Regularized Learning with Flexible Constraints

Eyke Hüllermeier

Informatics Institute, Marburg University, Germany
eyke@informatik.uni-marburg.de

Abstract. By its very nature, inductive inference performed by machine learning methods is mainly data-driven. Still, the consideration of background knowledge – if available – can help to make inductive inference more efficient and to improve the quality of induced models. Fuzzy set-based modeling techniques provide a convenient tool for making expert knowledge accessible to computational methods. In this paper, we exploit such techniques within the context of the regularization (penalization) framework of inductive learning. The basic idea is to express knowledge about an underlying data-generating model in terms of flexible constraints and to penalize those models violating these constraints. Within this framework, an optimal model is one that achieves an optimal trade-off between fitting the data and satisfying the constraints.

1 Introduction

The common framework of learning from data assumes an underlying functional relation $f(\cdot)$ that is a mapping $\mathfrak{D}_X \to \mathfrak{D}_Y$ from an input space $\mathfrak{D}_X$ to an output space $\mathfrak{D}_Y$. One basic goal of learning from data is to approximate (or even recover) this function from given sample points $(x_i, y_i) \in \mathfrak{D}_X \times \mathfrak{D}_Y$. Depending on the type of function to be learned, several special performance tasks can be distinguished. For example, in *classification* (pattern recognition) $f(\cdot)$ is a mapping $\mathfrak{D}_X \to \mathcal{L}$ with $\mathcal{L}$ a finite set of labels (classes). If $f(\cdot)$ is a continuous function, the problem is one of *regression*. Here, the sample outputs are usually assumed to be noisy, e.g. corrupted with an additive error term: $y_i = f(x_i) + \varepsilon$.

On the basis of the given data, an approximation (estimation) $h_0(\cdot)$ of the function $f(\cdot)$ is chosen from a class $\mathcal{H}$ of candidate functions, called the *hypothesis space* in machine learning [9]. The adequate specification of the hypothesis space is crucial for successful learning, and the complexity of $\mathcal{H}$ is a point of special importance. In fact, if the class of candidate functions is not rich (flexible) enough, it is likely that $f(\cdot)$ cannot be approximated accurately: There might be no $h(\cdot) \in \mathcal{H}$ that is a good approximation of $f(\cdot)$. On the other hand, if $\mathcal{H}$ is too flexible, the learning problem might become ambiguous in the sense that there are several hypotheses that fit the data equally well. Apart from that, too much flexibility involves a considerable risk of *overfitting* the data. Roughly speaking, overfitting means reproducing the data in too exact a manner. This happens if the hypothesis space is flexible enough to provide a good or even exact approximation of any set of sample points. In such cases the induced hypothesis $h_0(\cdot)$

M.R. Berthold et al. (Eds.): IDA 2003, LNCS 2810, pp. 13–24, 2003.

is likely to represent not only the structure of the function $f(\cdot)$ but also the structure of the (noisy) data.

The specification of a hypothesis space allows for the incorporation of background knowledge into the learning process. For instance, if the relation between inputs and outputs is known to follow a special functional form, e.g. linear, one can restrict the hypothesis space to functions of this type. The calibration of such models through estimating free parameters is typical for classical statistical methods. However, if only little is known about the function $f(\cdot)$, restricting oneself to a special class of (simple) functions might be dangerous. In fact, it might then be better to apply *adaptive* methods which proceed from a very rich hypothesis space.

As pointed out above, for adaptive methods it is important to control the complexity of the induced model, that is, to find a good trade-off between the complexity and the accuracy of the hypothesis $h_0(\cdot)$. One way to accomplish this is to "penalize" models which are too complex. Thus, the idea is to quantify both the accuracy of a hypothesis as well as its complexity and to define the task of learning as one of minimizing an objective function which combines these two measures. This approach is also referred to as *regularization*.

Fuzzy set-based (linguistic) modeling techniques provide a convenient way of expressing background knowledge in learning from data [5]. In this paper, we propose fuzzy sets in the form of flexible constraints [4,3,7] as a means for regularization or, more specifically, for expressing background knowledge about a functional relation to be learned.

The rest of the paper is organized as follows: We briefly review the regularization framework in Section 2. The idea of expressing background knowledge about a model in terms of flexible constraints is detailed in Section 3. Section 4 is devoted to "constraint-based regularization", that is the application of flexible constraints in learning from data. We conclude the paper with some remarks in Section 5.

