A Variational Approach to Learning Curves

Dörthe Malzahn       Manfred Opper
Neural Computing Research Group
School of Engineering and Applied Science
Aston University, Birmingham B4 7ET, United Kingdom.
[malzahnd,opperm]@aston.ac.uk

Abstract

We combine the replica approach from statistical physics with a variational approach to analyze learning curves analytically. We apply the method to Gaussian process regression. As a main result we derive approximative relations between empirical error measures, the generalization error and the posterior variance.

1 Introduction

Approximate expressions for generalization errors for finite dimensional statistical data models can be often obtained in the large data limit using asymptotic expansions. Such methods can yield approximate relations for empirical and true errors which can be used to assess the quality of the trained model see e.g. [1]. Unfortunately, such an approximation scheme does not seem to be easily applicable to popular non-parametric models like Gaussian process (GP) models and Support Vector Machines (SVMs). We apply the replica approach of statistical physics to assess the average case learning performance of these kernel machines. So far, the tools of statistical physics have been successfully applied to a variety of learning problems [2]. However, this elegant method suffers from the drawback that data averages can be performed exactly only under very idealistic assumptions on the data distribution in the "thermodynamic" limit of infinite data space dimension. We try to overcome these limitations by combining the replica method with a variational approximation. For Bayesian models, our method allows us to express useful data averaged a-posteriori expectations by means of an approximate measure. The derivation of this measure requires no assumptions about the data density and no assumptions about the input dimension.

The main focus of this article is Gaussian process regression where we demonstrate the various strengths of the presented method. It solves some of the problems stated at the end of our previous NIPS paper [3] which was based on a simpler somewhat unmotivated truncation of a cumulant expansion. For Gaussian process models we show that our method does not only give explicit approximations for generalization errors but also of their sample fluctuations. Furthermore, we show how to compute corrections to our theory and demonstrate the possibility of deriving approximate universal relations between average empirical and true errors which might be of practical interest.

An earlier version of our approach, which was still restricted to the assumption of idealized data distributions appeared in [4].
2 Setup and Notation

We assume that a class of elementary predictors (neural networks, regressors etc.) is given by functions \( f(x) \). In a Bayesian formulation, we have a prior distribution over this class of functions \( f \). Assuming that a set of observations \( y_1, \ldots, y_m \) is conditionally independent given inputs \( x_1, \ldots, x_m \), we assign a likelihood term of the form \( \exp(-\beta h(f(x_i), y_i)) \) to each observation. Posterior expectations (denoted by angular brackets) of any functional \( F\{f\} \) are expressed in the form

\[
\langle F\{f\} \rangle = \frac{1}{Z_m} E \left( F\{f\} \exp \left( -\beta \sum_{i=1}^{m} h(f(x_i), y_i) \right) \right) \tag{1}
\]

where the partition function \( Z_m \) normalizes the posterior and \( E \) denotes the expectation with respect to the prior. We are interested in computing averages \( \langle F\{f\} \rangle_D \) of posterior expectations over different drawings of training data sets \( D = \{ (x_1, y_1), \ldots, (x_m, y_m) \} \) were all data examples are independently generated from the same distribution. In the next section we will show how to derive a measure which enables us to compute analytically approximate combined data and posterior averages.

3 A Grand-Canonical Approach

We utilize the statistical mechanics approach to the analysis of learning. Our aim is to compute the so-called averaged "free energy" \(-\ln Z_m \) which serves as a generating function for suitable data averages of posterior expectations. The partition function \( Z_m \) is

\[
Z_m = E \exp \left( -\beta \sum_{i=1}^{m} h(f(x_i), y_i) \right) \tag{2}
\]

To perform the average \( \ln Z_m \) we use the replica trick \( \ln Z_m = \lim_{n \to 0} \frac{\partial \ln Z_m^n}{\partial n} \), where \( Z_m^n \) is computed for integer \( n \) and the continuation is performed at the end [5]. We obtain

\[
Z_n(m) = [Z_m^n]_D = E_n \left( \left[ \exp \left( -\beta \sum_{a=1}^{n} h(f_a(x), y) \right) \right]_{(x,y)} \right)^m \tag{3}
\]

where \( E_n \) denotes the expectation over the replicated prior measure.

