9th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-9-2008) Nizhni Novgorod, Russian Federation, September, 14-20, 2008. Conference Proceedings, Volume 2

ESTIMATION OF HYPERPARAMETERS ON PARAMETRIC REGRESSION MODEL GENERATION[1]

V.V. Strijov[2]

[2] Computing Center of the Russian Academy of Sciences, Vavilova 40, Moscow, Russia, strijov@ccas.ru

The problem of the non-linear regression analysis is considered. The algorithm of the inductive model generation is described. The regression model is a superposition of given smooth functions. To estimate the model parameters two-level Bayesian Inference technique was used. It introduces hyperparameters, which describe the distribution function of the model parameters.

Introduction

Inductive model generation algorithms invoke the problem of models elements importance estimation. C. Bishop suggested a method [1] of evaluation the probability distribution function for the model parameters. The parameters of these functions are called hyperparameters. For each element of the model one must to estimate the probability distribution function and make a decision either particular element of the regression model important or not.

The problem of the model comparison using hyperparameters was advanced after papers by D. MacKay and I. Nabney. The papers [3–6] and [7] investigate hyperparameter optimization algorithms.

In this paper an inductive model generation algorithm is described. It consists of the following steps. Data set, namely the values of several independent variables and one dependent variable are given. The set of terminal functions and optionally the set of initial models are given. The model parameters and hyperparameters are tuned with an optimization algorithm. For each model, the importance of superposition elements is evaluated. The importance depends on the values of the hyperparameters. Several best generated models are selected according to a target function. The selected models are modified and new models are generated according to generation rules.

The hyperparameter values bring the information how to modify the models to improve them.

Problem statement

A sample set $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, of the independent variables $\mathbf{x} \in \mathbf{R}^P$ and corresponding depended variables $y \in \mathbf{R}$ is given. A set $G = \{g \mid g : \mathbf{R} \times \ldots \times \mathbf{R} \to \mathbf{R}\}$ of the smooth parametric functions $g = g(\mathbf{b}, \cdot, \ldots, \cdot)$ is given.

The set G inductively defines a set of superposition $F = \{f_i\}$. Each superposition f_i consists of no more than r functions g.

The superposition f defines a parametric regression model $f = f(\mathbf{w}, \mathbf{x})$. The vector $\mathbf{w} \in \mathbf{R}^W$ consists on concatenated parameter vectors of the functions $g_1, \ldots, g_r$, thus $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots \ldots \vdots \mathbf{b}_r$, where $\vdots$ is the vector concatenation.

One must to select from the set F a model f_i, which minimizes the given target function $p(\mathbf{w} \mid D, \alpha, \beta, f_i)$. This function depends on the sample set D, the model $f_i = f_i(\mathbf{w}, \mathbf{x})$ and additional parameters α, β. The shape of the target function is defined by hypotheses on the sample set distribution.

[1] Support of the grant RFBR 07-07-00181

The target function is defined as following. Let ν be a random variable of the regression problem $y = f_i(\mathbf{w}, \mathbf{x}) + \nu$. It has a Gaussian distribution of the variance σ_ν and expectation of zero. Then according to the maximum likelihood method the target function is

$$p(D \mid \mathbf{w}, \alpha, \beta, f_i) = \frac{\exp(-\beta E_D(D \mid \mathbf{w}, f_i))}{Z_D(\beta)},$$

where $\beta = \sigma_\nu^{-2}$ and $Z_D(\beta)$ is the normalizing constant. The error function E_D is the sum of residual squares of the model values f_i and dependent variable values,

$E_D = \sum_{n=1}^{N} (f_i(\mathbf{w}, \mathbf{x}) - y_n)^2$.

The model parameters $\mathbf{w}^{\mathrm{MP}}$, which brings the maximum to the target function are called the most probable parameters.

Inductive model generation

The models are generated with the set of the primitive functions G as following. The indices of the functions g_ν are in the set $\mathrm{V} = \{1, ..., V\}$. The mapping $\iota : \mathrm{V}^r \to \mathrm{A}$ is given. The elements $A_t \in \mathrm{A}$ are the every possible combinations of K from V, where $K = 1, ..., r$. The elements of the set $A_t = \{a_t(k)\}$ have the indexes $k = 1, ..., K_t$. Since $a \in \mathrm{V}$, the elements $a_t(k)$ correspond to the functions $g_\nu \in G$. For each A_t consider the set of the incidence matrices $\rho_i(A_t)$, $i \in \mathbf{N}$. The index i of the matrix ρ defines a unique superposition f_i of the functions g; denoted as $\rho_i = \rho_i(A_t)$. The number of the elements of this superposition equals K_t. The incidence matrix $\rho_i : \{1, ..., K_t\}^2 \to \{0,1\}$ defines the orgraph and the superposition f_i. The superposition is called *acceptable* if the following conditions are held.