2 The Regularization Framework

Without loss of generality, we can assume $\mathcal{H}$ to be a parameterized class of functions, that is $\mathcal{H} = \{h(\cdot, \omega) \mid \omega \in \Omega\}$. Here, ω is a parameter (perhaps of very high or even infinite dimension) that uniquely identifies the function $h = h(\cdot, \omega)$, and $h(x, \omega)$ is the value of this function for the input x.

Given observed data in the form of a (finite) sample

$$S = \big\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \big\} \in (\mathfrak{D}_X \times \mathfrak{D}_Y)^n,$$

the selection of an (apparently) optimal hypothesis $h_0(\cdot)$ is guided by some *inductive principle*. An important and widely used inductive principle is *empirical risk minimization* (ERM) [11]. The risk of a hypothesis $h(\cdot, \omega)$ is defined as

$$R(\omega) \doteq \int_{\mathfrak{D}_X \times \mathfrak{D}_Y} L(y, h(x, \omega))\, p(x, y)\, d(x, y),$$

where $L(\cdot)$ is a loss function and $p(\cdot)$ is a probability (density) function: $p(x, y)$ is the probability (density) of observing the input vector x together with the output y. The empirical risk is an approximation of the true risk:

$$R_{emp}(\omega) \doteq \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i, \omega)) \ . \tag{1}$$

The ERM principle prescribes to choose the parameter ω_0 resp. the associated function $h_0 = h(\cdot, \omega_0)$ that minimizes (1).

Now, let $\phi(\omega) = \phi(h(\cdot, \omega))$ be a measure of the complexity of the function $h(\cdot, \omega)$. The *penalized risk* to be minimized for the *regularization inductive principle* is then given by a weighted sum of the empirical risk and a penalty term:

$$R_{pen}(\omega) \doteq R_{emp}(\omega) + \lambda \cdot \phi(\omega). \tag{2}$$

The parameter λ is called the regularization parameter. It controls the trade-off between accuracy and complexity.

As already indicated above, the function $h(\cdot, \omega_0)$ minimizing (2) is supposed to have a smaller *predictive risk* than the function minimizing the empirical risk. In other words, it is assumed to be a better generalization in the sense that it predicts future outputs more accurately.

Without going into detail let us mention that the practical application of the regularization framework requires, among other things, optimization methods for the minimization of (2) as well as techniques for specifying an optimal regularization parameter λ.

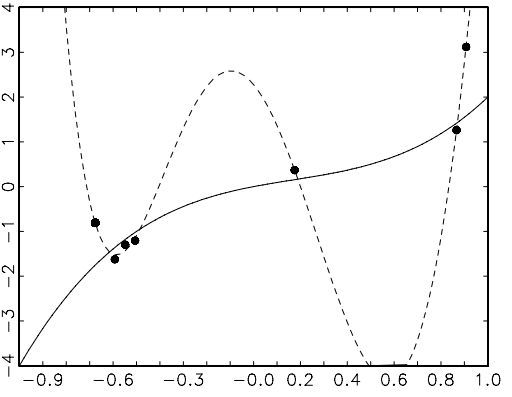

Fig. 1. The function $f(\cdot)$ to be learned (solid line) together with the given sample points and least squares approximation (dashed line).

To illustrate regularized learning, let us consider a very simple example. Suppose that the function to be learned is given by the polynomial $f : [0, 1] \to \mathfrak{R}$, $x \mapsto 2x^3 - x^2 + x$, and let the hypothesis space consist of all polynomials

$$x \mapsto \omega_0 + \sum_{k=1}^{m} \omega_k \cdot x^k$$

whose degree is at most $m = 5$. We have generated seven sample points at random, assuming a uniform distribution for x and a normal distribution with mean $f(x)$ and standard deviation $1/2$ for y. Fig. 1 shows the function $f(\cdot)$ together with the given data. Furthermore, the figure shows the optimal least squares approximation, i.e. the approximation $h_0(\cdot)$ that minimizes the empirical risk with loss function $L(y, h(x, \omega)) = \big(y - h(x, \omega)\big)^2$. As can be seen, $h_0(\cdot)$ does obviously overfit the data. In fact, the approximation of the true function $f(\cdot)$ is very poor.