Eq.(3) can be transformed into a simpler form by introducing the "grand canonical" partition function \( \Xi_n(\mu) \)

\[
\Xi_n(\mu) = \sum_{m=0}^{\infty} \frac{\mu^m}{m!} Z_n(m) = E_n \exp(-H_n) \tag{4}
\]

with the Hamiltonian

\[
H_n = -\mu \left[ \exp \left( -\beta \sum_{a=1}^{n} h(f_a(x), y) \right) \right]_{(x,y)} \tag{5}
\]

The density \( e^{-H_n} \) evaluates all \( n \) replicas \( f_1, \ldots, f_n \) of the predictor at the same data point \( (x, y) \) and the expectation \( \langle \cdots \rangle_{(x,y)} \) is taken with respect to the true data density \( p(x, y) \).

The "grand canonical" partition function \( \Xi_n(\mu) \) represents a "poissonized" version of the original model with fluctuating number of examples. The "chemical potential" \( \mu \) determines the expected value of \( m = \frac{\partial \ln \Xi_n(\mu)}{\partial \mu} \) which yields simply \( \mu = \ln m \) for \( n \to 0 \). For sufficiently large \( m \), we can replace the sum in Eq. (4) by its dominating term

\[
\ln Z_n(m) \approx \ln \Xi_n(\mu) + m(\ln m - 1) - m \mu \tag{6}
\]
The grand-canonical ensemble was chosen such that Eq.(5) can be rewritten as an integral

4.1 Variational Equations

The brackets \( \langle \ldots \rangle_0 \) denote averages with respect to the effective measure which is induced by the prior and \( e^{-H_n^0} \) and acts in the space of replicated variables. As is well known, the first two leading terms in Eq.(7) present an upper bound [6] to \( -\ln \Xi_n(\mu) \). Although differentiating the bound with respect to \( n \) will usually not preserve the inequality, we still expect \(^1\) that an optimization with respect to \( H_n^0 \) is a sensible thing to do [7].

4 Variational Approximation

For most interesting cases, the partition function \( \Xi_n(\mu) \) can not be computed in closed form for given \( n \). Hence, we use a variational approach to approximate \( H_n \) by a different tractable Hamiltonian \( H_n^0 \). It is easy to write down the first terms in an expansion of the "grand canonical" free energy with respect to the difference \( H_n - H_n^0 \)

\[
-\ln \Xi_n(\mu) = -\ln E_n e^{-H_n} + \langle H_n - H_n^0 \rangle_0 - \frac{1}{2} \left( \langle (H_n - H_n^0)^2 \rangle_0 - \langle H_n - H_n^0 \rangle_0^2 \right) + \ldots \quad (7)
\]

The brackets \( \langle \ldots \rangle_0 \) denote averages with respect to the effective measure which is induced by the prior and \( e^{-H_n^0} \) and acts in the space of replicated variables. As is well known, the first two leading terms in Eq.(7) present an upper bound [6] to \( -\ln \Xi_n(\mu) \). Although differentiating the bound with respect to \( n \) will usually not preserve the inequality, we still expect \(^1\) that an optimization with respect to \( H_n^0 \) is a sensible thing to do [7].

4.1 Variational Equations

The grand-canonical ensemble was chosen such that Eq.(5) can be rewritten as an integral over a local quantity in the input variable \( x \), i.e. in the form \( H_n = -m \int dx \mathcal{H}(x, \{ f_a(x) \}) \) with

\[
\mathcal{H}(x, \{ f_a(x) \}) = \int dy \ p(y, x) \exp \left( -\frac{1}{2} \sum_{a=1}^n h(f_a(x), y) \right) . \quad (8)
\]

We will now specialize to Gaussian priors over \( f \), for which a local quadratic expression

\[
H_n^0 = \int dx \sum_{ab} \frac{1}{2} \eta_{ab}(x) f_a(x) f_b(x) + \sum_a r_a(x) f_a(x) \quad (9)
\]

is a suitable trial Hamiltonian, leading to Gaussian averages \( \langle \ldots \rangle_0 \). The functions \( \eta_{ab}(x) \) and \( r_a(x) \) are variational parameters to be optimized. It is important to have an explicit dependence on the input variable \( x \) in order to take a non uniform input density into account.

To perform the variation of the first two terms in Eq.(7) we note that the locality of Eq.(8) makes the "variational free energy" \( -\ln E_n \exp(-H_n^0) + \langle H_n - H_n^0 \rangle_0 \) an explicit function of the first two local moments

\[
K_{ab}(x) = \langle f_a(x) f_b(x) \rangle_0 \quad r_a(x) = \langle f_a(x) \rangle_0 . \quad (10)
\]

Hence, a straightforward variation yields

\[
m \frac{d\langle H(x, f(x)) \rangle_0}{dK_{ab}(x)} = \eta_{ab}(x) \quad m \frac{d\langle H(x, f(x)) \rangle_0}{dr_a(x)} = r_a(x) . \quad (11)
\]

To extend the variational solutions to non-integer values of \( n \), we assume that for all \( a \) the optimal parameters are replica symmetric, i.e., \( r_a(x) = r(x) \) as well as \( \eta_{ab}(x) = \eta(x) \) for \( a \neq b \) and \( \eta_{aa}(x) = \eta_0(x) \). We also use a corresponding notation for \( K_{ab}(x) \) and \( r_a(x) \).

