Machine Learning and Statistical MAP Methods

Mark Kon[1], Leszek Plaskota[2], and Andrzej Przybyszewski[3]

[1] Boston University, Boston, MA 02215, USA
[2] Warsaw University, 02-097 Warsaw, Poland
[3] McGill University, Montreal, Quebec H3A 1B1, Canada

Abstract. For machine learning of an input-output function f from examples, we show it is possible to define an a priori probability density function on the hypothesis space to represent knowledge of the probability distribution of f, even when the hypothesis space H is large (i.e., nonparametric). This allows extension of maximum a posteriori (MAP) estimation methods nonparametric function estimation. Among other things, the resulting MAPN (MAP for nonparametric machine learning) procedure easily reproduces spline and radial basis function solutions of learning problems.

1 Introduction

In machine learning there are a number of approaches to solving the so-called function approximation problem, i.e., learning an input-output function $f(\mathbf{x})$ from partial information (examples) $y_i = f(\mathbf{x}_i)$ (see [6,9]). This is also the regression problem in statistical learning [12,8]. The problem has evolved from a statistical one dealing with low dimensional parametric function estimation (e.g., polynomial regression) to one which tries to extrapolate from large bodies of data an unknown element f in a nonparametric (large or infinite dimensional) hypothesis space H of functions. Recent nonparametric approaches have been based on regularization methods [12], information-based algorithms [9,10], neural network-based solutions [6], Bayesian methods [13], data mining [2], optimal recovery [5], and tree-based methods [3].

We will include some definitions along with a basic example. Suppose we are developing a laboratory process which produces a pharmaceutical whose quality (as measured by the concentration y of the compound being produced) depends strongly on a number of input parameters, including ambient humidity x_1, temperature x_2, and proportions $x_3, \ldots, x_n$ of chemical input components. We wish to build a machine which takes the above input variables $\mathbf{x} = (x_1, \ldots, x_n)$ and whose output predicts the desired concentration y. The machine will use experimental data points $y = f(\mathbf{x})$ to learn from previous runs of the equipment. We may already have a prior model for f based on simple assumptions on the relationships of the variables.

With an unknown i-o function $f(x)$, and examples $Nf \equiv (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ $= (y_1, \ldots, y_n) = \mathbf{y}$, we seek an algorithm ϕ which maps information Nf into the best estimate $\phi(Nf)$ of f. The new algorithm presented here (MAP for nonparametric machine learning, or MAPN) is an extension of methods common in parametric (finite dimensional) learning. In the approach, an a priori distribution P (representing prior knowledge) on the hypothesis space H of functions is given, and the function is learned by combining data Nf with a priori information μ.

One possible a posteriori estimate based on Nf is the conditional expectation $E(\mu|Nf)$ [7,10,9], which can be done in high (nonparametric) and low (parametric) dimensional situations. In low dimensions an easier estimation procedure is often done using maximum a posteriori (MAP) methods, in which a density function $\rho(x)$ of the probability measure P is maximized. In data mining on the other hand, a full (nonparametric) f must be estimated, and its infinite dimensional hypothesis space H does not immediately admit MAP techniques. We show that in fact densities $\rho(f)$ exist and make sense even for nonparametric problems, and that they can be used in the same way as in parametric machine learning. Given information $\mathbf{y} = Nf$ about an unknown $f \in H$, the MAPN estimate is simply $\widehat{f} = \arg\max_{f \in N^{-1}\mathbf{y}} \rho(f)$. Density functions $\rho(f)$ have some important advantages, including ease of use, ease of maximization, and ease of conditioning when combined with examples $(y_1, \ldots, y_n) = Nf$ (see examples in Section 3). Since they are also likelihood functions (representing our intuition of how "likely" a given guess f_1 is as compared to another f_2), they can be modified on a very intuitive basis (see also, e.g., [1]). For example, if we feel that we want our a priori guess at the unknown f to be smoother, we can weight the density function $\rho(f)$ (for the measure μ) with an extra factor $e^{-||Af||^2}$, with A a differential operator, in order to give less weight to "nonsmooth" functions with high values of $||Af||$. By the Radon-Nikodym theorem we will be guaranteed that the new (intuitively motivated) density $\rho(f)e^{-||Af||^2}$ will be the density of a bona fide measure ν, with $d\nu = e^{-||Af||^2} d\mu$.