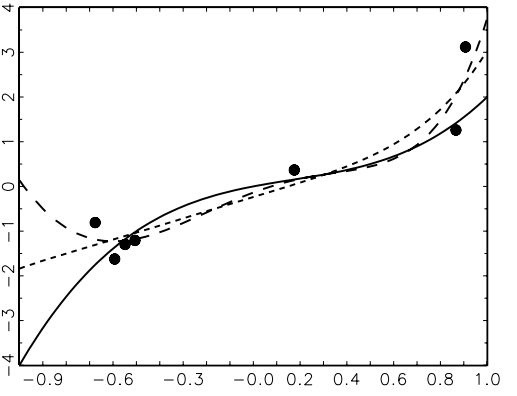

Fig. 2. The function $f(\cdot)$ to be learned (solid line), ridge regression with $\lambda = 0.01$ (dashed line) and $\lambda = 0.1$ (small dashes).

A special type of regularization is realized by so-called (standard) *ridge regression* [8] by using the sum of squared parameters $(\omega_k)^2$ as a penalty term. Thus, ridge regression seeks to minimize the penalized risk

$$R_{pen}(\omega) = \sum_{\imath=1}^{n} \big(y_\imath - h(x_\imath, \omega)\big)^2 + \lambda \cdot \sum_{k=1}^{m} (\omega_k)^2$$

that favors solutions with parameters being small in absolute value.[1] This type of regression problem can still be solved analytically. Fig. 2 shows the approximations for $\lambda = 0.01$ and $\lambda = 0.1$. The regularization effect becomes obvious in both cases. As can be seen, the larger the parameter λ is chosen, the smaller the variation of the approximating function will be.

[1] The intercept ω_0 is usually not considered in the penalty term.

3 Flexible Constraints

This section is meant to provide some background information on fuzzy sets and flexible constraints. Moreover, it will be shown how the framework of flexible constraints can be used for expressing properties of (real-valued) functions.

3.1 Background on Fuzzy Sets

A fuzzy subset of a set $\mathfrak{D}$ is identified by a so-called membership function, which is a generalization of the characteristic function $\mathbb{I}_A$ of an ordinary set $A \subseteq \mathfrak{D}$ [12]. For each element $x \in \mathfrak{D}$, this function specifies the degree of membership of x in the fuzzy set. Usually, membership degrees are taken from the unit interval $[0, 1]$, i.e. a membership function is a mapping $\mathfrak{D} \to [0, 1]$. We shall use the same notation for ordinary sets and fuzzy sets. Moreover, we shall not distinguish between a fuzzy set and its membership function, that is, $A(x)$ denotes the degree of membership of the element x in the fuzzy set A.

Is it reasonable to say that (in a certain context) 30°C is a high temperature and 29.9°C is not high? In fact, any sharp boundary of the set of high temperatures will appear rather arbitrary. Fuzzy sets formalize the idea of *graded membership* and, hence, allow for "non-sharp" boundaries. Modeling the concept "high temperature" as a fuzzy set A, it becomes possible to express, for example, that a temperature of 30°C is completely in accordance with this concept ($A(30) = 1$), 20°C is "more or less" high ($A(20) = 0.5$, say), and 10°C is clearly not high ($A(10) = 0$). As can be seen, fuzzy sets can provide a reasonable interpretation of linguistic expressions such as "high" or "low". This way, they act as an interface between a quantitative, numerical level and a qualitative level, where knowledge is expressed in terms of natural language.

Membership degrees of a fuzzy set can have different semantic interpretations [6]. For our purposes, the interpretation of fuzzy sets in terms of *preference* is most relevant. Having this semantics in mind, a (fuzzy) specification of the temperature as "high" means that, e.g., a temperature of 30°C is regarded as fully satisfactory, but that other temperatures might also be acceptable to some extent. For example, with the fuzzy set of high temperatures being specified as above, 20°C would be accepted to degree 0.5.

3.2 Flexible Constraints

In connection with the preference semantics, a fuzzy set defines a "flexible constraint" in the sense that an object can satisfy the constraint to a certain degree. Formally, let $C : \mathfrak{D}_X \to [0, 1]$ be a fuzzy set associated with a constraint on the variable X; $C(x)$ is the degree to which $x \in \mathfrak{D}_X$ satisfies the constraint.

