996).
8. Vapnik, V.N.: The Nature of Statistical Learning Theory, Springer-Verlag, New York (1995).
9. Williams, C. K. I. : private communication.