2 The maximization algorithm

Let P be a probability distribution representing prior knowledge about $f \in H$, with the hypothesis space H initially finite dimensional. Let λ be uniform (Lebesgue) measure on H, and define the probability density function (pdf) of P (assuming it exists) by

$$\rho(f) = \frac{dP}{d\lambda}. \tag{1}$$

It is possible to define ρ alternatively up to a multiplicative constant through

$$\frac{\rho(f)}{\rho(g)} = \lim_{\epsilon \to 0} \frac{P(B_\epsilon(f))}{P(B_\epsilon(g))}. \tag{2}$$

That is the ratio of densities of two measures at f equals the ratio of the measures of two small balls there. Here $B_\epsilon(f)$ is the set of $h \in H$ which are within distance ϵ from f. Though definition (1) fails to extend to (infinite dimensional) function spaces H, definition (2) does. Henceforth it will be understood that a density function $\rho(f)$ is defined only up to a multiplicative constant (note (2) only defines ρ up to a constant). The MAP algorithm ϕ maximizes $\rho(f)$ subject to the examples $\mathbf{y} = Nf$. Thus (2) extends the notion of a density function $\rho(f)$ to a nonparametric H. Therefore it defines a likelihood function to be maximized a posteriori subject to $\mathbf{y} = Nf$. It follows from the theorem below that this in fact can be done for a common family of a priori measures [10]. For brevity, the proof of the following theorem is omitted.

Theorem 1. *If μ is a Gaussian measure on the function space H with covariance C, then the density $\rho(f)$ as defined above exists and is unique (up to a multiplicative constant), and is given by $\rho(f) = e^{-\langle f, Af \rangle}$, where $A = C^{-1/2}$.*

Under the assumption of no or negligible error (we will later not restrict to this), the MAPN estimate of f given data $Nf = \mathbf{y}$ is $\phi(Nf) = \widehat{f} = \arg\max_{Nf=y} \rho(f)$. More generally, these ideas extend to non-Gaussian probability measures as well; the theorems are omitted for brevity.

3 Applications

We consider an example involving a financial application of the MAPN procedure for incorporating a priori information with data. We assume that a collection of 30 credit information parameters are collected from an individual borrower's credit report by a large bank. These include total debts, total credit, total mortgage balances, and other continuous information determined earlier to be relevant by a data mining program. We wish to map this information into a best estimated debt to equity ratio two years hence. A (limited) database of past information is available, containing recent information (as of the last year) on debt to equity ratios, together with data on the $d = 30$ parameters of interest We wish to combine this information with an earlier estimate (taken 4 years earlier), consisting of a function $f_0 : J^{30} \to I$ from the (normalized) credit parameters into a debt to equity ratio (also normalized), where $J = [-1, 1]$ and $I = [0, 1]$. In order to avoid boundary issues, we will extend f_0 smoothly to a periodic map $K^{30} \to I$, where $K = [-1.5, 1.5]$, with -1.5 identified with 1.5, so that smooth functions on K must match (as well as all their derivatives) at the endpoints ± 1.5. Similarly, a function on the torus K^{30} is smooth if it is periodic and smooth everywhere, including on the matching periodic boundaries. The purpose of this is to expand a differentiable function f on K^{30} in a Fourier series.