Now, consider constraints $C_1, C_2, \ldots, C_n$ with a common domain $\mathfrak{D}_X$ and let C be the conjunction of these constraints. The degree to which an object $x \in \mathfrak{D}_X$ satisfies C can then be defined as

$$C(x) \doteq C_1(x) \otimes C_2(x) \otimes \ldots \otimes C_n(x), \tag{3}$$

where $\otimes$ is a t-norm. A t-norm is a generalized logical conjunction, that is, a binary operator $[0,1] \times [0,1] \to [0,1]$ which is associative, commutative, non-decreasing in both arguments and such that $\alpha \otimes 1 = \alpha$ for all $0 \leq \alpha \leq 1$. Well-known examples of t-norms include the minimum operator $(\alpha, \beta) \mapsto \min\{\alpha, \beta\}$ and the Lucasiewicz t-norm $(\alpha, \beta) \mapsto \max\{\alpha + \beta - 1, 0\}$. Set-theoretically, C as defined in (3) is the intersection $C_1 \cap C_2 \cap \ldots \cap C_n$ of the fuzzy sets $C_\imath$.

Note that a common domain $\mathfrak{D}_X$ can be assumed without loss of generality. In fact, suppose that C_1' specifies a constraint on a variable Y with domain $\mathfrak{D}_Y$ and that C_2' specifies a constraint on a variable Z with domain $\mathfrak{D}_Z$. The fuzzy sets C_1' and C_2' can then be replaced by their *cylindrical extensions* to the domain $\mathfrak{D}_X \doteq \mathfrak{D}_Y \times \mathfrak{D}_Z$ of the variable $X = (Y, Z)$, that is by the mappings

$$C_1 : (y, z) \mapsto C_1'(y), \qquad C_2 : (y, z) \mapsto C_2'(z).$$

The definition of a cylindrical extension is generalized to higher dimensions in a canonical way.

In connection with the specification of fuzzy constraints, it is useful to dispose of further logical operators:

- The common definition of the negation operator $\neg$ in fuzzy logic is the mapping given by $\neg x = 1 - x$ for all $0 \leq x \leq 1$ (though other negations do exist).
- A t-conorm $\oplus$ is a generalized logical disjunction, that is, a binary operator $[0,1] \times [0,1] \to [0,1]$ which is associative, commutative, non-decreasing in both arguments and such that $\alpha \oplus 0 = \alpha$ for all $0 \leq \alpha \leq 1$. Given a t-norm $\otimes$, an associated t-conorm $\oplus$ can be defined by the mapping $(\alpha, \beta) \mapsto 1 - (1 - \alpha) \otimes (1 - \beta)$ or, more generally, as $(\alpha, \beta) \mapsto \neg(\neg\alpha \otimes \neg\beta)$. For example, the t-conorm associated with the Lucasiewicz t-norm is given by $(\alpha, \beta) \mapsto \min\{\alpha + \beta, 1\}$.
- A (multiple-valued) implication operator $\rightsquigarrow$ is a mapping $[0,1] \times [0,1] \to [0,1]$ which is non-increasing in the first and non-decreasing in the second argument, i.e., $\alpha \rightsquigarrow \beta \leq \alpha' \rightsquigarrow \beta$ for $\alpha' \leq \alpha$ and $\alpha \rightsquigarrow \beta \leq \alpha \rightsquigarrow \beta'$ for $\beta \leq \beta'$. Besides, further properties can be required [2]. An example of an implication is the Lucasiewicz operator $(\alpha, \beta) \mapsto \max\{1 - \alpha + \beta, 0\}$. Implication operators can be used for modeling *fuzzy rules*: Let C_1 and C_2 be two constraints. A new constraint C can then be expressed in terms of a fuzzy rule

  $$\text{IF } C_1(x) \text{ THEN } C_2(x).$$

  The degree to which an object x satisfies this constraint can be evaluated by means of an implication $\rightsquigarrow$, that is $C(x) = C_1(x) \rightsquigarrow C_2(x)$.

