Fuzzy Interface with Prior Concepts and Non–convex Regularization

Jörg C. Lemm[1]

Abstract

The fundamentals of an interface are presented to include prior knowledge in learning problems. This is done in the framework of regularization theory, working not with a restricted parameterization but directly with the function values itself. Technically, single components of prior knowledge are hereby represented by so called quadratic prior concepts. The error functional which has to be minimized is obtained by probabilistic or fuzzy logical operations of prior concepts. Commonly used regularization approaches correspond to a probabilistic implementaion of AND yielding convex error surfaces. The paper presents an approach to go beyond such classical regularization functionals by including OR–like combinations of prior concepts. Resulting in non–convex error functionals this requires non–linear stationarity equations to be solved. A great variety of corresponding learning algorithms can be constructed. The paper concentrates on the fundamental definitions.
Keywords: Prior information, Regularization, annealing methods, Landau–Ginzburg model.

1 Introduction

We choose in this paper the function approximation setting as formalization of learning. Hereby, we assume to have available a set of training data $D = \{(x_i, y_i) | 1 \le i \le n\}$. The learning problem consist in finding a function $h(x)$ so we can predict outcomes also for x not contained in the training data by $y = h(x)$. For a face detection task, for example, the independent or input variable x_i may be an array of grey level values representing example images, while the dependent or output variable y_i can simply be a binary variable encoding the attributes "image contains a face" and "image contains no face". In the analysis of time series, $t_i = x_i$ is a time of measurement while the measured value y_i might be an exchange rate of the Dollar at time t_i.

Learning, i.e. generalization from training data to new situations requires always the presence of prior knowledge. Besides stationarity assumptions for the empirical distribution of y for fixed x prior information has to specify the dependencies which allow to infer from one function value $y = h(x)$ to function values $y' = h(x')$ with $x' \ne x$. Hence prior information is especially needed in cases where not all values of independent variable x are available for training or x can take more values than training data are available. This is always the case for continuous x. In typical object recognition tasks, for example, the input space has very high dimensionality (e.g. images) while the number of available training data may be small (e.g. because costly to obtain). Nevertheless objects have often very specific characteristica and human observers may name a lot of them. Typical examples are constituents, like eyes, nose, and mouth for faces, and their, e.g. spatial, relations. Methods from fuzzy set theory and fuzzy logic [1] can be used to quantify such linguistic variables or "fuzzy" concepts, like "eyes" or "typical distance between eyes". Beginning with a set of elementary concepts, more complicated concepts can be created by a fuzzy

[1]Institut für Theoretische Physik I, Wilhelm–Klemm–Str.9, D–48149 Münster, Germany, phone: (+49)251/83–34922, Fax (+49)251/83–36328, lemm@uni-muenster.de. The paper has been accepted for the 5th International Workshop on Fuzzy–Neuro Systems 98.

logical combination of elementary concepts. For example we may require a face to have mouth, nose AND eyes, but all those parts may appear in different variants, e.g. an eye may be open OR closed. The problem consists 1. in the quantification of the elementary (fuzzy) concepts and 2. their probabilistic or fuzzy logical combination.

Despite the fact that generalization is essentially based on prior information most implementations of it are implicit in a chosen restricted parameterization (e.g. by choosing a specific neural network architecture) or implicit in the chosen algorithm (e.g. early stopping used for training of neural networks may be seen as regularization method). Implicit implementations are difficult to analyze and do usually not allow adaption to specific prior knowledge. As we want to deal with varying and relatively specific prior information we have to include it explicitly in the error functional. This is done, for example, in regularization theory (from a Bayesian point of view) but usually restricted to relatively simple (quadratic, i.e. convex) smoothness constraints.

The approach for explicit implementation of prior knowledge presented here has two steps. In a first step, components of prior information are formalized by defining template functions, being locally or globally defined function prototypes, and distances therefrom. In a second step the components are combined by the logical operations AND and OR, similar to fuzzy logical methods. Implementing a probabilistic AND yields the classical linear regularization methods, which include for example splines and Radial Basis Functions. For many problems knowledge formulated in terms of a probabilistic OR is available. Objects, for example consist of constituents, which may appear in one of several relatively well defined variants, e.g. eyes may be open OR closed, blue OR brown. Especially important are translation, scaling and other deformations. The implementation of a probabilistic OR yields nonlinear stationarity equations being the technical topic of this paper.