1. The orgraph ρ_i is acyclic.
2. The orgraph is one-connected, subject to $\sum_{l=1}^{K_t} \sum_{k=1}^{K_t} \rho_i(l,k) = \sum_{k=1}^{K_t} s(a_t(k))$, where $s = s(\nu)$ is the number of arguments of the function g_ν. The number of ones in the matrix ρ_i equals the overall number of arguments of the superposition f_i.
3. The number of arguments of every element of the superposition is equal to the number of arguments of the corresponded primitive function $\sum_{l=1}^{K_t} \rho_i(l,k) = s(a_t(k))$ for each $k = 1, ..., K_t$. The number of orgraph's vertices adjoined to the k-th node is the number $s(a_t(k))$ of arguments of the function g_ν, where $\nu = a_t(k)$.

Estimation of the model hyperparameters

Consider the set of the competitive models $f_1, ..., f_M$. When the data D have come, the posterior probability $P(f_i \mid D)$ of the model could be defined with the Bayes theorem $P(f_i \mid D) = \frac{p(D \mid f_i) P(f_i)}{\sum_{j=1}^{M} p(D \mid f_j) P(f_j)}$, where $p(D \mid f_i)$ are predictions, which model can make about the data : $p(D \mid f_i) = \int p(D \mid \mathbf{w}, f_i) p(\mathbf{w} \mid f_i) d\mathbf{w}$. It is called the evidence of the model.

The posterior probability of the parameters $\mathbf{w}$ of the model f_i given sample set D equals $p(\mathbf{w} \mid D, f_i) = \frac{p(D \mid \mathbf{w}, f_i) p(\mathbf{w} \mid f_i)}{p(D \mid f_i)}$, where $p(\mathbf{w} \mid f_i)$ is the prior probability of the parameters of the initial distribution, and $p(D \mid \mathbf{w}, f_i)$ is the likelihood function of the model parameters.

Introduce the regularization parameter α. It controls how well the model fits the data. The probability of the parameters given hyperparameter α equals

$$p(\mathbf{w} \mid \alpha, f_i) = \frac{\exp(-\alpha E_W(\mathbf{w} \mid f_i))}{Z_W(\alpha)},$$

where α corresponds the inverse variance of parameters, $\alpha = \sigma_W^{-2}$ and Z_W is the normalizing constant. The requirements to small parameter values suppose the Gaussian posterior distribution with zero-mean. For given values of the hyperparameters α and β the equation (*) for a given model f_i will be

[1] Support of the grant RFBR 07-07-00181

$$p(D \mid \mathbf{w}, \alpha, \beta, f_i) = \frac{p(D \mid \mathbf{w}, \beta, f_i) p(\mathbf{w} \mid \alpha, f_i)}{p(D \mid \alpha, \beta, f_i)} =$$

$$\frac{\exp(-S(\mathbf{w} \mid f_i))}{Z_S(\alpha, \beta)}, \text{where } S(\mathbf{w} \mid f_i) = \alpha E_W + \beta E_D$$

and Z_S is the normalizing constant. To estimate the optimal values of the parameters $\mathbf{w}$ and the hyperparameters α, β given model f_i consider an iterative algorithm. One must find the values of the hyperparameters, which bring maximum to the posterior probability of the parameters and then execute the other calculations include probability of the parameters given data with fixed values of the hyperparameters.