3.3 Flexible Constraints on Functions

Flexible constraints of the above type can be used for expressing knowledge about a functional relation $f : \mathfrak{D}_X \to \mathfrak{D}_Y$. For the sake of simplicity, and without loss of generality, we shall assume that $\mathfrak{D}_Y \subseteq \mathfrak{R}$. If $f(\cdot)$ is a vector-valued function with domain $\mathfrak{D}_Y \subseteq \mathfrak{R}^m$, $m > 1$, then each of its components $f_\imath(\cdot)$ can be considered

separately. Subsequently, we shall introduce some basic types of constraints and illustrate them through simple examples.

The most obvious type of constraint is a restriction on the absolute values of the function. Knowledge of this type can be expressed in terms of a fuzzy rule such as [IF x is close to 0 THEN $f(x)$ is approximately 1] or, more formally, as

$$\forall\, x \in \mathfrak{D}_X \;:\; C_1(x) \rightsquigarrow C_2(f(x)),$$

where the fuzzy set C_1 models the constraint "close to 0" and C_2 models the constraint "approximately 1". The degree of satisfaction of this constraint is given by

$$C(f) \doteq \inf_{x \in \mathfrak{D}_X} C_1(x) \rightsquigarrow C_2(f(x)). \tag{4}$$

Note that the infimum operator in (4) generalizes the universal quantifier in classical logic.

Constraints of such type can simply be generalized to the case of $m > 1$ input variables:

$$C(f) \doteq \inf_{x_1,\ldots,x_m \in \mathfrak{D}_X} (\, C_1(x_1) \otimes \ldots \otimes C_m(x_m)\,) \;\rightsquigarrow\; C_{m+1}(f(x_1,\ldots,x_m)).$$

Note that constraints on the input variables need not necessarily be "non-interactive" as shown by the following example: [IF x_1 is close to x_2 THEN $f(x_1, x_2)$ is approximately 1]. This constraint can be modeled in terms of an implication whose antecedent consists of a constraint such as $C_1 : (x_1, x_2) \mapsto \max\{1 - |x_1 - x_2|, 0\}$.

In a similar way, constraints on first or higher (partial) derivatives of $f(\cdot)$ can be expressed. For instance, let $f'(\cdot)$ denote the (existing) derivative of $f(\cdot)$. Knowing that $f(\cdot)$ is increasing then corresponds to the (non-fuzzy) constraint

$$C(f) \doteq \inf_{x \in \mathfrak{D}_X} f'(x) \geq 0.$$

A flexible version of this constraint could be specified as

$$C(f) \doteq \inf_{x \in \mathfrak{D}_X} C_1(f'(x)),$$

where C_1 is the fuzzy set of derivatives strictly larger than 0, e.g. $C_1(dx) = \min\{1, dx\}$ for $dx \geq 0$ and $C_1(dx) = 0$ otherwise. This version takes into account that $f(\cdot)$ can be increasing to different degrees and actually expresses that $f(\cdot)$ is *strictly* increasing. Of course, a constraint on the derivative of $f(\cdot)$ can also be local in the sense that it is restricted to a certain (fuzzy) subset of $\mathfrak{D}_X$. Thus, a constraint such as "For x close to 0, the derivative of $f(\cdot)$ is close to 1" could be modeled as follows:

$$C(f) \doteq \inf_{x \in \mathfrak{D}_X} C_1(x) \rightsquigarrow C_2(f'(x)),$$

where the fuzzy sets C_1 and C_2 formalize, respectively, the constraints "close to 0" and "close to 1".

Other types of constraints include restrictions on the relative values of the function. For example, the constraint "The values of $f(\cdot)$ for x close to 0 are much smaller than those for x close to 1" can be modeled as follows:

$$C(f) \doteq \inf_{x_1, x_2 \in \mathfrak{D}_X} \big(C_1(x_1) \otimes C_2(x_2)\big) \rightsquigarrow C_3(f(x_1), f(x_2)),$$

where the fuzzy relation $C_3 \subseteq \mathfrak{D}_Y \times \mathfrak{D}_Y$ might be defined, e.g., by $C_3(y_1, y_2) = \min\{1, y_2 - y_1\}$ for $y_1 \leq y_2$ and $C_3(y_1, y_2) = 0$ otherwise.