More than one approach is possible for implementing OR–like combinations. Interpreting the individual components as (shifted and rescaled) log–probabilities results in a mixture model over a function space, resembling a physical system at finite temperature. But one also can directly construct an effective total posterior by fuzzy logical–like combination of prior components. Introducing a temperature–like parameter this corresponds to the Landau–Ginzburg treatment of phase transitions. The resulting equations are similar, for example, to equations obtained in quantum mechanical scattering theory.

2 Templates and concepts

Technically, one is interested in function approximation in finding a function or hypothesis $h(x)$ which minimizes a given error/energy functional $E[h]$. Using a Bayesian interpretation[2] we understand the error functional up to a constant as proportional to the posterior log–probability for the function h given training data D and prior informations D_0, i.e.

$$p(h|D,D_0) \propto e^{-\beta E[h]}. \tag{1}$$

The error functional usually depends as well on a finite number of training data D as also on additional prior informations D_0. Special cases of function approximation include density estimation where h has to fulfill an additional normalization condition or classification or pattern recognition where the function $h(x)$ takes only discrete values representing the possible classes or patterns of x.

Let us consider as example an error functional with mean square data terms and a typical smoothness constraint for d–dimensional x

$$E[h] = \frac{1}{2}\sum_i^n (y_i - h(x_i))^2 + \frac{\lambda}{2}\int_{-\infty}^{\infty} d^d x \, \sum_{l=1}^{d} \left(\frac{\partial h(x)}{\partial x_l}\right)^2 . \tag{2}$$

[2]Minimizing the error functional $E[h]$ corresponds to a maximum posterior or saddle point approximation for the expected risk in a full Bayesian decision theory.

We now rewrite the terms to show their common form. A mean square error term can be written using the bra–ket notation for scalar products and matrix elements of symmetric operators

$$(y_i - h(x_i))^2 = \int_{-\infty}^{\infty} d^d x\, \delta(x - x_i)(h(x) - t_i(x))^2 = < h - t_i | P_i | h - t_i > \qquad (3)$$

with projector $P_i(x,x') = \delta(x - x_i)\delta(x - x')$ and t_i the constant function $t_i(x) \equiv y_i$. Thus, (3) is a square distance of h from t_i at point x_i.

For the smoothness constraint in (2) we find the quite similar form

$$\int_{-\infty}^{\infty} d^d x \sum_{l=1}^{d} \left(\frac{\partial h(x)}{\partial x_l}\right)^2 = - < h - t_0 | \Delta | h - t_0 > \qquad (4)$$

with zero function $t_0(x) \equiv 0$ and $\Delta(x,x') = \delta(x - x') \sum_{l=1}^{d} \frac{\partial^2 h(x)}{\partial x_l^2}$ the negative (semi) definite d–dimensional Laplacian. Here we used partial integration assuming vanishing boundary terms. Also (4) can be interpreted as distance of h from the zero function t_0 but now over all x. Hence, we obtain for functional (2)

$$E[h] = \frac{1}{2} \sum_{i}^{n} < h - t_i | P_i | h - t_i > - \frac{\lambda}{2} < h - t_0 | \Delta | h - t_0 > . \qquad (5)$$

These examples motivate the following general definitions. Let H denote a Hilbert space[3] of hypothesis functions $h(x)$ of d–dimensional x.

