Quadratic Concepts

Jörg C. Lemm
Institut für Theoretische Physik I, Universität Münster
D–48149 Münster, Germany
E-mail: lemm@uni-muenster.de

Abstract

Gaussian processes have recently become popular in the empirical learning community. They encompass many classical methods of statistics, e.g., radial basis functions or various splines, and are technically convenient due to the fact that Gaussian integrals can be performed analytically and the corresponding saddle point equations (to be solved for a maximum a–posteriori approximation or empirical risk minimization) are linear. At the same time, however, this technical advantage implies a severe practical limitation. Linear equations, i.e., quadratic and therefore convex error surfaces, forbid the implementation of genuine non–convex prior knowledge. For example, one may want to implement the belief that individual earthquakes or electrocardiograms tend to be similar to either a prototype A OR a prototype B. Quadratic concepts (or non–zero mean Gaussian processes in Bayesian interpretation) are used as building blocks for implementation of non–convex prior knowledge exploring possibilities to go beyond Gaussian processes. Continuous data functions are introduced based on the observation that learning is, implicitly or explicitly, based on an infinite number of data. It is shown how empirical measurement of an infinite amount of data is possible by a–posteriori control.

1 Introduction

In the setting of empirical learning available training data are used to obtain information about new test situations. Clearly, this generalization from training to test data requires knowledge about their dependencies. In this paper, such knowledge concerning dependencies between training and test data, in some contexts also known as rules or axioms, will be called *prior knowledge*.

To be specific, we consider a typical function approximation problem. Assume a given a set of training data $D_T = \{(x_i, y_i|1 \le i \le n\}$ sampled i.i.d. from an unknown but fixed "true state of Nature". The aim is to obtain an approximation function $h(x)$ to predict unknown outcomes y for test situations $x \in \mathcal{X}$ by $y = h(x)$.

Relying on the fact that the generalization ability of any learning system is crucially based on the dependencies it implements, our goal has to be a strict empirical measurement and control of the prior or "dependency" data D_P which represent our prior knowledge. It is interesting to note that for the common situation with an infinite number of potential test situations $x \in \mathcal{X}$ also the number of dependencies to be controlled empirically becomes infinite.

Empirical measurement of an infinite number of data, however, seems at first glance impossible. On the other hand, for an infinite set $\mathcal{X}$ of test situations any learning system has to use an infinite number of data, either explicitly or implicitly. To discuss this empirical measurement problem let us have a closer look at two examples:

1. A simple bound on $h(x)$ for every x like $h(x) \le a$, $\forall x \in \mathcal{X}$ corresponds for infinite $\mathcal{X}$ to an infinite number of data.

2. Similarly, deviations from exact symmetries may be bounded. Let σ denote a one–to–one transformation on $\mathcal{X}$. Then, $||h(x) - h(\sigma(x))|| < a$ describes a bound on the deviation from an exact symmetry under σ, also corresponding to an infinite number of data. The prototypical example is smoothness i.e., approximate symmetry under infinitesimal translations.

Like the number of training data can only be finite for practical reasons also the number of actually appearing test situations in the future can only be finite. The key point is now, that from all possible dependencies only those related to actual test situations have to be controlled empirically. Measurement devices, for example, only have to be active at the, always finite, number of actually appearing test situations. Thus, there is an easy way to enforce bounds without actually measuring an infinite number of times. To be specific, bounds are often the consequence of using realistic, non–ideal measurement devices:

1. A simple bound $h(x) < a$ is implemented by using a measurement device with cut–off at a.

2. Assume that, in addition to an upper and lower bound on h, input noise or input averaging with respect to σ is present in the measurement device we use. That means that we do not have perfect control over the value of x. In that situation fixing x at the measurement device still allows that $\sigma(x)$ produces the observable result y. The resulting effective function is necessarily smooth with respect to the transformations σ which generate the input noise. (Compare [1, 4, 5].)

Thus, one can say: *Infinite a–priori information can be empirically measured by a–posteriori control at the time of testing.* From this point of view, related to that of constructivism, (also infinite) a-priori information can (and should) be explicitly related to empirical control of the application situation.