4 Regularization with Flexible Constraints

The regularization framework of Section 2 and the framework of flexible constraints as presented in Section 3 can be combined in order to express background knowledge about a function in learning from data. The basic idea of "constraint-based regularization" is to replace the term $\phi(\omega)$ in the penalized risk functional (2) by a fuzzy constraint $C(\omega) = C(h(\cdot, \omega))$. Thus, one arrives at the following functional:

$$R_{pen}(\omega) \doteq R_{emp}(\omega) - \lambda \cdot C(\omega). \tag{5}$$

As can be seen, an evaluation $R_{pen}(\omega)$ is a trade-off between the accuracy of the hypothesis $h(\cdot, \omega)$, expressed by $R_{emp}(\omega)$, and the extent to which $h(\cdot, \omega)$ is in accordance with the background knowledge, expressed by $C(\omega)$.

4.1 Example

To illustrate, let us return to our small example presented in Section 2. Again, suppose that the function to be learned is given by the polynomial $f : x \mapsto 2x^3 - x^2 + x$ and let the hypothesis space consist of all polynomials whose degree is at most five. Moreover, the same sample points as before shall be given (cf. Fig. 1). Finally, suppose the prior knowledge about the function $f(\cdot)$ to be reflected by the following flexible constraints:

- If x is almost 0, then $f(x)$ is also close to 0.
- $f'(x)$ is approximately 3 for x around 0.8.
- $f'(x)$ is roughly between 6 and 7 for x close to -0.8.

The fuzzy sets used to model these constraints (i.e. the linguistic terms "almost", "close to", ...) are specified in terms of trapezoidal membership functions (we omit details due to reasons of space). Moreover, as generalized logical operators we employed the Lucasiewicz t-norm, t-conorm, and implication. Fig. 3 shows the function $h(\cdot, \omega_0)$ minimizing the penalized risk functional (5) for $\lambda = 1$. The optimal parameter $\omega_0 = (0.08, 1.17, -0.17, 2.62, -0.53, -0.51)$ has been found by means of a general stochastic search algorithm. As can be seen, the induced function $h(\cdot, \omega_0)$ is already quite close to the true function $f(\cdot)$.

It is worth mentioning that constraint-based regularization naturally supports a kind of "interactive" learning and optimization process. Namely, the

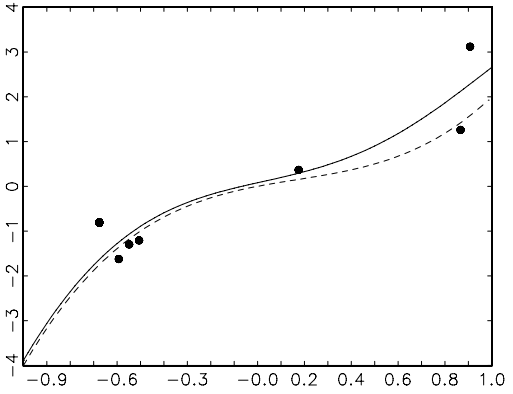

Fig. 3. Optimal approximation (solid line) of the underlying function (dashed line) for $\lambda = 1$.

optimal approximation obtained for a certain set of constraints might give the human expert cause to modify these constraints or to adapt the regularization parameter. That is, the inspection of the apparently optimal function might suggest weakening or strengthening some of the original constraints, or adding further constraints not made explicit so far. This way, learning a function from data is realized in the form of an iterative process.

4.2 Search and Optimization

As already pointed out, the minimization of the penalized risk functional (5) will generally call for numerical optimization techniques, such as stochastic search methods. The search process can often be made more efficient by exploiting the fact that parameters ω with $C(\omega) = 0$ can usually be left out of consideration.

Let ω_1 denote the optimal solution to the unconstrained learning problem, that is, the parameter vector minimizing the empirical risk $R_{emp}(\cdot)$. This vector can often be found by means of analytic methods, at least for common types of loss functions. Moreover, let ω_2 minimize the penalized risk (5) over the subspace

$$\Omega_0 \doteq \{\, \omega \in \Omega \mid C(\omega) > 0 \,\}.$$

For the overall optimal solution ω_0, we obviously have

$$\omega_0 = \begin{cases} \omega_1 & \text{if} \quad \omega_1 \leq \omega_2 \\ \omega_2 & \text{if} \quad \omega_2 < \omega_1 \end{cases}.$$

Consequently, the learning problem can be decomposed into two steps (followed by a simple comparison):

- Find the solution ω_1 to the unconstrained learning task.
- Find the optimal solution ω_2 to the constrained problem in Ω_0.