Definition 1 ((prior) concept (t,d) with (prior) template t and template distance d): *A prior concept is a pair (t,d) with $t(x)$ a function in H and $d : (H \times H) \to \{0\} \cup \mathbb{R}^+$ a ("distance") functional with $d[t,t] = 0$. Note that this allows $d[t,h] = 0$ for $h \neq t$. We write $d[t,h] = d[h]$. The function $t(x) \in H$ will be called a (prior) template.*

Template functions will be used to represent function prototypes. Notice that templates include standard training data which is the reason for the brackets around the word "prior". We are especially interested in distances quadratic in h, for which the functional derivative with respect to h is linear. Such distances can be defined by positive semi–definite operators O. Such operators have a decomposition $O = W^T W$ with W invertible if O positive definite. More precisely, $\sqrt{< h|O|h >} = ||h||_O = ||Wh||$ defines a semi–norm on H with $||h||_O = 0$ if h is in the zero space of O, i.e. if $Oh = 0$. Typical W are projectors into the space of training data like in (3) and generators of infinitesimal transformations of continuous Lie groups, like the gradient ∇ for translations in (4) with $\nabla^T \nabla = -\Delta$ under appropriate boundary conditions. Thus, we define:

Definition 2 (quadratic (prior) concept (t,O) with template distance operator O): *A quadratic (prior) concept is a pair (t,O) with $t(x)$ a function in H and a symmetric and positive semi–definite operator $O = W^T W$ which will be called a template distance operator. O defines the square template distance:*

$$d^2[h] = < h - t | O | h - t > = < W(h - t) | W(h - t) > = ||W(h - t)||^2 = ||h - t||_O^2 . \qquad (6)$$

Thus, a quadratic concept defines a d–dimensional Gaussian process with $p(h) = e^{-d^2[h]/2}$ and covariance operator O^{-1}. Its matrix elements are sometimes also called Greens function, propagator or two-point correlation function. The Laplacian (4), for example, corresponds to the Wiener measure known from Brownian motion or

[3] We neglect here specific problems appearing for infinite Hilbert spaces and use for example the δ–function freely in the manner of physicists. We may also remark that we have to resort finally to a finite representation of functions $h(x)$ in all numerical calculations.

diffusion and is also used as kinetic energy for Euclidean scalar fields in physics (see for example [2]. The zero modes of O represent the projections of h which do not contribute to $d^2[h]$. The projector P_i in the mean square error term (3), for example, measures the distance only at (training data) point x_i. Also continuous template functions may be restricted to subspaces, e.g. parts of an image or a specific resolution.

Definition 3 (template space H_t and template projectors P_t): *The maximal subspace on which the positive semi–definite O is positive definite will be called the template space H_t of O. The corresponding hermitian projector P_t in this subspace H_t, i.e. $P_t(H) = H_t$, $P_t(H \setminus H_t) = 0$, and $P_t^2 = P_t$ will be called template projector.*

Hence O commutes with the template projector $[P_t, O] = P_t O - O P_t = 0$. Maximality of H_t means that $1 - P_t$ is the projector in the zero space of O i.e. $O(H \setminus H_t) = 0$. Our aim is to built an error functional $E[h]$ depending on h over square distances $d^2[h]$.

3 Combining quadratic concepts

Having defined the fundamentals of prior concepts in a bit more detail we now give a short and rather compact outline of combinations of concepts. A more detailed paper is in preparation. We also include some comments on learning algorithms.

Commonly one wants h to approximate training data AND an additional regularization (e.g. smoothness) constraint. Interpreting the functional $E[h]$ up to a constant as posterior log–probability, an AND of independent events is realized by a sum. A sum of quadratic terms $d_i^2 = <h - t_i|O_i|h - t_i>$ is again quadratic and one finds easily

Proposition 1 (probabilistic AND) *AND–ing quadratic concepts by adding their square distances d_i^2 results in $E = \sum_i^N d_i^2[h] = d^2[h] + E_0$ with square distance $d^2[h]$ $= <h - \bar{t}|O|h - \bar{t}>$ from template average $\bar{t} = O^{-1}\tilde{t}$ where $\tilde{t} = \sum_i^N O_i t_i$, $O = \sum_i O_i$ and h–independent minimal component energy $E_0 = \sum_i^N <t_i|O_i|t_i> - <\bar{t}|O|\bar{t}>$, which has, up to a factor N, the structure of a variance. The linear stationarity equation for a functional $E = d^2[h]/2 + E_0$ reads $0 = O(\bar{t} - h) = \tilde{t} - Oh$.*

The inverse exists for positive definite O giving the solution $h = \bar{t} = O^{-1}\tilde{t}$. This case includes for example the classical regularization approaches of Radial Basis Functions (RBF) and various spline methods [3, 4, 5, 6].