2 Quadratic concepts

As a first step empirical dependency control requires an explicit and readable formulation of prior knowledge. By explicit formulation we mean in the following the expression of prior knowledge directly in terms of the functions values $h(x)$, like it is done in regularization theory or for stochastic processes. In an implicit implementation of dependencies, on the other hand, single function values $h(x)$ are not parameterized independently. Examples include neural networks, linear, additive or tensor models. Also the realization of learning algorithms can induce dependencies, e.g., due to restricted initial conditions and stopping rules.

In the regularization framework an approximation $h(x)$ is chosen to minimize a regularization or error functional $E(h)$. Prior knowledge can be represented by a regularization term added to the training error. A typical example of a smoothness related regularization functional for d–dimensional $x = \{x_{(1)}, \cdots, x_{(d)}\}$ is

$$E(h) = \sum_i^n \Big(y_i - h(x_i)\Big)^2 + \int_{-\infty}^{\infty} d^d x \sum_{l=1}^d \sum_k \gamma_k \left(\frac{\partial^k h(x)}{\partial x_{(l)}^k}\right)^2 . \tag{1}$$

It is well known that (for approximation problems with log–loss [5]) a regularization or error functional can be interpreted as up to a constant proportional to the negative logarithm of the posterior probability $p(h|D) \propto p(D|h) \propto e^{-\beta E(h)}$ with $D = D_T \cup D_P$. For differentiable functionals E the optimal solution h^* is found by setting the functional derivative of E with respect to h to zero. Clearly, the functional E must depend on each single function value $h(x)$ to determine a complete function $h \in \mathcal{H}$. Thus, for an infinite set $\mathcal{X}$ the x–integral (or a sum over x, respectively) corresponds to an infinite number of (prior) data terms. For quadratic terms like in Eq.(1) the common structure of the training data and prior terms can best be seen by writing them as scalar products. Let angular brackets denote scalar products and matrix elements of symmetric operators, e.g. $\langle t|h\rangle = \int dx\, t(x)h(x)$ and $\langle t\,|K|\,h\rangle = \int dx\, dx' t(x)K(x,x')h(x') = \langle t|h\rangle_K$ where x can be d–dimensional. We obtain for a mean–square training error term

$$(y_i - h(x_i))^2 = \int_{-\infty}^{\infty} d^d x\, \delta(x - x_i)\Big(h(x) - t_i(x)\Big)^2 = \langle h - t_i\,|P_i|\,h - t_i\rangle,$$

with $t_i \equiv y_i$ and P_i the projector into the space of functions on x_i. Similarly, a typical prior term becomes by partial integration for vanishing boundary terms

$$\int_{-\infty}^{\infty} d^d x \sum_{l=1}^d \left(\frac{\partial^k h(x)}{\partial x_l^k}\right)^2 = -\big\langle h - t_0\,\big|\Delta^k\big|\,h - t_0\big\rangle,$$

with $t_0 \equiv 0$ and negative (semi) definite d–dimensional Laplacian Δ with kernel $\Delta(x,x') = \delta(x - x')\sum_{l=1}^d \frac{\partial^2}{\partial x_l^2}$. *Notice that here, analogously to finite training data y_i, a "continuous data" or template function function $t_0 \equiv 0$ corresponding to infinite data has been introduced explicitly.* Indeed, there is no reason except technical convenience why the special function $t_0 \equiv 0$ should be preferred over other functions $t(x)$ which are not identically equal to zero. As long as only one continuous data or template function $t(x)$ enters the functional, $t_0 \equiv 0$ can always be achieved by solving for $h(x) - t(x)$ instead for $h(x)$. This is possible within classical regularization functionals of the form (1). If, however, more than one template function appears in a functional, then only one of them can be set to zero. Such cases will be studied below. We summarize: *Regularization* *functionals have to contain continuous data (template functions) to determine a function h completely. In the special case of classical functionals of form (1) the template function can be chosen identically zero, so it does not appear explicitly in the formalism. In the general case there is no reason not to consider non–zero template functions.* Having motivated the use of continuous data or template functions we now define a quadratic concept. Assume h to be in some Hilbert space $\mathcal{H}$.