\(^1\) Guided by the success of the method in physical applications, for instance in polymer physics.
4.2 Interpretation of $H_n^0$

Note, that our approach is not equivalent to a variational approximation of the original posterior. In contrast, $H_0$ contains the full information of the statistics of the training data. We can use the distribution induced by the prior and $e^{-H_0}$ in order to compute approximate combined data and posterior averages. As an example, we first consider the expected local posterior variance $\sigma^2(x) \doteq (\langle f(x) \rangle - \langle f(x) \rangle)^2 p$. Following the algebra of the replica method (see [5]) this is approximated within the variational replica approach as

$$\sigma^2(x) = \lim_{n \to 0} \left( \langle f_n^2(x) \rangle_0 - \langle f_n(x) f_n(x) \rangle_0 \right) = K_0(x) - K(x).$$

Second, we consider the noisy local mean square prediction error of the posterior mean predictor $\hat{f}(x) = \langle f(x) \rangle$ which is given by $\epsilon(x, y) = \langle (f(x) - y)^2 \rangle_D$. In this case

$$\epsilon(x, y) = \lim_{n \to 0} \left( \langle f_n(x) f_n(x) \rangle_0 + y^2 - 2y\langle f_n(x) \rangle_0 \right) = K(x) - \sigma^2(x) + \langle r(x) - y \rangle^2.$$

We can also calculate fluctuations with respect to the data average, for example

$$[(\hat{f}(x) - y)^2 (\hat{f}(x') - y')^2]_D = \lim_{n \to 0} \left( \langle f_n(x) - y \rangle \langle f_n(x') - y' \rangle \right)_0.$$

5 Regression with Gaussian Processes

This statistical model assumes that data are generated as $y_i = f(x_i) + \xi_i$, where $\xi$ is Gaussian white noise with variance $\beta^{-1}$. The prior over functions has zero mean and covariance $C(x, x') = E[f(x)f(x')]$. Hence, we have $h(f, y) = \frac{1}{2}(y - f(x))^2$. Using the definitions Eqs.(12,13), we get

$$\frac{\partial \langle H(x, \{f_n(x)\}) \rangle_0}{\partial n} = -\frac{p(x)}{2} \left\{ \ln(1 + \beta\sigma^2(x)) + \frac{\int dy \, p(y|x) \epsilon(x, y)}{\beta^{-1} + \sigma^2(x)} \right\}$$

which yields the set of variational equations (11). They become particularly easy when the regression model uses a translationally invariant kernel $C(x-x')$ and the input distribution is homogeneous in a finite interval. The variational equations (11) can then be solved in terms of the eigenvalues of the Gaussian process kernel.

[8, 9] studied learning curves for Gaussian process regression which are not only averaged over the data but also over the data generating process $f$, using a Gaussian process prior on $f$. Applying these averages to our theory and adapting the notation of [9] simply replaces in Eq.(15) the term $\int dy \, p(y|x) \epsilon(x, y)$ by $\epsilon(x) + \sigma^2(x)$ while $\sigma^2(x) \equiv \epsilon(x)$.

5.1 Learning Curves and Fluctuations

Practical situations differ from this “typical case” analysis. The data generating process is unknown but assumed to be fixed. The resulting learning curve is then conditioned on this particular “teacher” $f$. The left panel of Fig.1 shows an example. Displayed are the mean square prediction error $\epsilon$ (circle and solid line) and its sample fluctuations (error bars) $\Delta \epsilon$ with respect to the data average (cross and broken line). The target $f_\ast$ was a random but fixed realization from a Gaussian process prior with a periodic Radial Basis Function kernel $C(x, x') = \sum_k \exp(-(x-x'-k)^2/2l^2), l = 0.1$. We keep the example simple, e.g the Gaussian process regression model used the same kernel and noise $\beta^{-1} = \sigma^2_\ast = 0.01$. The inputs are one dimensional, independent and uniformly distributed $x \in [0, 1]$. Symbols represent simulation data. A typical property of our theory (lines) is that it becomes very accurate for sufficiently large number of example data.