As example for OR–like combinations. consider an old, only partly conserved image. Parts of it might be trees OR persons. Having available a collection of example images of trees and persons one may wish to use this information to improve the reconstruction of the image. One has to be careful not to reconstruct a tree as person. Also we do not want to simply replace objects by their prototypes, but to reconstruct the original image as good as possible.

A probabilistic OR for disjunct events appears for probabilities as sum, i.e. as mixture model, $p(h) = \sum_i^{N_i} p(h,i) = \sum_i^{N_i} p_i\, p(h|i)$. Assuming $p(h|i)$ to be easier to specify than $p(h,i)$ and using the notations $p(h|i) = e^{-\beta E_i}/Z_i$, $Z_i = \int dh\, e^{-\beta E_i}$, $p_i = e^{-b e_i}/z$, $z = \sum_i e^{-b e_i}$ and $E_i = (1/2)\sum_{j_i}^{n_{j_i}} d_{i j_i}^2$ we obtain

Proposition 2 (probabilistic OR) *Modeling a probabilistic OR for concepts by a mixture (or finite temperature) model*

$$E^M[h] = \ln\left(\sum_i^{N_i} e^{-\beta E_i[h] - b e_i - \ln Z_i(\beta)}\right), \tag{7}$$

the stationarity condition becomes $0 = t^M[h] - O^M[h]h$, with $O^M = \sum_i p(h,i) O_i$ and $t^M = \sum_i p(h,i) O_i \bar{t}_i = \sum_i p(h,i)\tilde{t}_i$. It is in general nonlinear.

Here we introduced a convexity parameter β (inverse temperature, Lagrange multiplier to determine the average error) controlling the number of local minima of E^M and an analogous b for mixture probabilities p_i. (See for example [7] and, similarly, for clustering [8].)

Instead of a mixture of Gaussian ("free") models one may specify non–concave ("interacting") log–probabilities by fuzzy logical combinations. Choosing the product as a simple realization of a fuzzy logical OR for distances we can consider polynomial interactions:

Proposition 3 ("a fuzzy OR") *A product implementation of OR gives the polynomial model*

$$E^F = \frac{1}{2} \sum_{i=1}^{n_i} \beta_F^{n_{j_i}-1} \prod_{j_i=1}^{n_{j_i}} d^2_{ij_i}, \tag{8}$$

with stationarity equation $O^F[h]h = t^F[h]$. *Here* $O^F[h] = \sum_i \sum_{j_i} M_{ij_i}[h] O^{ij_i}$ *and* $t^F[h] = \sum_i \sum_{j_i} M_{ij_i}[h] O^{ij_i} t^{ij_i}$ *with* $M_{ij_i}[h] = \beta_F^{n_{j_i}-1} \prod_{k_i \neq j_i} d^2_{ik_i}[h]$.

The error surface becomes convex in the limit $\beta_F \to 0$. This parallelizes the phenomenological Landau–Ginzburg treatment of phase transitions in statistical physics.

4 Iteration procedures = learning algorithms

The stationarity equations of both models are in general inhomogeneous integral equations. Similar equations appear for example in quantum mechanical scattering theory, where the inhomogeneities, in analogy to templates or data, represent the measurable asymptotic states ("channels") of the system [9]. As nonlinear equations they have to be solved by iteration. An iteration procedure ("learning algorithm") for solving a (linear or nonlinear) equation of the form $Oh = t$ is obtained by selecting a possibly iteration step k and h^k–dependent operator A and relaxation factor η and using the updating rule $h^{k+1} = h^k + \eta A^{-1}(t^k - O^k h^k)$. For any positive definite A the error decreases till reaching a local minimum provided η is chosen small enough. An A equal to the identity, for example, corresponds to the gradient algorithm and requires no inversion. A Gaussian A^{-1} can approximate local correlations induced by differential operators. Choosing $A = O^k$ corresponds for error functional E^M to the EM algorithm and selecting the negative Hessian results in the Newton method.