Essentially, this means that the search process can focus on the subspace $\Omega_0 \subseteq \Omega$, which will often be much smaller than Ω itself. In order to construct Ω_0, or an outer approximation thereof, the constraints on the function $f(\cdot)$ must be translated into constraints on the parameter ω. This translation clearly depends on which type of constraints and parameterized functions are used.

To illustrate the basic principle, consider an absolute value constraint C of the form

$$\text{IF} \quad x \in A \quad \text{THEN} \quad f(x) \in B, \tag{6}$$

where A and B are fuzzy subsets of $\mathfrak{D}_X$ and $\mathfrak{D}_Y$, respectively. Since $1 \rightsquigarrow 0 = 0$ for any implication operator $\rightsquigarrow$ satisfying the neutrality property, we have $C(\omega) = 0$ as soon as $A(x) = 1$ and $B(h(x,\omega)) = 0$, i.e. as soon as

$$\begin{aligned} x \in cr(A) &\doteq \{x \in \mathfrak{D}_X \mid A(x) = 1\}, \\ h(x,\omega) \notin sp(B) &\doteq cl\{x \in \mathfrak{D}_Y \mid B(y) > 0\}. \end{aligned}$$

Here, cr and sp denote, respectively, the so-called core and support of a fuzzy set, and clX is the closure of the set X. It follows that

$$\forall\, x \in cr(A) \,:\, h(x,\omega) \in sp(B) \tag{7}$$

is a necessary (though not sufficient) condition for $\omega \in \Omega_0$.

Now, suppose the functions $h(\cdot,\omega)$ to be expressed in terms of basis functions $\beta_\imath(\cdot)$, that is

$$h(x,\omega) = \sum_{\imath=0}^{m} \alpha_\imath \cdot \beta_\imath(x).$$

Moreover, let $sp(B) = [l,u]$ and

$$l_\imath \doteq \inf_{x \in cr(A)} \beta_\imath(x), \qquad u_\imath \doteq \sup_{x \in cr(A)} \beta_\imath(x)$$

for $0 \leq \imath \leq m$. The following conditions are necessary for the parameters $\alpha_\imath$ to satisfy (7):

$$l \leq \sum_{\imath=0}^{m} l_\imath \alpha_\imath, \qquad u \geq \sum_{\imath=0}^{m} u_\imath \alpha_\imath. \tag{8}$$

Thus, the original absolute value constraint (6) can be translated into constraints on the parameter ω, which are given in the form of linear inequalities.

In our example, we have $cr(T) = [\beta,\gamma]$ and $sp(T) = [\alpha,\delta]$ for a trapezoidal fuzzy set

$$T[\alpha,\beta,\gamma,\delta] : \; x \;\mapsto \begin{cases} (x-\alpha)/(\beta-\alpha) & \text{if} \quad \alpha \leq x < \beta \\ 1 & \text{if} \quad \beta \leq x \leq \gamma \\ (\delta - x)/(\delta - \gamma) & \text{if} \quad \gamma < x \leq \delta \\ 0 & \text{if} \quad x \notin [\alpha,\delta] \end{cases}.$$

Moreover, the basis functions are of the form $\beta_\imath(x) = x^\imath$, hence $l_\imath$ and $u_\imath$ are attained at 0 or at the boundary points of the interval $[\beta, \gamma]$.

Finally, note that $0 \otimes t = 0$ for any t-norm $\otimes$ and $0 \leq t \leq 1$. Therefore, conditions on the parameter ω derived from different constraints can be combined in a conjunctive way. Thus, if $\Omega_{0,\imath}^{appr}$ denotes an outer approximation of the set $\Omega_{0,\imath} \doteq \{\omega \mid C_\imath(\omega) > 0\}$, then

$$\Omega_0^{appr} \doteq \bigcap \Omega_{0,\imath}^{appr} \supseteq \Omega_0$$

is an outer approximation of Ω_0. Particularly, this means that linear inequalities (8) stemming from different constraints can simply be lumped together.