On the belief that the current form $f \colon K^{30} \to I$ of the desired function is different from the (a priori) form f_0 earlier estimated, we make the prior assumption that there is a probability distribution P for the sought (currently true) f_1 centered at the earlier estimate f_0, having the form of a Gaussian on H, the set of square integrable functions from K^{30} to I. This a priori measure P favors deviations from f_0 which are sufficiently smooth to be well-defined pointwise (but not too smooth) and small, and so P is given the form of a Gaussian measure with a covariance C defined on the orthonormal basis (here a is a normalization constant) $\{b_{\mathbf{k}} = ae^{\frac{2}{3}\pi i\mathbf{x}\cdot\mathbf{k}}\}_{\mathbf{k}\in\mathbf{Z}^{30}}$ ($\mathbf{Z}$ is the integers) for $L^2(K^{30})$ by $C(e^{\frac{2}{3}\pi i\mathbf{x}\cdot\mathbf{k}}) = \frac{1}{(1+|\mathbf{k}|)^{31}}e^{\frac{2}{3}\pi i\mathbf{x}\cdot\mathbf{k}}$ with $\mathbf{k} = (k_1, \ldots, k_{30})$ a multiinteger, and $\mathbf{x} \in K^{30}$ (note that P forms a Gaussian measure essentially concentrated on functions $f \in L^2(K^{30})$ with 15.5 square integrable derivatives, which guarantees that such functions' pointwise values are well-defined, since $15.5 > \frac{d}{2}$). We uniquely define the operator A by $C = A^{-2}$; A satisfies $A(e^{\frac{2}{3}\pi i\mathbf{x}\cdot\mathbf{k}}) = |\mathbf{k}|^{31/2}e^{\frac{2}{3}\pi i\mathbf{x}\cdot\mathbf{k}}$. To simplify notation and work with a Gaussian centered at 0, we denote the full new i-o function we are seeking by $f_1(\mathbf{x})$. We will seek to estimate the change in the i-o function, i.e., $f = f_1 - f_0$. With this subtraction the function f we seek is centered at 0 and has a Gaussian distribution with covariance C. Our new i-o data are $y_i = f(\mathbf{x}_i) = f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)$, where $f_1(\mathbf{x}_i)$ are the measured debt to equity ratios, and are immediately normalized by subtracting the known $f_0(\mathbf{x}_i)$. Thus y_i sample the change $f(\mathbf{x}_i)$ in the i-o function.

We first illustrate the algorithm under the hypothesis that data $y_i = f(\mathbf{x}_i)$ are exact (the more realistic noisy case is handled below). In this exact information case the MAPN algorithm finds the maximizer of the density $\rho(f) = e^{-\|Af\|^2}$ (according to Theorem 1) restricted to the affine subspace $N^{-1}(\mathbf{y})$. This is equivalent to minimizing $\|Af\|$ subject to the constraint $\mathbf{y} = Nf = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, (where $f(\mathbf{x}_i)$ is the outcome for example $\mathbf{x}_i$), which yields the spline estimate

$$\widehat{f} = \sum_{j=1}^{n} c_j C L_j, \tag{3}$$

where for each j, the linear functional $L_j(f) = f(\mathbf{x}_j)$, and where $c_i = S\mathbf{y}$ is determined from $\mathbf{y}$ by a linear transformation S (see [9] for the construction of such spline solutions). We have (here δ denotes the Dirac delta distribution) $CL_j = C\delta(\mathbf{x} - \mathbf{x}_j) = C\left(a^2 \sum_k e^{\frac{2}{3}\pi i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}_j)}\right) = \sum_k a^2 C e^{\frac{2}{3}\pi i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}_j)}$ $= \sum_k \frac{a^2}{|\mathbf{k}|^{31}} e^{\frac{2}{3}\pi i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}_j)} = G(\mathbf{x} - \mathbf{x}_j)$ is a radial basis function (equivalently, a B-spline) centered at $\mathbf{x}_j$. So the estimated regression function is $\widehat{f} = \sum_{j=1}^{n} c_j G_j(\mathbf{x} - \mathbf{x}_j) = \sum_{j=1}^{n} c_j \sum_{\mathbf{k}} \frac{a^2}{|\mathbf{k}|^{31}} e^{\frac{2}{3}\pi i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}_j)}$. By comparison, a standard algorithm for forming a (Bayesian) estimate for f under the average case setting of information-based complexity theory using information $Nf = (y_1, \ldots, y_n)$ is to compute the conditional expectation $\phi(Nf) = E_\mu(f|N(f) = (y_1, \ldots, y_n))$. For a Gaussian measure this expectation is known also to yield the well-known spline estimate (3) for f [9,10]. The regularization algorithm [12] can be chosen to minimize the norm $\|Af\|$ subject to $Nf = \mathbf{y}$, again yielding the spline solution (3).