Figure (1) summarizes the temperature dependence which have been found in numerical studies of a model with error $E^M_{(2)} = e^{-\frac{\beta}{2}(d_D^2+d_1^2)} + e^{-\frac{\beta}{2}(d_D^2+d_2^2)}$. This corresponds to a data template d_D AND–ed to a probabilistic OR of two continuous prior templates d_i^2. Here the data template d_D^2 is a sum of standard mean square error terms and the prior concepts $d_i^2 = - < h - t_i | \Delta | h - t_i >$ measure the ("Laplace") distance between $h(x)$ and prior template $t_i(x)$. The actual data and the two continuous prior templates t_i are displayed on the right hand side of the figure.

Conclusion

Quadratic prior concepts have been introduced to provide a tractable link between prior knowledge of fuzzy form and regularization theory.

Acknowledgments

The author wants to thank Tomaso Poggio and Federico Girosi for useful discussions and the hospitality during the time he spent at the Center for Biological and Computational Learning, Massachusetts Institute of Technology. He was supported by a Postdoctoral Fellowship (Le 1014/1–1) from the Deutsche Forschungsgemeinschaft and a NSF/CISE Postdoctoral Fellowship.

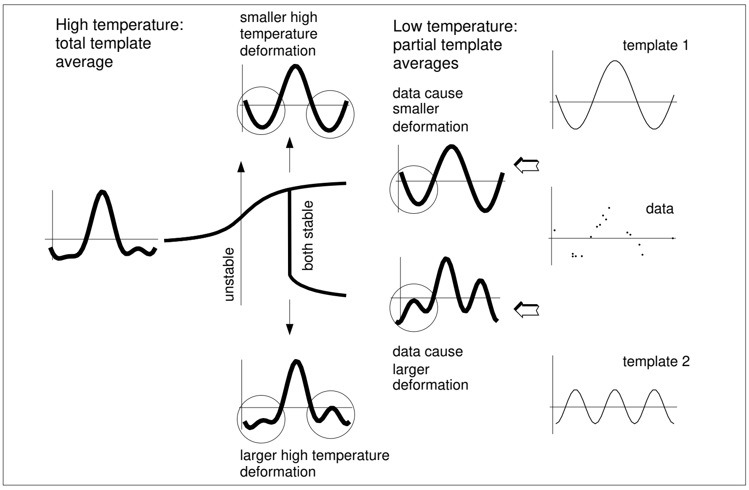

Figure 1: Scheme of temperature ($=\beta^{-1}$) dependence of minima h of error functional $E^M_{(2)}[h]$ in a model with data template (=training data) and two continuous prior templates combined by OR and Laplace operator as smoothness distance. For low temperatures two local minima appear corresponding to one of the two continuous prior templates deformed by the training data. At high temperature all templates are effectively AND–ed and the solution is the total template average of all three templates.

References

[1] Klir, G.J. & Yuan, B. (eds.): Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems. World Scientific, 1996.

[2] Itzykson, C.& Drouffe, J–M: Statistical field theory (vol 1 & vol 2). Cambridge University Press, 1989.

[3] Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035–1038, 1963.

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

[5] Poggio, T. & Girosi, F.: Networks for Approximation and Learning. Proceedings of the IEEE, Vol 78, No. 9, 1990.

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

[7] Yuille, A.L. & Kosowski, J.J.: Statistical Physics Algorithm That Converge. *Neural Computation* **6** (3), 341–356, 1994.

[8] Rose, K., Gurewitz, E., & Fox, G.C.: Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* **65**, 945–948, 1990.

[9] Lemm, J.C.: Inhomogeneous Random Phase Approximation for Nuclear and Atomic Reactions. *Annals of Physics* **244** (1), 136–200 & 201–238, 1995.