4.3 Relation to Bayesian Inference

It is illuminating to compare constraint-based regularization with other types of inductive principles. In this respect, Bayesian inference appears to be particularly interesting. In fact, just like constraint-based regularization, Bayesian inference [1] is concerned with the incorporation of background knowledge in learning from data.[2] (As opposed to this, other frameworks such as *structural risk minimization* [11] or *minimum description length* [10] are more data-driven.) In fact, it can be shown that constraining the class of models as detailed above can be interpreted as specifying a prior probability distribution for Bayesian inference. (Again, we refrain from a detailed discussion due to reasons of space.) This is quite interesting, since the acquisition of a reasonable prior in the Bayesian approach is generally considered a quite difficult problem. Especially, it should be noted that a prior distribution is generally given as a global (parameterized) function on Ω. The specification of such a function is hence difficult for high-dimensional input spaces. For instance, in our example above, one has to define a density over the space of polynomials of degree 5, i.e. over a six-dimensional parameter space. Apart from that, the relation between a parameter vector, such as $(\omega_0, \omega_1, \ldots, \omega_5)$ in our example, and the induced function is often not obvious. In fact, in order to define a distribution over Ω, one first has to "translate" known properties of a function into properties of the parameters. For instance, what does it mean for the prior distribution of $(\omega_0, \omega_1, \ldots, \omega_5)$ that "the derivative of $f(\cdot)$ is small for x close to 1"? In this respect, our constraint-based approach is more flexible and by far less demanding as it allows one to specify constraints that are local and that refer to aspects of the function $f(\cdot)$ directly.

5 Concluding Remarks

We have introduced (fuzzy) constraint-based regularization where the basic idea is to embed fuzzy modeling in regularized learning. This approach provides a simple yet elegant means for considering background knowledge in learning from

[2] The same is true for inductive logic programming and knowledge-based neurocomputing.

data. Using fuzzy set-based (linguistic) modeling techniques, such knowledge can be expressed in terms of flexible constraints on the model to be learned.

Due to limited space, the paper could hardly go beyond presenting the basic idea and its "ingredients" (regularized learning and fuzzy modeling), and the method obviously needs further elaboration. An important aspect of ongoing work concerns its practical realization, namely the development of suitable modeling tools and efficient optimization methods. A first prototype offering a restricted language for expressing constraints already exists. Such restrictions, concerning the type of fuzzy sets and logical operators that can be used as well as the type of constraints that can be specified, are reasonable not only from a modeling point of view, they also enable the development of specialized and hence efficient optimization methods.

There are also several theoretical questions which are worth further investigation. For instance, one such question concerns the sensitivity of constraint-based regularization, that is the dependence of the induced function $h(\cdot, \omega_0)$ on the specification of fuzzy sets and the choice of logical operators.

References

1. J. Bernardo and A. Smith. *Bayesian Theory.* J. Wiley & Sons, Chichester, 1994.
2. B. Bouchon-Meunier, D. Dubois, L. Godo, and H. Prade. Fuzzy sets and possibility theory in approximate reasoning and plausible reasoning. In J.C. Bezdek, D. Dubois, and H. Prade, editors, *Fuzzy Sets in Approximate Reasoning and Information Systems*, pages 15–190. Kluwer, 1999.
3. D. Dubois, H. Fargier, and H. Prade. The calculus of fuzzy restrictions as a basis for flexible constraint satisfaction. In *Second IEEE International Conference on Fuzzy Systems*, pages 1131–1136, 1993.
4. D. Dubois, H. Fargier, and H. Prade. Propagation and satisfaction of flexible constraints. In R.R. Yager and L. Zadeh, editors, *Fuzzy Sets, Neural Networks, and Soft Computing*, pages 166–187. Van Nostrand Reinhold, 1994.
5. D. Dubois, E. Hüllermeier, and H. Prade. Fuzzy set-based methods in instance-based reasoning. IEEE *Transactions on Fuzzy Systems*, 10(3):322–332, 2002.
6. D. Dubois and H. Prade. The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90(2):141–150, 1997.
7. H.W. Guesgen. A formal framework for weak constraint satisfaction based on fuzzy sets. Technical Report TR-94-026, ICSI Berkeley, June 1994.
8. A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.
9. T.M. Mitchell. *Machine Learning.* McGraw-Hill, Boston, Massachusetts, 1997.
10. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
11. V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, second edition, 2000.
12. L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.