Noisy information: It is much more realistic, however, to assume the information $Nf = (y_1, \ldots, y_n)$ in the above example is noisy, i.e., that if $f = f_1 - f_0$ is the sought change in the 2 year debt to equity ratio, then $y_i = f(\mathbf{x}_i) + \epsilon_i$ where ϵ_i is a normally distributed error term. In this case the MAP estimator is given by $\widehat{f} = \arg\sup_f \rho(f|\mathbf{y})$. However, note that (as always, up to multiplicative constants) $\rho(f|\mathbf{y}) = \frac{\rho_{\mathbf{y}}(\mathbf{y}|f)\rho(f)}{\rho_{\mathbf{y}}(\mathbf{y})}$ so that if the pdf of $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is Gaussian, i.e., has density $\rho_\epsilon(\epsilon) = K_1 e^{-\|B\epsilon\|^2}$ with B linear and K a constant, then $\rho(f|\mathbf{y}) = K_2 \frac{e^{-\|B(Nf-\mathbf{y})\|^2} e^{-\|Af\|^2}}{\rho_{\mathbf{y}}(\mathbf{y})} = K_3 e^{-\|B(Nf-\mathbf{y})\|^2 - \|Af\|^2}$ where K_3 can depend on the data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. MAP requires that this be maximized, so

$$\widehat{f} = \arg\min \|Af\|^2 + \|B(Nf - \mathbf{y})\|^2. \tag{4}$$

This maximization can be done using Lagrange multipliers, for example. This again is a spline solution for the problem with error [7]. In addition, again, the minimization of (4) is the same as the regularization functional minimization approach in statistical learning theory [12]. It yields a modified spline solution as in (3), with modified coefficients c_j.

References

1. N. Friedman and J. Halpern (1996) Plausibility measures and default reasoning. In Thirteenth National Conf. on Artificial Intelligence (AAAI).
2. T. Hastie, R. Tibshirani and J. Friedman (2001) The elements of statistical learning: data mining, inference and prediction. Springer-Verlag.
3. M. Jordan. and M. Meila (2000) Learning with mixtures of trees. Journal of Machine Learning Research **1**, 1-48.
4. M. Kon (2004) Density functions for machine learning and optimal recovery. preprint.
5. C. Micchelli and T. Rivlin (1985) Lectures on optimal recovery. Lecture Notes in Mathematics, 1129, Springer-Verlag, Berlin, 1985, 21-93.
6. T. Mitchell (1997) Machine Learning. McGraw-Hill, NY.
7. L. Plaskota (1996) Noisy Information and Complexity. Cambridge Univ. Press.
8. T. Poggio and C. Shelton (1999) Machine Learning, Machine Vision, and the Brain. The AI Magazine **20**, 37-55.
9. J. Traub, G. Wasilkowski, and H. Wozniakowski (1988), Information-Based Complexity. Academic Press, Boston.
10. J. F. Traub and A. Werschulz (2001) Complexity and Information. Cambridge University Press, Cambridge.
11. J. Traub and H. Wozniakowski (1980) A General Theory of Optimal Algorithms. Academic Press, New York.
12. V. Vapnik (1998) Statistical Learning Theory. Wiley, New York.
13. G. Wahba (1999) Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. in B. Schoelkopf, C. Burges & A. Smola, eds, Advances in Kernel Methods Support Vector Learning, MIT Press, 69-88.