To specify the posterior probability $p(D \mid \mathbf{w}, \alpha, \beta)$, which uses the posterior distribution of parameters, one must approximate error function $S(\mathbf{w})$ with the second degree Taylor series: $S(\mathbf{w}) \approx S(\mathbf{w}^{\mathrm{MP}}) + 2^{-1}(\mathbf{w} - \mathbf{w}^{\mathrm{MP}}) A (\mathbf{w} - \mathbf{w}^{\mathrm{MP}})$, where the Hessian matrix $A = \nabla^2 S = \beta \nabla^2 E_D + \alpha I.$ Substitute the approximate value of $S(\mathbf{w})$ into (*) and denote $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^{\mathrm{MP}}$, obtain

$$p(\mathbf{w} \mid D, \alpha, \beta) = Z_S^{-1} \exp(-S(\mathbf{w}) - 2^{-1} \Delta\mathbf{w}^T A \Delta\mathbf{w}).$$

Evaluate the constant Z_S, which contain the hyperparameters. To estimate the hyperparameters one must optimize the function $p(D \mid \alpha, \beta)$ subject to α and β: $\ln p(D \mid \alpha, \beta) =$

$$-E_W^{\mathrm{MP}} - 2^{-1} \sum_{j=1}^{W} (\lambda_j + \alpha)^{-1} + W 2^{-1} \alpha^{-1}.$$ Set the last statement equal zero and transform it. The statement for evaluation α is $2\alpha E_W^{\mathrm{MP}} = W - \sum_{j=1}^{W} \alpha (\lambda_j + \alpha)^{-1}.$ Denote the deduction of the right part as $\gamma = \sum_{j=1}^{W} \alpha (\lambda_j + \alpha)^{-1}$. Then the optimal value of β equals $2\beta E_D^{\mathrm{MP}} = N - \gamma.$

Model generation using hyperparameters

The inductive generation of the regression models is executed iteratively. It involves the generated models and the set of the primitive functions. Before it starts, the set of the measured data D and the set of the smooth functions G were given. The initial set of the competitive models $\{f_1, ..., f_M\}$ are given. Each model f_i in the set is a superposition of the functions $g_{ij}, j = 1, ..., r_i \le r.$ The hyperparameter α_{ij} corresponds the element g_{ij} of the model f_i. It describes the initial probability distribution of the parameter vector $\mathbf{b}_{ij}$ of this function. The hyperparameter β_i corresponds to the model f_i. The initial values of the hyperparameter for *i*-th model are predefined according to the prior noise probability distribution function parameters. After the algorithm starts the following sequence of steps is executed. The sequence repeats the given number of iterations.

1. Minimize the error functions $S_i(\mathbf{w})$ for each model f_i with the Levenberg–Marquardt method [2]. Estimate the parameters $\mathbf{w}_i^{\mathrm{MP}}$ of the models.
2. Define new values of the hyperparameters $\alpha_{ij}^* = (W - \gamma_i) E_W^{-1}(\mathbf{b}_{ij})$, $\beta_i^* = (N - \gamma_i) E_D^{-1}(f_i)$. They based on the initial values of the hyperparameters. Repeat the steps 1 and 2 until the parameters will be converged.
3. According to the error function values select $2^{-1}M$ best models to the further modification. Modify each model: find the element of the superposition with minimal value of the hyperparameters α_{ij}^*; replace it for the arbitrary primitive function $g \in G$.
4. Use the selected and the modified models in the next iterations.

Conclusion

The method of the inductive generation of the parametric regression models is described. The models are superposition of the given primitive functions. The model generation algorithm uses hyperparameters of the models. The hyperparameters are estimations of the model parameters distribution function. They show the importance of the models elements. The parameters and the hyperparameters are estimated with non-linear optimization methods.

[1] Support of the grant RFBR 07-07-00181

References

1. Bishop, C.M., Tipping, M.E. Bayesian regression and classification // Suykens, J., Horvath, G. et. al., eds. Advances in Learning Theory: Methods, Models and Applications, Volume 190. IOS Press, NATO Science Series III: Computer and Systems Sciences, 2000. P 267–285.
2. Kelley, C. T., Iterative Methods for Optimization, SIAM Frontiers in Applied Mathematics, no 18, 1999.
3. MacKay, D. Information, inference, learning algorithms. Cambridge University Press, 2003.
4. MacKay, D. Hyperparameters: optimise or integrate out? // Heidberger, G., ed. Maximum entropy and Bayesian Methods. Santa Barbara, Dordrecht: Kluwer, 1993.
5. MacKay, D. Bayesian interpolation // Neural Computation 4(3), 1992. P. 415-447.
6. MacKay, D. Choice of basis for Laplace approximation // Machine Learning, vol. 33(1), 1998.
7. Nabney, I. T. NETLAB: Algorithms for pattern recognition. Springer, 2004. P. 330.

[1] Support of the grant RFBR 07-07-00181