Definition. (Quadratic concept.) A quadratic *concept* is a pair (t, K) consisting of a *template function* $t(x) \in \mathcal{H}$ and a real symmetric, positive semi–definite operator on $\mathcal{H}$, the *concept operator* K. The operator K defines a *concept distance* $d^2(h) = \langle h - t\,|K|\,h - t\rangle = ||h - t||_K^2$ on subspaces where it is positive definite. The maximal subspace in which the positive semi–definite K is positive definite is the *concept space* H_K of K. The corresponding hermitian projector P_K in this subspace H_K is the *concept projector*.

Remark 1 (Approximate symmetries): Typical concept operators K are related to symmetries. Let for example $Sh(x) = h(\sigma(x))$ with σ a permutation within $\mathcal{X}$, i.e., one–to–one. Then $K = (I - S)^T (I - S)$, with identity I and ^{T} denoting the transpose, defines a symmetry concept $d_S^2(h) = \langle h - Sh|h - Sh\rangle$. Similarly, assume $S(\theta) = e^{\theta s}$ to be a continuous symmetry (Lie) group, parameterized by θ and with infinitesimal generators s. Then an infinitesimal symmetry concept for $S \approx 1 + \theta s$ is defined by $d_s^2(h) = \langle sh|sh\rangle$. For infinitesimal translations (smoothness) $s = \partial/\partial x$.
Remark 2 (Gaussian processes): A quadratic concept defines a Gaussian process according to $p(D|h) \propto e^{-d^2(h)/2}$ with covariance operator K^{-1}. We remark however, that while Gaussian processes can also be defined for continuous $\mathcal{X}$ [2, 7] we do not discuss continuum limits for the non–Gaussian extensions below. In these cases we refer to a lattice approximation.
Remark 3 (Support vector machine): Expanding $h = \sum_m n_m \Phi_m(x)$ in a basis of eigenfunctions of $K = \sum_m \lambda_m \Phi_m \Phi_m^T$ one obtains $E(h) = \sum_i \left(y_i - h(x_i)\right)^2 + \langle h\,|K|\,h\rangle = \sum_i \left(\sum_m n_m \Phi_m(x) - y_i\right)^2 + \lambda_m |n_m|^2$. Replacing the mean–square training error by Vapnik's ϵ–insensitive error yields a support vector machine with kernel K [6]. In general, flat regions of $E(h)$ as they appear in the ϵ–insensitive error and other robust error functions have the technical advantage that they do not contribute to the gradient and can be ignored within a saddle point approximation.
Remark 4 (Templates): Templates $t(x)$ can be constructed directly by experts. They also can represent a structural hypothesis realized by a parameterized learning system like a neural network. Templates can be used for transfer by choosing them as the output of a learning system trained for a similar situation. For finite spaces templates can in principle be estimated by sampling.
Remark 5 (Covariances): Covariances K^{-1} can be given directly by experts. They do not necessarily have to be local but can include non–local correlations. As already remarked they are often constructed from symmetry considerations. For finite spaces covariances can in principle also be estimated by sampling.

3 Combination of concepts

Classical regularization functionals consist of a sum of quadratic concepts. In a probabilistic interpretation this corresponds to a combination by AND. Typically, for example, a training data term AND a prior term is approximated by h. The sum of quadratic concepts, however, is again a quadratic concept. Analogously, a product of Gaussians is Gaussian. Straightforward calculation shows that a sum of squared distances d_i^2 with concept operators K_i can be written $E(h) = \sum_i^N E_i(h) = \sum_i^N d_i^2(h)/2 = d^2(h)/2 + V$, with squared distance $d^2(h) = \langle h - \bar{t}\,|K|\,h - \bar{t}\rangle$, template average $\bar{t} = K^{-1}\tilde{t}$, $\tilde{t} = \sum_i^N K_i t_i$, $K = \sum_i K_i$, and h–independent minimal component energy $V = \left(\sum_i^N \langle t_i\,|K_i|\,t_i\rangle - \langle \bar{t}\,|K|\,\bar{t}\rangle\right)/2$, which has the structure of a variance up to a factor $2N$. The linear stationarity equation for a functional $E = d^2(h)/2 + V$ reads $0 = K(\bar{t} - h) = \tilde{t} - Kh$. For positive definite, i.e., invertible K, this has solution $h = \bar{t} = K^{-1}\tilde{t}$ which can be solved in a space with dimension smaller or equal to n [7, 3] . Now let us present types of non–convex error functionals.

Example: Consider an image reconstruction task where we expect the image of a face. Thus, we may choose concepts with partial template functions for eyes, nose and mouth and require the reconstructed image to approximate the given pixel data AND the eye, nose and mouth templates. Typically, however, the constituents of a face can appear in many different variations. Eyes may be open OR closed, blue OR brown but also translated, scaled or otherwise deformed. Such OR–like combinations of alternative concepts are examples of non–convex prior knowledge.

In a probabilistic interpretation of alternative concepts representing disjunct events indexed by i, this yields the mixture model

$$E_M(h) = -\ln \sum_i^N p(D, i|h) - c = -\ln \left(\sum_i^N p(i) e^{-\beta(E_i(h) - F_i) + c} \right), \qquad (2)$$

with component energies $E_i(h) = d_i^2(h)/2 = \langle h - t_i\,|K_i|\,h - t_i\rangle/2$, arbitrary constant c and $p(D, i|h) = p(i)e^{-\beta(E_i(h) - F_i)}$. If i–dependent, the normalization integrals $\beta F_i = -\ln\left(\int dt_i e^{-\beta\, E_i}\right)$ have to be calculated so they do not interfere with the mixture probabilities $p(i)$. In that case the model has the structure of a disordered ("spin–glass–like") system. The model has the stationarity condition $0 = t_M(h) - K_M(h)h$, with $K_M = \sum_i p(D, i|h)K_i$ and $t_M = \sum_i p(D, i|h)K_i t_i$. The parameter β is known as inverse temperature and interpolates between a convex AND at high temperature and a non–convex OR at low temperature.

Products are another possibility to implement OR–like structures leading to technically convenient polynomial models

$$E_{LG}(h) = \frac{1}{2} \prod_{i=1}^N d_i^2(h). \qquad (3)$$

The stationarity equation is $K_{LG}(h)h = t_{LG}(h)$ where $K_{LG}(h) = \sum_i M_i(h)K_i$ and $t_{LG}(h) = \sum_i M_i(h)K_i t_i$ with $M_i(h) = \prod_{k\neq i} d_k^2(h)$. The model resembles the Landau–Ginzburg treatment of phase transitions in statistical mechanics. Numerical studies of solvable mixture and polynomial models have been performed and will be reported elsewhere.

4 Conclusions

The a–priori information required for function approximation corresponds to an infinite number of data. The paper discusses template functions as example of an explicit implementation of infinite data. An infinite number of data can be empirically measured by a–posteriori control. While in classical convex regularization functionals template functions representing infinite number of data can be chosen identically zero, general non–convex a–priori information requires explicitly non–zero template functions.

Acknowledgements The author was supported by a Postdoctoral Fellowship (Le 1014/1–1) from the Deutsche Forschungsgemeinschaft and a NSF/CISE Postdoctoral Fellowship at the Massachusetts Institute of Technology. He also wants to thank Federico Girosi, Tomaso Poggio, Jörg Uhlig, Achim Weiguny, and Chris Williams for discussions.

References

[1] Bishop, C.M.: Training with noise is equivalent to Tikhonov regularization. *Neural Computation* **7** (1), 108–116, 1995.

[2] Doob, J.L.: *Stochastic Processes.* Wiley, New York, 1953.

[3] Girosi, F., Jones, M., and Poggio, T.: Regularization Theory and Neural Networks Architectures. *Neural Computation* **7** (2), 219–269, 1995.

[4] Leen, T.K.: From Data Distributions to Regularization in Invariant Learning. *Neural Computation* **7**, 974–981, 1995.

[5] Lemm, J.C. Prior Information and Generalized Questions. A.I.Memo No. 1598, C.B.C.L. Paper No. 141, Massachusetts Institute of Technology, 1996, (available at http://planck.uni–muenster.de:8080/~lemm).

[6] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, 1995.

[7] Wahba, G.: Spline Models for Observational Data. SIAM, 1990.